ELSEVIER

Journal of Clinical Epidemiology 64 (2011) 1060−1069

Journal of Clinical Epidemiology

Statistically significant meta-analyses of clinical trials have modest credibility and inflated effects

Tiago V. Pereira[a,b], John P.A. Ioannidis[a,c,d,*]

[a]Department of Hygiene and Epidemiology, Clinical Trials and Evidence-Based Medicine Unit, University of Ioannina School of Medicine, Ioannina 45110, Greece

[b]Laboratory of Genetics and Molecular Cardiology, Heart Institute (InCor), São Paulo, 05403-000 Brazil

[c]Stanford Prevention Research Center, Stanford University School of Medicine, Stanford, CA 94305, USA

[d]Institute for Clinical Research and Health Policy Studies, Department of Medicine, Tufts Medical Center, Tufts University School of Medicine, Boston, MA 02111, USA

Accepted 10 December 2010

Abstract

Objective: To assess whether nominally statistically significant effects in meta-analyses of clinical trials are true and whether their magnitude is inflated.

Study Design and Setting: Data from the Cochrane Database of Systematic Reviews 2005 (issue 4) and 2010 (issue 1) were used. We considered meta-analyses with binary outcomes and four or more trials in 2005 with $P < 0.05$ for the random-effects odds ratio (OR). We examined whether any of these meta-analyses had updated counterparts in 2010. We estimated the credibility (true-positive probability) under different prior assumptions and inflation in OR estimates in 2005.

Results: Four hundred sixty-one meta-analyses in 2005 were eligible, and 80 had additional trials included by 2010. The effect sizes (ORs) were smaller in the updating data (2005−2010) than in the respective meta-analyses in 2005 (median 0.85-fold, interquartile range [IQR]: 0.66−1.06), even more prominently for meta-analyses with less than 300 events in 2005 (median 0.67-fold, IQR: 0.54−0.96). Mean credibility of the 461 meta-analyses in 2005 was 63−84% depending on the assumptions made. Credibility estimates changed >20% in 19−31 (24−39%) of the 80 updated meta-analyses.

Conclusions: Most meta-analyses with nominally significant results pertain to truly nonnull effects, but exceptions are not uncommon. The magnitude of observed effects, especially in meta-analyses with limited evidence, is often inflated.

Keywords: Meta-analysis; Bias; Treatment effect; Bayes factor; Winner's curse; Outcomes

1. Introduction

Meta-analyses are often considered as the highest level of evidence for evaluating interventions in health care [1,2] and are very influential in the literature and in practice [3]. However, there has been some debate on whether meta-analyses provide reliable evidence. For example, in an analysis that stirred intense discussion and criticism, LeLorier et al. [4] evaluated 19 meta-analyses and pointed out that these studies had only modest ability to predict the results of subsequent large clinical trials. Meta-analyses with limited evidence, biased studies, and poor-quality trials are considered to be more prone to unreliable results [5−10]. Other investigators have pointed out that the current interpretation of statistically significant results in meta-analyses ignores the fact that studies are added one at a time, thus one needs more conservative rules to claim statistical significance [7,10]. When corrections for sequential testing are made, many statistically significant meta-analyses lose their nominal significance [11].

Based on these concerns, clinicians, patients, and policy makers are left with some uncertainty about how they should interpret a meta-analysis, when they see that it has a P-value < 0.05 and its 95% confidence intervals (CIs) exclude the null. How likely is it that there is some genuine treatment effect rather than a ‘‘false positive’’? Moreover, if

Competing interests: None.

Funding: T.V.P. is funded by grants from the Fundação de Amparo à Pesquisa do Estado de São Paulo (State of São Paulo Research Foundation, FAPESP) and the Coordenação de Aperfeiçoamento Pessoal de Nível Superior (CAPES, Brazilian Ministry of Education).

* Corresponding author. Stanford Prevention Research Center, Stanford University School of Medicine, 251 Campus Drive, Stanford, CA 94305, USA. Tel.: 650-725-5465; fax: 650-725-6247.

