The impact of outcome reporting bias in randomised controlled trials on a cohort of systematic reviews

Jamie J Kirkham,[1] Kerry M Dwan,[1] Douglas G Altman,[2] Carrol Gamble,[1] Susanna Dodd,[1] Rebecca Smyth,[3] Paula R Williamson[1]

[1]Centre for Medical Statistics and Health Evaluation, University of Liverpool, Liverpool L69 3GS
[2]Centre for Statistics in Medicine, University of Oxford, Oxford OX2 6UD
[3]Population, Community and Behavioural Sciences, University of Liverpool, Liverpool L69 3GB
Correspondence to: P R Williamson prw@liv.ac.uk
Accepted: 26 October 2009

Cite this as: *BMJ* 2010;340:c365
doi: 10.1136/bmj.c365

Objective To examine the prevalence of outcome reporting bias—the selection for publication of a subset of the original recorded outcome variables on the basis of the results—and its impact on Cochrane reviews.

Design A nine point classification system for missing outcome data in randomised trials was developed and applied to the trials assessed in a large, unselected cohort of Cochrane systematic reviews. Researchers who conducted the trials were contacted and the reason sought for the non-reporting of data. A sensitivity analysis was undertaken to assess the impact of outcome reporting bias on reviews that included a single meta-analysis of the review primary outcome.

Results More than half (157/283 (55%)) the reviews did not include full data for the review primary outcome of interest from all eligible trials. The median amount of review outcome data missing for any reason was 10%, whereas 50% or more of the potential data were missing in 70 (25%) reviews. It was clear from the publications for 155 (6%) of the 2486 assessable trials that the researchers had measured and analysed the review primary outcome but did not report or only partially reported the results. For reports that did not mention the review primary outcome, our classification regarding the presence of outcome reporting bias was shown to have a sensitivity of 88% (95% CI 65% to 100%) and specificity of 80% (95% CI 69% to 90%) on the basis of responses from 62 trialists. A third of Cochrane reviews (96/283 (34%)) contained at least one trial with high suspicion of outcome reporting bias for the review primary outcome. In a sensitivity analysis undertaken for 81 reviews with a single meta-analysis of the primary outcome of interest, the treatment effect estimate was reduced by 20% or more in 19 (23%). Of the 42 meta-analyses with a statistically significant result only, eight (19%) became non-significant after adjustment for outcome reporting bias and 11 (26%) would have overestimated the treatment effect by 20% or more.

Conclusions Outcome reporting bias is an under-recognised problem that affects the conclusions in a substantial proportion of Cochrane reviews. Individuals conducting systematic reviews need to address explicitly the issue of missing outcome data for their review to be considered a reliable source of evidence. Extra care is required during data extraction, reviewers should identify when a trial reports that an outcome was measured but no results were reported or events observed, and contact with trialists should be encouraged.

Selective reporting bias in a study is defined as the selection, on the basis of the results, of a subset of analyses to be reported. Selective reporting may occur in relation to outcome analyses,[1] subgroup analyses,[2] and per protocol analyses, rather than in intention to treat analyses,[3] as well as with other analyses.[4] Three types of selective reporting of outcomes exist: the selective reporting of some of the set of study outcomes, when not all analysed outcomes are reported; the selective reporting of a specific outcome—for example, when an outcome is measured and analysed at several time points but not all results are reported; and incomplete reporting of a specific outcome—for example, when the difference in means between treatments is reported for an outcome but no standard error is given.

A specific form of bias arising from the selective reporting of the set of study outcomes is outcome reporting bias, which is defined as the selection for publication of a subset of the original recorded outcome variables on the basis of the results.[5] Empirical research on randomised controlled trials shows strong evidence of an association between significant results and publication: studies that report positive or significant results ($P<0.05$) are more likely to be published, and outcomes that are statistically significant have higher odds of being fully reported than those that are not significant (range of odds ratios: 2.2 to 4.7).[6] An analysis of studies that compared trial publications with protocols found that 40-62% of trials changed, introduced, or omitted at least one primary outcome.[6]

The systematic review process has been developed to minimise biases and random errors in the evaluation of healthcare interventions.[7] Cochrane systematic reviews are internationally recognised as among the best sources, if not the best source, of reliable up to date information on health care.[8 9] Meta-analysis, a statistical technique for combining results from several related but independent studies, can make important contributions to medical research—for example, by showing that there is evidence to support treatments not widely used[10] or that evidence is lacking to support treatments that are in wide use.[11]

Missing outcome data can affect a systematic review in two ways. Publication bias, where a study is not published on the basis of its results, can lead to bias in the analysis of a particular outcome in a review, especially if the decision not to submit or publish the study is related to the results for that outcome. In a published study that has been identified by the reviewer, outcome reporting bias can arise if the

Table 1 | Example of a review outcome matrix displaying the information available in trial reports

Trial ID (author, year of publication)	Review primary outcome	Other review outcomes		Additional outcomes (reported in any of the eligible trials)		Reason for exclusion
	Live birth rate	Chemical pregnancy rate	Clinical pregnancy rate	Ectopic pregnancy rate	Birth weight of baby	
12345678.1 (Smith, 1999)	o	×	√	×	×	—
12345678.2 (Lowe, 2001)	√	o	×	√	×	—
12345678.3 (Biggs, 2004)	×	√	√	×	√	—
...						
Excluded trials						
1234578.9 (Johns, 2006)	×	×	×	×	×	No relevant outcome data
...						

√ Full reporting of results for treatment comparison of interest.
× No reporting of results for treatment comparison of interest.
o Partial reporting of results for treatment comparison of interest.

outcome of interest in the review had been measured and analysed but not reported on the basis of the results.

Little is known about the impact of outcome reporting bias on systematic reviews. One previous study examined a small cohort of nine Cochrane reviews of randomised trials.[1] Although outcome reporting bias in the review primary outcome was suspected in several individual randomised trials, the impact of such bias on the conclusions drawn in the meta-analyses was minimal. This study used a very select set of reviews, however, and highlighted the need for a larger study.

In this paper we report the findings of the Outcome Reporting Bias in Trials (ORBIT) study, in which we applied a new classification system for the assessment of selective outcome reporting and evaluated the validity of the tool. We used the classification system to estimate the prevalence of outcome reporting bias and its impact on an unselected cohort of Cochrane reviews. To our knowledge, this is the first systematic empirical study of the impact of outcome reporting bias in randomised controlled trials on the results of systematic reviews.