E-mail address: jioannid@stanford.edu (J.P.A. Ioannidis).

0895-4356/$ - see front matter © 2011 Elsevier Inc. All rights reserved.
doi: 10.1016/j.jclinepi.2010.12.012

What is new?

- Most statistically significant results from meta-analyses of clinical trials are more likely to reflect truly nonnull effects than false-positive results.
- It is more probable that the credibility of the updated meta-analyses increases rather than decreases.
- Data added to the existing meta-analysis in a 5-year window (2005–2010) indicated less prominent effects than did the summary estimates in 2005.
- The median fold change in these summary estimates was 0.85, but the reduction was greater for meta-analyses with less cumulative data (median reduction of 0.67-fold).

there is some effect, is the statistically significant meta-analysis estimate reliable or inflated—and, if so, by how much? Often clinicians and policy makers use nominal statistical significance as a first prerequisite before even considering an intervention for implementation. Then, they may also ask for a sufficiently large treatment effect size. However, there is evidence from diverse fields that, when one focuses on statistically significant results that pass a given threshold of significance (e.g., $P < 0.05$), some of them are false positives [5] and effect size estimates are inflated on average because of the winner's curse phenomenon [12]. The winner's curse refers to the situation where we select results based on the fact that they cross a threshold of significance and at the same time we try to obtain an effect size estimate. It is then mathematically expected that, on average, these estimates are exaggerated [12]. The extent of inflation of effect sizes varies substantially across different studies and scientific fields and is more prominent when the sample size is smaller [12–14]. False positives and inflation of effects for meta-analyses of clinical trials require more systematic study. Both false positives and inflated effects could cause misleading impressions about an intervention and wrong treatment choices.

Here, we evaluated empirically whether nominally statistically significant results in meta-analyses of clinical trials are credible and the effect sizes from such meta-analyses are potentially inflated. We estimated the credibility (the posterior probability of true-positive results) in independent meta-analyses that had nominal statistical significance in the Cochrane Database of Systematic Reviews (CDSR) in late 2005. Then, we evaluated the change in the credibility of these meta-analyses that had data from additional trials included by early 2010. Moreover, we estimated whether the updating data suggested smaller effects than the initial meta-analyses.

2. Material and methods

2.1. Databases of meta-analyses

We have previously collected data on all 1,011 independent meta-analyses from the CDSR (issue 4, 2005), with binary outcomes and four or more trials [15,16]. Briefly, one meta-analysis has been used per systematic review (the one with the largest number of trials or the largest number of events, if there were two or more with similar number of studies). Further detailed information on selection criteria appears elsewhere [15–17]. In these 1,011 meta-analyses, we summarized results using the odds ratio (OR) as the metric of effect by applying a random-effects model [18] and selected those meta-analyses that had a nominally statistically significant summary effect ($P <$ 0.05). These meta-analyses are referred herein as "Significant meta-analyses—2005." For each of them, we searched for the respective versions of these meta-analyses in the CDSR, issue 1, 2010 and isolated meta-analyses where additional trials had been included. The meta-analyses from 2005 comprise the "Meta-analyses with updating—2005" data set, their updated versions are called the "Updated meta-analyses—2010" data set, and the extra data included in the 2005–2010 window are the "2005–2010 Update" data set. All data were electronically exported from the Cochrane Library to avoid errors in manual data extraction. We focused only on binary outcomes for consistency in the effect metric.

2.2. Meta-analyses calculations for effect sizes

We used both fixed-effects and random-effects models to calculate the summary estimates for meta-analyses. Fixed-effects calculations were carried out using the general inverse-variance method, and random-effects results were obtained by the DerSimonian–Laird method, which incorporates the between-study variance τ^2 estimated using the method of moments [18]. Statistical significance was set at $\alpha = 5\%$. For consistency, all comparisons were coined to yield ORs greater than 1.0 for meta-analyses in the Significant meta-analyses—2005 data set, and the same direction of comparison was maintained for the other data sets. Results refer primarily to random effects, unless stated otherwise, because there is evidence for between-study heterogeneity in the effects of some medical interventions.

When one selects results based on a statistical significance threshold (here $P < 0.05$), it is expected that on average the effect sizes of these results would be inflated compared with the true effects. This is known as the winner's curse phenomenon [12,19]. The inflation is expected to be greater when the evidence is more limited [12,14]. If additional data are obtained for these significant effects, the effect estimates of these additional data should be unaffected by the winner's curse phenomenon. We compared the effect sizes in the "Meta-analyses with updating—2005" data set vs. the "Updated meta-analyses—2010" data set and vs. the

"2005–2010 Update" data set using the Wilcoxon test for paired observations and the Mann-Whitney *U* test for independent samples when appropriate. We also estimated the fold change in OR estimates in the 2005–2010 Update vs. the 2005 estimate and whether this correlated with the weight $(1/\sigma^2(\widehat{\theta}))$, total number of trials, total number of subjects, and number of events of the evidence in 2005.

2.3. Credibility

The credibility (C), the proportion of true positives, represents a posterior probability that an effect exists (it is nonnull, i.e., OR is not 1.00) [6,20]. C can be computed as:

$$C = \frac{R}{R+B}$$

where R is the prior odds of the effect being true (nonnull) and B is the Bayes factor. B is obtained from the evidence provided by the meta-analysis. B-values below 1 mean that the meta-analysis increases the chances that a nonnull effect exists. The smaller the B-value, the larger the increase in the odds that a nonnull effect exists.

There are different approaches to calculate B [21,22]. Here, we used a method that has been proposed previously [6]. In brief, B is calculated by the following formula:

$$B = \sqrt{1+\frac{\pi\theta_A}{\sigma^2(\hat{\theta})}} \times \mathrm{e}\left\{-\left(\frac{\hat{\theta}}{\sigma(\hat{\theta})}\right) \Big/ \left[2\left(1+\frac{\sigma^2(\hat{\theta})}{\pi\theta_{\mathrm{A}}^2}\right)\right]\right\}$$

where $\widehat{\theta}$ and $\sigma^2(\widehat{\theta})$ are the observed summary logarithm of the OR (lnOR) and its variance, respectively, in the meta-analysis, whereas θ_{A} stands for the mean lnOR that is anticipated for true (nonnull) treatment effects. Thus, in this framework, one needs to specify the mean effect anticipated for interventions that do have a nonnull effect.

We have performed analyses considering different magnitudes of anticipated mean effects. One may anticipate that effect sizes for medical interventions vary depending on the type of outcome. For example, large reductions in mortality are difficult to achieve, whereas large effect sizes may be realistic for outcomes that refer to pain control, harms, or laboratory tests. Therefore, we first classified the outcomes of the 461 meta-analyses in the following categories: mortality (including outcomes where death is a composite with other major clinical events), withdrawals (including dropouts and loss to follow up, for any reason), toxicity (harms), pain response, efficacy outcomes that are determined by laboratory tests without any clinical component, and all other efficacy outcomes.

In the main analysis, for each of these outcome categories, we specified a priori θ_{A}-values that correspond to a mean OR of 1.10 for mortality, 2.00 for harms, pain response, and laboratory-determined efficacy, and 1.25 for withdrawals and other efficacy outcomes. This is commensurate with anticipated treatment effects used in previous empirical evaluations of meta-analyses [11] for mortality and withdrawals/efficacy and allowing for substantially larger effects for harms, pain, and laboratory-determined outcomes.