Methods

We examined an unselected cohort of new reviews from 50 of the 51 Cochrane collaboration review groups published in three issues of the *Cochrane Library* (Issue 4, 2006, Issue 1, 2007, and Issue 2, 2007). For each review, two investigators (JJK and SD) independently examined the "types of outcome measures" section to determine whether the review specified a single primary outcome. For those reviews where either no primary outcome was detailed or multiple primary outcomes were specified, the lead reviewer was contacted and asked to select a single primary outcome from those listed. When no contact could be established or the reviewer(s) could not define a single primary outcome, two investigators (PRW and SD) independently selected and agreed upon a single primary outcome from those listed.

Assessment of systematic reviews

Two investigators (JJK and SD) scrutinised all 33 reviews from Issue 4, 2006 that specified a single primary outcome and agreed on the need for further assessment of all but two reviews. Both disagreements were related to whether the reasons for exclusion were suggestive of outcome reporting bias. Each remaining review was read by one investigator (JJK) to check whether all included trials fully reported the review primary outcome. The reason for exclusion of any trial (in the "characteristics of excluded studies" section) was also checked for any suggestion of potential outcome reporting bias. For example, a trial excluded because there was "no relevant outcome data" required further scrutiny because the relevant outcome might have been measured but not reported. Any uncertainties regarding the excluded studies were referred to PRW.

Reviews that did not identify any randomised controlled trials were not assessed further. Similarly, reviews were not assessed further if no standard definition of the primary outcome exists, because outcome reporting bias assessment in this situation would be impossible. One example is relapse in schizophrenia trials, for which definitions include a change in symptom score and hospital readmission.

Classification of randomised controlled trials in systematic reviews

For each review, an outcome matrix was constructed showing the reporting of the primary outcome and other outcomes in each trial included, distinguishing full, partial, or no reporting. An example of an outcome matrix is given in table 1. For this example, "live birth" was the review primary outcome. The matrix was completed using the information in the review and revised accordingly in light of any extra information obtained from the trial reports or through contact with the trialists. Outcomes for which the data could be included in a meta-analysis were considered to be fully reported. Such data may have been in the trial report or may have been calculated indirectly from the results. For example, the number of events may have been calculated from the proportion of events and the number of patients in the treatment group, or the standard error of the treatment effect may have been calculated from the estimate of effect and the associated P value.

A classification system was developed to assess the risk of bias when a trial was excluded from a meta-analysis, either because the data for the outcome were not reported or because the data were reported incompletely (for example, just as "not significant"). The system was refined over the initial few months of the study, but if an amendment was made all previous classifications were reviewed and adjusted as appropriate to ensure consistency of application. The categories reflect the stages of assessing whether an outcome was measured, whether an outcome was analysed, and, finally, the nature of the results presented (table 2). The system identifies whether there is evidence that the outcome was measured and analysed but only partially reported (A to D classifications), whether the outcome was measured but not necessarily analysed (E and F), if it is unclear whether the outcome was measured (G and H), or if it is clear the outcome was not measured (I).

For each classification category, an assessment was made of the risk of outcome reporting bias arising from the lack of inclusion of non-significant results. A "high risk" classification was awarded when it was either known or suspected that the results were partially or not reported because the treatment comparison was statistically non-significant (P>0.05).

Table 2 | The Outcome Reporting Bias In Trials (ORBIT) study classification system for missing or incomplete outcome reporting in reports of randomised trials

	Description	Level of reporting	Risk of bias*
Clear that the outcome was measured and analysed			
A	Trial report states that outcome was analysed but only reports that result was not significant (typically stating P>0.05)	Partial	High risk
B	Trial report states that outcome was analysed but only reports that result was significant (typically stating P<0.05)	Partial	No risk
C	Trial report states that outcome was analysed but insufficient data were presented for the trial to be included in meta-analysis or to be considered to be fully tabulated	Partial	Low risk
D	Trial report states that outcome was analysed but no results reported	None	High risk
Clear that the outcome was measured			
E	Clear that outcome was measured but not necessarily analysed. Judgment says likely to have been analysed but not reported because of non-significant results	None	High risk
F	Clear that outcome was measured but not necessarily analysed. Judgment says unlikely to have been analysed but not reported because of non-significant results	None	Low risk
Unclear whether the outcome was measured			
G	Not mentioned but clinical judgment says likely to have been measured and analysed but not reported on the basis of non-significant results	None	High risk
H	Not mentioned but clinical judgment says unlikely to have been measured at all	None	Low risk
Clear that the outcome was not measured			
I	Clear that outcome was not measured	NA	No risk

*Risk of bias arising from the lack of inclusion of non-significant results when a trial was excluded from a meta-analysis or not fully reported in a review because the data were unavailable.

A "low risk" classification was awarded when it was suspected, but not actually known, that the outcome was either not measured, measured but not analysed, or measured and analysed but either partially reported or not reported for a reason unrelated to the results obtained. A "no risk" classification was reserved for cases where it was known that the outcome was not measured, known that it was measured but not analysed, or known that it was measured and analysed but the reason for partial or no reporting was not because the results were statistically non-significant. For cases where the outcome was measured but not necessarily analysed, judgment was needed as to whether it was likely (E) or unlikely (F) that the measured outcome was analysed and not reported because of non-significant results. When it was unclear whether the outcome was measured, judgment was needed as to whether it was likely that the outcome was measured and analysed but not reported on the basis of non-significant results (G) or unlikely that the outcome was measured at all (H). Trials classified as A/D/E/G, C/F/H, and B/I were assumed to be at high, low, and no risk of outcome reporting bias, respectively, in relation to the review primary outcome. Examples of each of the classifications in the ORBIT study are shown in web table A.

On the basis of all identified publications for a trial, one investigator (JJK, SD, or KD) and the corresponding review author independently classified any trial that did not report or partially reported results for the review primary outcome (table 2). All trials excluded from the review but selected for assessment were also classified. For each classification, justification for the classification was recorded in prose to supplement the category code, including verbatim quotes from the trial publication whenever possible. The agreed classification, with the justification, was then reviewed by the senior investigator (PRW). Any discrepancies were discussed until a final overall classification was agreed for each trial and the justification for the classification documented in full. When the corresponding review author and coauthors were unable to assist with our assessments and the clinical area proved to be challenging, help was sought from medical colleagues at the University of Liverpool.

To assess how many reviewers had considered the possibility of outcome reporting bias, we searched the text of included reviews for the words "selective" and "reporting."