We also performed two sensitivity analyses with different assumptions. In a first sensitivity analysis, we used θ_{A}-values that correspond to the mean lnOR that was observed for each category of outcomes based on the meta-analyses of all data until 2005 (Significant meta-analyses—2005), that is, assuming that the meta-analyses of nominally statistically significant results provide an unbiased estimate of the effect sizes. In a second sensitivity analysis, we used θ_{A}-values that correspond to the mean lnOR that was observed for each category of outcomes based on the 2005–2010 Update data set, that is, assuming that the updating data provide an unbiased estimate of the effect sizes.

We considered two different values of R. First we assumed $R = 0.5$. This means that when an intervention goes into extensive clinical trials testing (at least four trials), it is considered a priori to have a 33% chance (1:2 odds) to have an effect for the outcome of interest. This may be a reasonable assumption in the current environment where clinical trials are performed only for interventions that have a substantial chance of showing some effect, given the high cost of trial research. In a second analysis, we assumed $R = 0.1$, that is, an intervention having 1:10 odds (9% chance) to have an effect a priori, a more pessimistic assumption. The range covered by these two R-values is commensurate with empirical data from cohorts of randomized trials [23–26].

2.4. Software

All meta-analysis calculations were performed in Stata 8.0 (College Station, TX, USA). A Stata module has been written, adapting the B and C calculations discussed above for a meta-analysis context. A Stata module has been written, implementing all calculations discussed below. This program is available at the author's website (http://www.dhe.med.uoi.gr/software.htm). A simple Excel-based calculator can be found at www.dhe.med.uoi.gr/software.htm.

3. Results

3.1. Evaluated meta-analyses

Among the 1,011 meta-analyses, 461 had nominally statistically significant effects in 2005 by random-effects calculations. Of the 461 meta-analyses, 199 belonged to a systematic review that had been updated between 2005 and 2010. Eighty of these 199 meta-analyses included data from additional trials in the 2005–2010 window. Appendix shows the comparisons and outcomes for these 80 meta-analyses, the amount of information available until 2005 and in the 2005–2010 Update (see Appendix on the journal's Web site at www.elsevier.com). Table 1 shows summary characteristics for the different sets of meta-analyses. The meta-analyses that had updates did not differ

Table 1
Summary characteristics of the examined meta-analyses

Characteristics	Significant meta-analyses—2005 ($N = 461$)	Meta-analyses with updating—2005 ($N = 80$)	Meta-analyses without updating—2005 ($N = 381$)	Updated meta-analyses—2010 ($N = 80$)	Update 2005−2010 ($N = 80$)
Trials, median (IQR)	8 (5−14)	12 (7−20)	8 (5−12)	15 (8−28)	2 (1−7)
Participants, median (IQR)	1,234 (630−3,396)	2,202 (836−6,039)	1,131 (597−2,905)	3,158 (1,126−9,216)	577 (190−1,883)
Type of outcome, *n* (%)					
Efficacy—clinical	282 (61.2)	55 (68.8)	227 (59.6)	55 (68.8)	55 (68.8)
Efficacy—laboratory	29 (6.3)	3 (3.8)	26 (6.5)	3 (3.8)	3 (3.8)
Harms	57 (12.4)	7 (8.8)	50 (13.1)	7 (8.8)	7 (8.8)
Pain response	26 (5.6)	5 (6.3)	21 (5.5)	5 (6.3)	5 (6.3)
Withdrawals	39 (8.5)	6 (7.5)	33 (8.7)	6 (7.5)	6 (7.5)
Mortality	28 (6.1)	4 (5)	24 (6.3)	4 (5)	4 (5)
$-\log_{10}$ (*P*-value), median (IQR)					
Random effects	3.20 (2.02−6.14)	3.46 (2.06−7.12)	3.20 (2.01−6.13)	3.74 (2.28−7.63)	0.74 (0.33−3.42)
Fixed effects	4.80 (2.67−9.95)	4.99 (2.65−10.94)	4.73 (2.69−9.62)	5.21 (2.66−10.7)	0.99 (0.35−3.98)
OR, geometric mean (range)[a]					
Random effects	2.73 (1.08−99.9)	2.35 (1.08−11.4)	2.82 (1.10−99.9)	2.21 (1.08−11.4)	2.03 (0.74−10.6)
Fixed effects	2.52 (1.05−99.9)	2.20 (1.08−9.1)	2.59 (1.05−99.9)	2.06 (1.08−9.1)	1.88 (0.74−10.6)

Abbreviations: OR, odds ratio; IQR, interquartile range.