Accuracy of classification

For trials for which it was uncertain whether the review primary outcome had actually been measured and/or analysed (E, F, G, or H classification; table 2), the trialists were contacted via email (address obtained from either the trial report or a search of PubMed or Google) and asked to confirm whether the review primary outcome was measured and analysed. If so, the reason for not reporting the results was requested. Non-responders were contacted a second time if a reply was not received within three weeks. Trialists were not contacted if a reviewer had previously approached them for the relevant information.

Two separate sensitivity and specificity analyses were performed. The first analysis considered only G and H classifications and aimed to determine how good our classification system was at judging whether the primary outcome of interest in the review had been measured when it was not mentioned in the trial report. For this analysis only, we incorporated an extra category of G classification for trials with binary outcomes where we predicted that the outcome was measured but it was not reported because there were no events.

The second analysis compared our classifications with information from the trialists to establish whether we could predict if biased reporting had occurred. Implicitly, E and G classifications suggested that bias was likely because it was either clear or assumed that the outcome had been measured and possible that non-reporting could have been influenced by the non-significance of the result. These classifications were taken to imply bias on the basis of the lack of inclusion of non-significant results. The specificity was calculated taking F and H classifications to indicate no bias. This analysis excluded any studies classified as F that were ongoing because it is difficult to assess bias until a study is completed. Confidence intervals for sensitivity and specificity estimates were calculated using standard formulae.[12]

Amount and impact of missing trial data

The amount of missing data per review was calculated, firstly on the basis of trials that omitted data for any reason and secondly only using those trials where data omission was suspected on the basis of the results (that is, outcome reporting bias was suspected). The maximum bias bound approach was used in a sensitivity analysis[13 14] to estimate the impact of outcome reporting bias on the review meta-analysis. This approach calculates an upper bound for the bias resulting from the number of eligible studies suspected of outcome reporting bias, and assumes that on average smaller studies (lower precision) will have a higher probability of not reporting the outcome of interest than larger studies (higher precision). This method was applied only to reviews that had a single meta-analysis of the review primary outcome, because

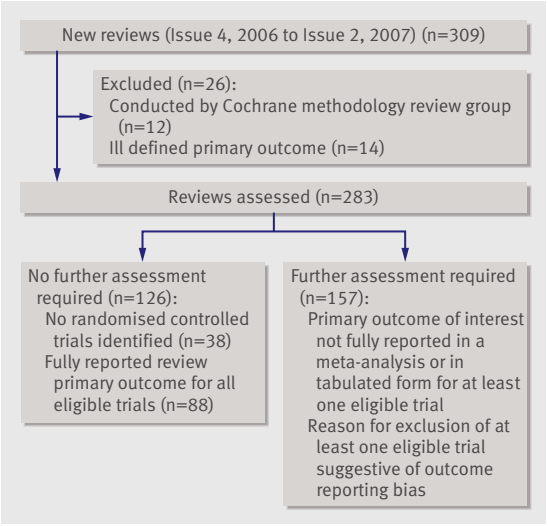

Fig 1 | Flow diagram for Outcome Reporting Bias In Trials (ORBIT) study

if there were multiple meta-analyses it would be difficult to ascertain to which analyses the trial with suspected outcome reporting bias would relate without discussion with a clinical expert. The impact was not assessed for trials with H or I classifications, where it was suggested that the review primary outcome had not been measured, or G classifications where the explanation was that there were no events. The impact was assessed both in terms of the percentage change in the treatment effect estimate and the change in the statistical significance of the treatment effect estimate after adjustment.

Results

Assessments of systematic reviews

The *Cochrane Library* published 309 new reviews in Issue 4, 2006, Issue 1, 2007, and Issue 2, 2007 (fig 1). We excluded 12 reviews by the Cochrane Methodology Review Group. Single primary outcomes were specified in 103 reviews, whereas lead reviewers or co-reviewers were asked to select a single primary outcome for the remaining 194 reviews. In 173 cases reviewers were willing to do so, with 127 (73%) choosing the first outcome listed. For the remaining 21 reviews a single primary outcome was selected by the research team (PRW and SD). On further scrutiny, however, 14 reviews were excluded because the review primary outcome was not well defined.

Among the remaining 283 reviews, the median number of reviews from an individual Cochrane review group was five (range 1 to 21, interquartile range (IQR) 2 to 7). The five groups with most reviews were the hepato-biliary group (21 reviews), the pregnancy and childbirth group (18), the neonatal group (14), the oral health group (13), and the menstrual disorders and subfertility group (12). The median number of randomised controlled trials per review was five (range 0 to 134, IQR 2 to 10).

A total of 126 reviews did not require further assessment: 38 did not identify any randomised controlled trials and 88 fully reported the primary outcome for all eligible trials. This left 157 reviews requiring further assessment—that is, 55% (157/283) of reviews did not include full data on the primary outcome of interest from all eligible trials.

By text searching for the words "selective" and "reporting," 20 (7%) of the 283 reviews assessed were found to have mentioned outcome reporting bias, the proportion being similar in reviews requiring and those not requiring further assessment.

Full reporting of review primary outcomes in trials

Figure 2 shows a flow diagram for the assessment of the 2562 trials included in the study cohort of 283 systematic reviews. Seventy-six trial reports could not be assessed because the articles were not in English. Seventy-one per cent (1774/2486) of the remaining trials fully reported the review primary outcome in the trial report.

Table 3 provides information on 177 trial reports that gave full data on the primary outcome of interest that was not included in the review. For 59 trials, the data were not included in the review for a reason unrelated to outcome reporting bias. For 118 trials (7% of the 1774 trials that fully reported the review primary outcome), the review primary outcome data were fully reported in the publication but were not included in the review. Information on missed outcome data was fed back to the reviewers for inclusion in a review update.

Classification of trials

For 788 (31%) of the 2562 trials included in our study, the review primary outcome was either partially reported or not reported (fig 2). Seventy-six trial reports could not be assessed because the articles were not in English, leaving 2486 assessable trials and 712 trial reports requiring a classification (545 included in reviews and 167 excluded from reviews). Table 4 shows the classification of these 712 trials.

For 155 (6%) of the 2486 assessable trials, it was clear that the review primary outcome was measured and analysed (A, B, C, or D classification), but partial reporting meant the data could not be included in a meta-analysis. Trials classified as C were grouped according to the nature of the missing data (web table B).

A total of 359 (50%) of the 712 trials with missing data were under high suspicion for outcome reporting bias (A, D, E, or G classification; table 4). The prevalence of reviews

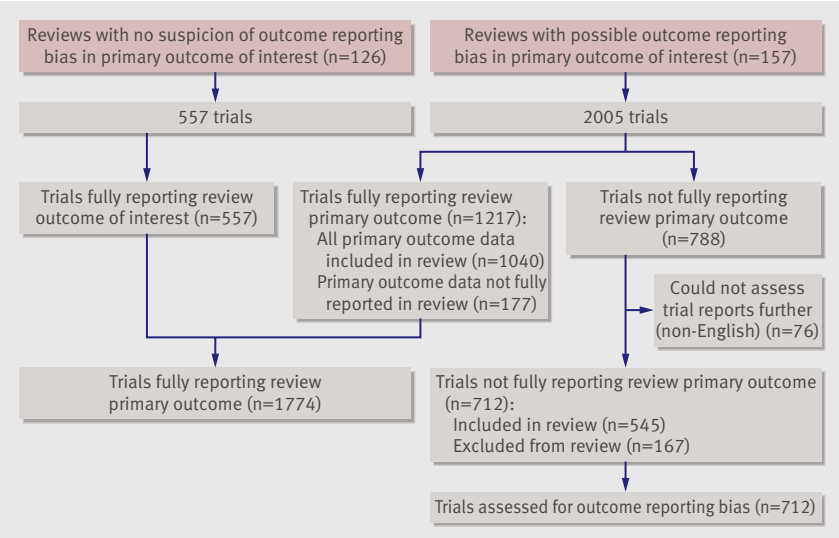

Fig 2 | Assessment of randomised controlled trials within reviews

Table 3 | Reasons for omission of data from trials fully reporting review primary outcome (n=177)

	Reason	Number of trials
Data not included in review for a reason unrelated to outcome reporting bias (n=59)		
Invalid measurement scales	In some reviews only certain validated measurement scales were allowed. Cases in which the primary outcome was deemed to be fully reported using a non-validated scale and there was no apparent evidence of outcome reporting bias were accepted as full reporting.	4
Poor reporting of time to event data	For inclusion in a time to event meta-analysis, the log hazard ratio and a measure of its variance is required. Although this information was not reported in these trials, enough information was reported to rule out outcome reporting bias. a. Review tabulates median time to event (no meta-analysis considered), whereas trials fully report the number of events as a binary outcome in each treatment arm (n=19) b. Review reports the number of events in a binary outcome meta-analysis, whereas the trials report the median time to event, Kaplan-Meier plot, and significance of the difference in survival curves for each treatment arm using log rank test (n=12)	31
Quality issues	The review primary outcome was fully reported in the trial report but the results were not included in the review owing to methodological shortcomings (for example, the trial was a crossover trial with no washout period). This was acceptable as full reporting if the primary outcome data were fully reported and the reasons for these shortcomings were discussed by the reviewer. These methodological shortcomings were considered to be quality issues not related to outcome reporting bias.	24
Data not included in review despite being fully reported in trial (n=118)		
Not fully reported in the review text*	The results were fully reported in the trial report but only partially reported in the review text (no meta-analysis undertaken). a. Review reported the P values only (n=19) b. Review reported the magnitude of treatment effect (group means or medians, or difference in means) but with no measure of precision or variability (confidence interval, standard deviation, or standard error for means; interquartile or other range for medians; n=7) c. Review reported the number of participants with the event for each group (or percentages) but did not give sample sizes for the denominators (n=2) d. Review reported the results from the main intervention arm only (n=1)	29
No event*	The primary outcome was not observed in any patient throughout the trial, which was mentioned in the trial report but not in the review.	31
No results reported in review*	The results were fully reported in the trial report but nothing was reported in the review. a. The outcome data were missed during data extraction (perhaps reported in a supplementary article rather than the main publication; n=42) b. The outcome data were available but extraction was not straightforward—that is, perhaps some calculation was involved before the data were in a suitable format for inclusion in a review meta-analysis (n=8) c. Trial reported results from a non-parametric analysis and review included only parametric results for use in a meta-analysis (n=8)	58

*This information was forwarded to the reviewer.

Table 4 | Trials assessed for outcome reporting bias (n=712)

Classification	Number of fully published trials	Number of abstracts	Total number of trials (%)
A	23	7	30 (4)
B	2	6	8 (1)
C	113	4	117 (16)
D	0	0	0 (0)
E	113	9	122 (17)
F	24	9	33 (5)
G	192	15	207 (29)
H	148	28	176 (25)
I	15	4	19 (3)
Total	**630**	**82**	**712**

Table 5 | Accuracy of judgment as to whether the review primary outcome was measured (G or H classification)

			Information from trialist		
			Primary outcome measured	Primary outcome not measured	Total
ORBIT assessment	Primary outcome measured	G classification	4	7*	11
		G classification (no event)	19	0	19
		Total	**23**	**7**	**30**
	Primary outcome not measured	H classification	2*	23	25
Total			**25**	**30**	**55**

*Reasons for these disagreements are given in table 6.

containing at least one trial with high outcome reporting bias suspicion was 34% (96/283).

Accuracy of classification

Information on whether the outcome of interest was measured and analysed was lacking in 538 trial reports (E, F, G, or H classification). We found the email addresses of 167 (31%) authors and contacted these individuals. Responses were received from 65 authors (39%): 26% (9/34) of authors whose trial had an E classification; 33% (1/3) who got an F classification; 42% (30/71) who got a G classification; and 42% (25/59) of individuals from trials with an H classification.

To determine whether the outcome of interest was measured or not, we compared our assessments against the trialists' information for 55 trials for which the outcome had not been mentioned in the trial report (G or H classification). The sensitivity for predicting that the outcome had been measured was 92% (23/25, 95% CI 81% to 100%), whereas the specificity for predicting that the outcome had not been measured was 77% (23/30, 95% CI 62% to 92%; table 5). Details of the nine incorrect classifications are provided in table 6.

To measure our judgment on whether outcome reporting bias occurred or not, we compared our assessments against the trialists' information for 62 trials for which the outcome was either clearly measured but not necessarily analysed (E and F classification) or had not been mentioned in the trial report (G or H classification). Three ongoing studies were excluded from this analysis. The sensitivity of our classification system for detecting bias was calculated to be 88% (7/8, 95% CI 65% to 100%), whereas the specificity was 80% (43/54, 95% CI 69% to 90%; table 7).

Amount and impact of missing trial data

The median amount of review primary outcome data missing from trials for any reason was 10%. For the 96 reviews that included at least one trial with a high suspicion of outcome reporting bias, the median amount of missing data was 43%.

Of the 283 reviews in our study cohort, 81 included a single meta-analysis of the review primary outcome and were included in the assessment of the impact of outcome reporting bias on the review meta-analysis. Table 8 lists the reasons for excluding reviews from the assessment of impact.