[a] OR estimates were coined to be greater than one for consistency in the meta-analyses in 2005, and the same direction of comparison also was used in the other data sets.

markedly from those without updates in effect sizes in 2005 ($P = 0.054$), *P*-values in 2005 ($P = 0.74$), or proportion of different types of outcomes ($P = 0.74$), but had a substantially higher number of trials ($P = 0.0001$) and participants ($P = 0.0014$) in 2005.

In most of these 461 meta-analyses, fixed-effects estimates were very similar to random effects. However, in 2005, for a total of 47 meta-analyses (10.2%), the random-effects estimate was ≥1.25-fold higher than the fixed-effects estimate, whereas for a total of three meta-analyses (0.65%), the random-effects estimate was ≤0.8-fold the fixed-effects estimate (for effects coined so that random-effects ORs are ≥1.00).

3.2. Effect sizes

For the 80 meta-analyses that had updating information, the point estimates of the effect sizes changed significantly toward lower values between 2005 and 2010 ($P = 0.001$). Likewise, the effect sizes of the updating data were smaller than that of the respective meta-analyses in 2005 ($P = 0.007$). The median change in the OR was 0.85-fold (interquartile range [IQR]: 0.66−1.06) in the 2005−2010 Update vs. the 2005 estimate (Fig. 1). Results were similar with fixed-effects calculations ($P = 0.001$ and $P = 0.008$, respectively; median effect size change 0.85-fold).

The change in the effect size in the 2005−2010 Update vs. the 2005 estimate correlated inversely with the weight $(1/\sigma^2(\hat{\theta}))$ (Spearman's correlation coefficient, $\rho_s = -0.30$, $P = 0.006$), total number of trials ($\rho_s = -0.41$, $P = 0.002$), total number of subjects ($\rho_s = -0.39$, $P = 0.003$), and number of events ($\rho_s = -0.32$, $P = 0.004$) of the evidence in 2005. The median change in the OR was 0.88-fold (IQR: 0.80−1.09) in the 2005−2010 Update vs. the 2005 estimate for the 40 meta-analyses with higher weights and 0.65-fold (IQR: 0.54−0.96) for the 40 meta-analyses with lower weights. The median change in the OR was 0.87-fold (IQR: 0.80−1.08) in the 2005−2010 Update vs. the 2005 estimate for the 40 meta-analyses with higher total number of subjects (>2,202) and 0.68-fold (IQR: 0.55−0.96) for the 40 meta-analyses with lower total number of subjects. Similar results were obtained when meta-analyses were grouped into two with total number of events above and below the median (300 events), respectively: the median

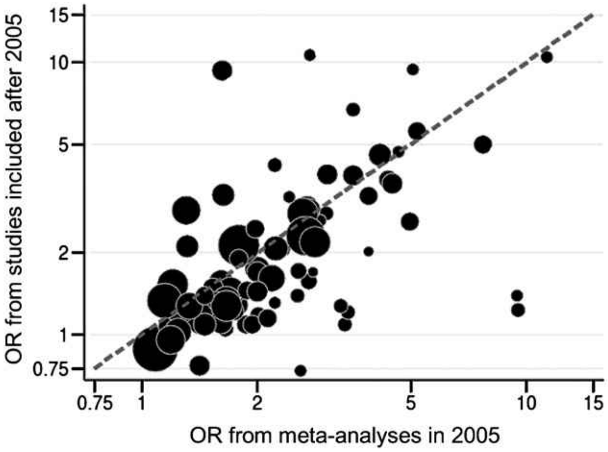

Fig. 1. Summary odds ratio (OR) estimates observed in trials added in 2005−2010 prior vs. ORs observed in 2005 ($N = 80$). Each comparison is represented by a circle, whose area is proportional to the inverse of the standard error (larger areas are given for comparisons with more precision) in 2005. Summary estimates are derived from random-effect calculations (DerSimonian−Laird method). The discontinuous line diagonal corresponds to the points where the OR in 2005 is the same as in the 2005−2010 data.