Table 6 | Reasons for incorrect judgment as to whether the outcome of interest was measured in a trial (G or H classification)

ORBIT study classification	Primary outcome of interest	Information from trialist	Reason for incorrect classification
Likely to have been measured			
G	Cause specific survival	Data were not reported on this outcome, only on overall survival	We thought it possible that the cause of death would have been recorded if it was breast cancer, which patients in the trial had been diagnosed with
G	Cognitive development	No data on cognitive development, only evaluated motor development	A number of trials in this review reported on both cognitive development and motor development
G	Bone fractures*	Bone fractures not measured, only bone mineral density	Although "bone fractures" is a long term outcome, one short term trial included in the review reported no bone fractures. It was thought that all similar trials would have the ability to detect a bone fracture even though it is unlikely that an event occurred. There is also lack of consensus between experts in this field on whether it is plausible to accurately detect bone fractures using the technology used in these trials
G	Pain response to bisphosphonates	Pain was only looked at through analgesic consumption but was not measured using the visual analogue scale required for review	It was clear that pain was an outcome domain of interest in this trial, and most other included trials reported pain using a visual analogue scale scale
G	Improvement in nerve function	This outcome was not measured. The trial assessed function using only a clinician's judgment, as would happen in clinical practice. Changes in skin, motor, sensory, and autonomic function are complex to measure, and reliability of measurement varies. It is difficult to determine what changes would be considered clinically significant to individual patients, so the study was not based on such measurements	It was clear that nerve function was an outcome domain of interest in this trial. We thought that since the other two trials included in this review reported on this outcome by using validated sensory and muscle testing scores, then this outcome would have also been measured in this trial in addition to the clinical assessments. There is lack of consensus between clinical experts on the validity and reliability of using validated test scores in this clinical field
Not likely to have been measured			
H	Mean weekly alcohol consumption	Mean weekly alcohol consumption was measured in the original study, but the primary results paper had not been written at the time of the early publication cited in the review. Results still not published	This study was excluded because no prespecified outcomes were mentioned. The trial report included in the review looked only at healthcare utilisation and did not report any outcome data that suggested that the review primary outcome would have been measured
H	Live birth rate	Data were collected on live birth rate but were not complete at time of publication. All pregnancies resulted in a live birth. Result still not published but data now analysed (P>0.05)	The primary end point for this trial was the number of clinical pregnancies diagnosed at 12 weeks' gestation. On the basis of the studies included in this review, it seems that trials in this area often do not follow-up to birth

*This reason applied to three separate trials within the same review.

Table 7 |Accuracy of judgment as to whether outcome reporting bias occurred (E, F, G, or H classification)

		Information from trialist		
		Bias	No bias	Total
ORBIT assessment	High risk	7(4* + 3†)	11(7‡ + 4§)	18
	Low risk	1¶	43(24** + 19††)	44
Total		**8**	**54**	**62**

*Review primary outcome measured but not analysed owing to small number of events.
†Review primary outcome measured and analysed but result not significant (P>0.05; one case), or result analysed but trialist would not share significance of result until article published (two cases).
‡Review primary outcome not measured (all incorrectly predicted—see all G classifications, table 6).
§Review primary outcome measured but not analysed because it was not a specific end point in the trial (one case), was measured in a small subset of patients in one treatment arm but not analysed (one case), or was analysed but favoured intervention (P<0.05; two cases).
¶See "Live birth rate" example, table 6.
**Review primary outcome not measured.
††Review primary outcome measured but no events recorded.

Table 8 | Reasons for excluding reviews from the assessment of impact

Reason	Number of reviews
Total number of reviews identified	**309**
Preliminary exclusions	
Study by Cochrane methods group	12
Primary outcome not well defined	14
Total number reviews included in the study	**283**
Exclusions from assessment of impact	
Review identified no randomised controlled trials	38
Language restrictions	2
No meta-analysis	45
Primary outcome measured in different ways (for example, weight might have been reported as BMI or change in weight)	20
Longitudinal study	15
Studies not combined owing to clinical heterogeneity	4
Review included several meta-analyses (owing to different intervention comparisons)	78
Total number of reviews included in assessment of impact	**81**

A total of 52 of the 81 reviews included in the assessment of impact included at least one trial that had a high suspicion of outcome reporting bias. In 27 of these 52 reviews, no sensitivity analysis was undertaken because classifications for all trials with missing data suggested that the review primary outcome seemed not to have been measured or it was suspected that there were no events (H and some G classifications, respectively; 17 reviews), or the reviewer or review text suggested that the missing studies would not have been combined with the other trials in the meta-analysis for reasons not related to outcome reporting bias (10 reviews). For the other 25 reviews that could be assessed, the maximum bias bound sensitivity analysis indicated that the statistically significant conclusions of eight of these reviews were not robust to outcome reporting bias—that is, the treatment effect estimate changed from a significant result favouring treatment (95% confidence interval excludes the null value) to a non-significant result (reviews one to eight; table 9). In a further eight analyses, the result was robust to outcome reporting bias—that is, the result for the adjusted pooled estimate was also statistically significant (P<0.05). The remaining nine analyses had non-significant treatment effect estimates for which the application of the sensitivity analysis produced no substantial change in three analyses and a change from favouring one group to moving the effect estimate closer to the null value of no difference in treatment effect in six analyses. For all the 25 reviews assessed, the median percentage change in the treatment effect estimates after the adjustment based on the maximum bias bound was 39% (IQR 18% to 67%).

Our sensitivity analysis indicates that of the 81 reviews where there was a single meta-analysis of the review

Table 9 | Sensitivity analysis to assess the robustness of the conclusions of the review to outcome reporting bias (n=25 reviews)

Review	Intervention*	Number of trials with results fully reported in meta-analysis (n)	Number of eligible trials missing from meta-analysis and suspected of outcome reporting bias (m)	Proportion of missing data (%)†	Original pooled estimate (95% confidence interval)	Conclusion	Adjusted pooled estimate (95% confidence interval)‡	Change in estimate (%)
1	Active treatment *v* placebo/nothing	6	3	45	HR 0.57 (0.39 to 0.82)	Favours active treatment	HR 0.73 (0.51 to 1.06)§	37¶
2	Active treatment *v* placebo/nothing	4	4	11	RR 0.49 (0.26 to 0.90)	Favours active treatment	RR 0.79 (0.42 1.46)§	59
3	Active treatment *v* placebo/nothing	3	3	81	WMD 0.39 (0.11 to 0.67)	Favours active treatment	WMD 0.21 (−0.07 to 0.49)§	46
4	Active treatment *v* placebo/nothing	4	2	20	SMD 0.66 (0.20 to 1.12)	Favours active treatment	SMD 0.41 (−0.05 to 0.88)§	38
5	Active treatment *v* placebo/nothing	9	4	10	RR 0.49 (0.32 to 0.74)	Favours active treatment	RR 0.67 (0.45 to 1.02)§	35
6	Active treatment 1 *v* active treatment 2	29	9	18	RD -0.04 (−0.07 to −0.01)	Favours active treatment 1	RD −0.02 (−0.05 to 0.01)§	50
7	Active treatment 1 *v* active treatment 2	5	1	7	RR 0.27 (0.09 to 0.81)	Favours active treatment 2	RR 0.38 (0.13 to 1.12)§	15
8	Active treatment *v* placebo/nothing	14	1	3	RR 0.31 (0.11 to 0.91)**	Favours active treatment	RR 0.39 (0.13 to 1.12)§	12
9	Active treatment *v* placebo/nothing	1	4	78	WMD 1.09 (0.48 to 1.70)	Favours active treatment	WMD 0.66 (0.05 to 1.27)	39
10	Active treatment *v* placebo/nothing	2	1	30	WMD 0.42 (0.14 to 0.69)	Favours active treatment	WMD 0.31 (0.03 to 0.58)	26
11	Active treatment 1 *v* active treatment 2	1	9	81	RR 0.55 (0.40 to 0.76)	Favours active treatment 1	RR 0.63 (0.46 to 0.87)	18
12	Active treatment 1 *v* active treatment 2	21	1	2	OR 0.24 (0.18 to 0.30)	Favours active treatment 1	OR 0.25 (0.19 to 0.32)	1
13	Active treatment 1 *v* active treatment 2	4	1	18	RD −0.17 (−0.24 to −0.10)	Favours active treatment 1	RD −0.09 (−0.21 to −0.07)	47
14	Active treatment *v* placebo/nothing	34	16	50	WMD −1.27 (−1.58 to −0.97)	Favours active treatment	WMD −0.79 (−1.10 to −0.49)	38
15	Active treatment *v* placebo/nothing	13	3	11	RR 0.62 (0.52 to 0.75)	Favours active treatment	RR 0.69 (0.58 to 0.83)	18
16	Active treatment *v* placebo/nothing	9	3	44	WMD 3.70 (−1.19 to 8.60)	Favours active treatment	WMD 0.69 (−4.20 to 5.59)	81
17	Active treatment *v* placebo/nothing	13	3	19	SMD −0.87 (−1.37 to −0.36)	Favours active treatment	SMD −0.57 (−1.08 to −0.06)	34
18	Active treatment 1 *v* active treatment 2	13	2	8	RR 0.85 (0.68 to 1.07)	Favours active treatment 1	RR 0.93 (0.73 to 1.17)	53
19	Active treatment *v* placebo/nothing	2	1	66	Peto's OR 1.51 (0.79 to 2.87)	Favours active treatment	Peto's OR 1.17 (0.61 to 2.23)	67
20	Active treatment 1 *v* active treatment 2	9	2	17	RR 0.77 (0.48 to 1.22)	Favours active treatment 1	RR 0.99 (0.62 to 1.57)	96
21	Active treatment *v* placebo/nothing	2	1	67	WMD 0.38 (−0.39 to 1.15)§	Favours placebo/ nothing	WMD 0.08 (−0.69 to 0.85)	79
22	Active treatment 1 *v* active treatment 2	1	1	18	RR 1.13 (0.85 to 1.49)	Favours active treatment 2	RR 1.01 (0.76 to 1.33)	92
23	Active treatment *v* placebo/nothing	3	1	50	RR 1.15 (0.80 to 1.65)	Favours placebo/ nothing	RR 1.00 (0.70 to 1.44)	100
24	Active treatment 1 *v* active treatment 2	13	1	1	RD −0.03 (−0.1 to 0.03)	Favours active treatment 1	RD −0.01 (−0.08 to 0.05)	67
25	Active treatment *v* placebo/nothing	4	1	20	OR 1.12 (0.72 to 1.73)	Favours placebo/ nothing	OR 0.98 (0.64 to 1.52)	13

*"Placebo/nothing" implies that the intervention was given as an add on therapy—that is, patients in both arms received standard care.
†Calculated as participants in trials missing from meta-analysis and suspected of outcome reporting bias divided by participants in trials missing from meta-analysis and suspected of outcome reporting bias plus participants in trials with results fully reported.
‡The maximum bias bound was calculated and then added to or subtracted from the original pooled estimate to move it closer towards the null.
§Indicates loss of significance.
¶Calculated as (0.73−0.57)/(1−0.57).
**Subtotals not combined in review; subtotals combined here for this analysis.
Abbreviations: HR, hazard ratio; OR, odds ratio; RD, risk difference; RR, relative risk; SMD, standardised mean difference; WMD, weighted mean difference.

primary outcome, the significance of the results was not robust to outcome reporting bias in eight (10%) cases and the treatment effect estimate was reduced by more than 20% in 19 (23%) reviews. If only the 42 meta-analyses with a statistically significant result are considered, however, then eight (19%) become non-significant after adjustment for outcome reporting bias and 11 (26%) overestimated the treatment effect estimate by 20% or more.

Discussion

Outcome reporting bias was suspected in at least one randomised controlled trial in more than a third of the systematic reviews we examined (35%), which is substantially higher than the number of reviews in which a reference to the potential for outcome reporting bias was found (7%), thus demonstrating under-recognition of the problem. We have also shown through sensitivity analysis that outcome report-

ing bias affects the treatment effect estimate in a substantial proportion of Cochrane reviews.

Strengths and limitations of the study

The strengths of this study are that we evaluated a large, unselected cohort of reviews, review authors were involved in the assessment of outcome reporting bias, and the authors of the trials included in the reviews were contacted for information. In addition, the textual justification for each trial classification was checked by a senior investigator.

We undertook an internal pilot study of 33 reviews to determine the level of agreement between two researchers on the need for further assessment of a review for suspicion of outcome reporting bias. Given that agreement was high, we concluded that it would be sufficient for a single reviewer to assess the remainder of the reviews, provided a second reviewer checked the reasons for excluded studies where there was uncertainty.

For the majority of trials that were missing outcome data, judgment was needed regarding the potential for outcome reporting bias. We believe we have shown that sufficiently accurate assessments are possible. This conclusion, however, rests on the assumption that the trialists we contacted provided accurate information to us. A previous study suggested that trialists may be reluctant to admit selective reporting.[15] In our study, the response rate for those trialists for whom an email address was obtained was similar in trials with a high risk classification and those with a low risk classification. If response bias was operating, we would expect the sensitivity of our classifications to be underestimated (as a result of trialists with high risk classifications being less likely to respond if they have selectively reported outcomes) and the specificity overestimated (as a result of trialists with low risk classifications being more likely to respond if they have not selectively reported outcomes). With such response bias, the number of selectively reported trials in a review would be underestimated; thus the impact of outcome reporting bias on the conclusions of the reviews studied here may have been underestimated.

Our classifications of trials for outcome reporting bias facilitated an assessment of the robustness of review conclusions to such bias.[13 14] The maximum bias bound approach was the method chosen to examine this source of bias because it can be applied to any outcome type. Although only 81 (29%) of the 283 reviews studied comprised a single meta-analysis of the primary outcome of interest and were thus included in the assessment, there is no reason to believe the results of this assessment would not be generalisable to those reviews containing multiple meta-analyses of the primary outcome relating to different treatment comparisons. However, a limitation of our study is that it has not examined how the impact of outcome reporting bias should be assessed in reviews that do not include a meta-analysis.

Comparison with other studies

We are only aware of one previous study that used similar methods to examine the prevalence of outcome reporting bias and its impact on systematic reviews.[1] This study used a highly selected set of nine reviews, however, in which 10 or more trials had been included in the meta-analysis of a binary outcome. Although outcome reporting bias of the primary outcome of interest was suspected in several individual randomised trials, the impact of such bias on the conclusions drawn in the meta-analyses was minimal. The findings from that study, in terms of the potential for outcome reporting bias to impact on the conclusions of a review and the degree of impact being related to the amount of missing outcome data, were similar to the current study.

A second study of meta-analyses in Cochrane reviews demonstrated a weak positive association between the amount of outcome data missing from the source trial reports and the treatment effect estimate.[16] Our study goes further by reviewing excluded studies and classifying the likelihood of outcome reporting bias in a review on the basis of the individual trial reports.

Implications for systematic reviews

The reliability of systematic reviews can be improved if more attention is paid to outcome data missing from the source trial reports. Trials should not be excluded because there is "no relevant outcome data" as the outcome data may be missing as a direct result of selective outcome reporting. Increasing the accuracy of data extraction, possibly by involving a second reviewer, could reduce the amount of missing data. If a high proportion of data is missing, reviewers should be encouraged to contact the trialists to confirm whether the outcome was measured and analysed and, if so, obtain the results. More than a third of the trialists contacted in this study responded to requests for information, 60% within a day and the remainder within three weeks. Similar response rates were observed with trials published in the past five years compared with those published earlier. In addition, some review authors did not declare when a trial report stated that no events were observed in any group. We believe that reviewers should report all such data in their review.

Review authors will need to use their judgment regarding the potential for outcome reporting bias. Unfortunately, we believe there are few practical alternatives to this approach, since to do nothing is unacceptable and to contact trialists for the information or data is recommended but is not always feasible or successful. To support their judgment, reviewers should justify fully in the text of their report the classification assigned and should include verbatim quotes from the trial publication whenever possible.

The classification system that we used in this study has been presented and applied by participants during workshops that we have developed and delivered at international Cochrane colloquia and the UK Cochrane meetings. The feedback from these workshops has so far not indicated any major shortcomings of this classification system or that any additional categories are required. Adoption of the new Cochrane risk of bias tool,[11] which includes a judgment of the risk of selective outcome reporting, should also help to raise awareness of outcome reporting bias.

If a sensitivity analysis used to assess the impact of outcome reporting bias on an individual review shows that the results are not robust to outcome reporting, the review conclusions may need to be amended. Even if the results appear robust, the reviewer should still consider the potential for bias caused by unpublished studies. An example of this approach is described in a recent tutorial paper (Dwan KM et al, submitted manuscript, 2009).

SUMMARY POINTS

Empirical research indicates that statistically significant outcomes are more likely to be fully reported than non-significant results in published reports of randomised controlled trials

Little is known about the impact of outcome reporting bias in source trial reports on the conclusions of systematic reviews

Few review authors mentioned the potential problem of outcome reporting bias

Outcome reporting bias was suspected in at least one trial in more than a third of reviews

In a sensitivity analysis, nearly a fifth of statistically significant meta-analyses of the review primary outcome were affected by outcome reporting bias and a quarter would have overestimated the treatment effect by 20% or more

Implications for trials

Recent long term initiatives could reduce the problem of outcome reporting bias in trials. For example, registration of randomised controlled trials before initiation[17] and advance publication of detailed protocols document that the trials exist and ensure their planned outcomes are specified. Reviewers can search registries to locate unpublished trials eligible to be included in a systematic review. Trialists should be encouraged to describe all changes to the outcomes stated in the protocol.

The standardisation of outcome measures in specific clinical areas, if implemented, will reduce the potential for bias.[18 19] In our sample, 18% (51/283) of reviews contained at least one trial where it was either clear or suspected that the review primary outcome was not measured (H or I classification). This represents 31 790 (4%) of the 836 689 trial patients studied. There was a missed opportunity to measure a core outcome in these individuals because this study focused on review primary outcomes.

A recent review of trial funders' guidelines[20] has identified gaps in relation to outcome reporting bias. Current recommendations, however, state that all prespecified primary and secondary outcomes should be fully reported; any changes to the prespecified outcomes from the protocol should be explained in the final report; and the choice of outcomes included in the final report should not be based on the results.[20]

The members of the World Health Organization International Clinical Trials Registry Platform working group on the reporting of findings of clinical trials have proposed that "the findings of all clinical trials must be made publicly available."[21] From 2008, the US Food and Drug Administration Act has required that results from clinical trials are made publicly available on the internet in a "registry and results databank" within a year of completion of the trial, whether the results have been published or not.[22] US public law requires "a table of demographic and baseline characteristics of the study participants as well as a table of primary and secondary outcome measures for each arm of the clinical trial, including the results of scientifically appropriate tests of the statistical significance."

We hope that such strategies, coupled with activities to raise awareness of the issues, will reduce the prevalence of outcome reporting bias in clinical research.

Future research

Our study undoubtedly underestimates the influence of outcome reporting bias on this cohort of Cochrane reviews. Review primary outcomes are less likely to be prone to outcome reporting bias than secondary outcomes as primary outcomes are usually chosen on the basis of clinical importance—thus increasing the measurement and reporting of the outcome—or because they are the most frequently reported variables. In an associated study, a sample of trialists identified in this study was interviewed about differences between the trial protocol and the trial report in order to understand outcome reporting bias across all primary and secondary outcomes (Smyth R et al, submitted manuscript, 2009). Future work is planned to assess the prevalence and impact of outcome reporting bias across all outcomes in a cohort of reviews.

In reviews that do not include a meta-analysis, outcome reporting bias may still be operating and may affect the conclusions. Guidance is needed on how to address this problem.

The authors are grateful to the many Cochrane reviewers who collaborated in some way to make this research possible. Their input includes defining the review primary outcome of interest, forwarding trial reports from their reviews, and establishing the outcome reporting bias classification for particular trials within their reviews. We thank the Cochrane Steering Group, and Cochrane statisticians and clinicians from the University of Liverpool who have provided expert knowledge when further assistance was required. We also acknowledge the trialists who answered specific queries about the reporting of outcomes in their trials. Finally, we thank Steve Taylor for undertaking some of the assessments and Chris Braithwaite for designing the study database.

Contributors: PRW, DGA, and CG designed the study protocol and developed the classification system for assessing outcome reporting bias in trials. The study case report form was designed by JJK, SD, and PRW. JJK contacted all review authors to confirm the chosen review primary outcome for use in the study. PRW and SD identified and agreed on the review primary outcome when no contact with the review authors was achieved. Assessments of outcome reporting bias (including obtaining trial reports and completing the outcome reporting bias classifications and justifications with the guidance from the systematic reviewers) were completed by JJK, SD, and KMD. All assessment justifications were checked by PRW. All the data were entered into the database by JJK. JJK and RS contacted trialists to verify if the review primary outcome was measured and/or analysed when outcome reporting bias was suspected. JJK and KMD undertook the data analysis under the supervision of PRW. JJK prepared the initial manuscript. JJK, PRW, DGA, and KMD were all involved in the substantial revision of this manuscript. All authors commented on the final manuscript before submission. PRW is the guarantor for the project.

Funding: The Outcome Reporting Bias In Trials (ORBIT) project was funded by the Medical Research Council (grant number G0500952). The funders had no role in the study design, data collection and analysis, decision to publish, or preparation of this manuscript.

Competing interests: PRW, CG, DGA, KMD, and RS are members of the Cochrane Collaboration. JJK and SD declare no competing interests.

Ethical approval: No ethics committee opinion was required for this study.

Provenance and peer review: Not commissioned; externally peer reviewed.

1 Williamson PR and Gamble C. Identification and impact of outcome selection bias in meta-analysis. *Stat Med* 2005;24:1547-61.
2 Hahn S, Williamson PR, Hutton JL, Garner P, Flynn EV. Assessing the potential for bias in meta-analysis due to selective reporting of subgroup analyses within studies. *Stat Med* 2000;19:3325-36.
3 Moreno SG, Sutton AJ, Turner EH, Abrams KR, Cooper NJ, Palmer TM, et al. Novel methods to deal with publication biases: secondary analysis of antidepressant trials in the FDA trial registry database and related trial publications. *BMJ* 2009;339:b2981.
4 Higgins JPT, Green S, eds. *Cochrane handbook for systematic reviews of interventions: version 5.0.1*. The Cochrane Collaboration, 2008.
5 Hutton JL, Williamson PR. Bias in meta-analysis due to outcome variable selection within studies. *Appl Stat* 2000;49:359-70.
6 Dwan K, Altman DG, Arnaiz JA, Bloom J, Chan A, et al. Systematic review of the empirical evidence of study publication bias and outcome reporting bias. *PloS ONE* 2008;3:e3081.
7 Egger M, Davey Smith G, Altman DG. *Systematic reviews in healthcare: meta-analysis in context*. 2nd ed. BMJ Books, 2001.

8 Shea B, Moher D, Graham I, Pham B, Tugwell P. A comparison of the quality of Cochrane reviews and systematic reviews published in paper-based journals. *Eval Health Prof* 2002;25:116-29.
9 Moher D, Tetzlaff J, Tricco AC, Sampson M, Altman DG. Epidemiology and reporting characteristics of systematic reviews. *PLoS Med* 2007;4:e78.
10 Clarke MJ, Hopewell S, Juszczak E, Eisinga A, Kjeldstrom M. Compression stockings for preventing deep vein thrombosis in airline passengers. *Cochrane Database Syst Rev* 2006;(2):CD004002.
11 Alderson P, Bunn F, Lefebvre C, Li Wan Po A, Li L, Roberts I, et al. Human albumin solution for resuscitation and volume expansion in critically ill patients. *Cochrane Database Syst Rev* 2004;(4):CD001208.
12 Altman DG. Diagnostic tests. In: Altman DG, Machin D, Bryant TN, Gardner MJ, eds. *Statistics with confidence*. 2nd ed. BMJ Books, 2000.
13 Williamson PR, Gamble C. Application and investigation of a bound for outcome reporting bias. *Trials* 2007;8:9.
14 Copas J, Jackson D. A bound for publication bias based on the fraction of unpublished studies. *Biometrics* 2004;60:146-53.
15 Chan A, Altman D. Identifying outcome reporting bias in randomised trials on PubMed: review of publications and survey of authors. *BMJ* 2005;330:753.
16 Furukawa TA, Watanabe N, Omori IM, Montori VM, Guyatt GH. Association between unreported outcomes and effect size estimates in Cochrane meta-analyses. *JAMA* 2007;297:468-70.
17 World Health Organization. *World Health Organization international clinical trials registry platform: unique ID assignment*. WHO, 2005.
18 Sinha I, Jones L, Smyth RL, Williamson PR. A systematic review of studies that aim to determine which outcomes to measure in clinical trials in children. *PLoS Med* 2008;5:96.
19 Clarke M. Standardising outcomes in paediatric clinical trials. *PLoS Med* 2008;5:e120.
20 Dwan K, Gamble C, Williamson PR, Altman DG. Reporting of clinical trials: a review of research funders' guidelines. *Trials* 2008;9:66.
21 Ghersi D, Clarke M, Berlin J, Gulmezoglu M, Kush R, Lumbiganon P, et al. Reporting the findings of clinical trials: a discussion. *Bull World Health Organ* 2008;86:492-3.
22 United States Code (2008) US Public Law 110-85: Food and Drug Administration Amendments Act 2007. http://frwebgate.access.gpo.gov/cgi-bin/getdoc.cgi?dbname=110_cong_public_laws&docid=f:publ085.110.pdf.