change in the OR was 0.87-fold (IQR: 0.80–1.08) for the former, whereas this change was 0.67-fold (IQR: 0.54–0.96) for the latter. The decrease in the effect size in the 2005–2010 Update vs. the 2005 estimate, conversely, did not correlate with the amount of evidence added in 2005–2010 (e.g., $\rho_s = -0.02$, $P = 0.86$ and $\rho_s = -0.09$, $P = 0.41$ for the number of participants and trials in the evidence in 2005–2010, respectively).

The observed effects differed significantly across the six types of outcomes ($P = 0.0001$, Table 2) with largest effect sizes seen for harms, pain response, and laboratory efficacy and the smallest effects were seen for mortality. This ranking is consistent with the a priori anticipated average effects for these types of outcomes. However, the magnitude of the effects for each type of outcome in 2005 was larger than our *prior* expectations. The effects seen on the 2005–2010 updates were smaller for all types of outcomes and modestly closer to our a priori expectations (Table 2). Inferences based on a fixed-effects model yielded similar results (not shown).

3.3. Credibility

In the main analysis, the median *B*-values in 2005 suggested a median 24-fold increase in the odds that the effect is true (IQR: 3.57–3,571-fold) (Table 3). In the two sensitivity analyses, the median (IQR) increase was 50-fold (4.35–32,258) assuming the average effects to be those observed in the 2005 meta-analyses and 50-fold (4.2–18,248) assuming the average effects to be those observed in the 2005–2010 Update data.

Assuming a prior odds of effects being true of 0.5, the median credibility of statistically significant effects in 2005 was 0.92 (0.96 in the two sensitivity analyses), whereas the mean was 0.82 (0.84 in both the sensitivity analyses). A total of 425 (92%) effects had more than 50% credibility (409 [89%] and 414 [90%] in the two sensitivity analyses, respectively), and this did not change much by fixed-effects calculations (444 [96%], 426 [92%], and 433 [94%], respectively).

Assuming a prior odds of effects being true of 0.1, the median credibility of statistically significant effects in a meta-analysis in 2005 was 0.71 (0.84 and 0.82 in the two sensitivity analyses, respectively) and the mean was 0.63 (0.68 and 0.67 in the two sensitivity analyses, respectively). A total of 264 (57%) effects had more than 50% credibility (309 [67%] and 303 [66%] in the two sensitivity analyses, respectively). The estimates improved modestly by fixed-effects calculations (351 [76%], 368 [80%], and 363 [79%], respectively).

3.4. Evolution of credibility in 2010 vs. 2005

In the main analysis, the Bayes factor suggested better credibility in 56 of the 80 meta-analyses and worse credibility in the other 24 (including five meta-analyses that lost nominal statistical significance) with the inclusion of the 2005–2010 updates as compared with 2005. Table 4 presents summary estimates and the extent of the heterogeneity found in these five systematic reviews.

Thus, it was significantly more likely for the credibility of the updated meta-analyses to increase rather than decrease ($P = 0.0005$) with accumulation of more data over the period 2005–2010. Assuming different expected effects, in the first sensitivity analysis 51 Bayes factors improved and 29 worsened ($P = 0.019$), whereas in the second sensitivity analysis 53 Bayes factors improved and 27 worsened ($P = 0.0052$).

Fig. 2 shows the credibility estimates in 2010 vs. that in 2005. It can be noticed that the variability in credibility estimates was larger when we assumed $R = 0.1$ rather than $R = 0.5$. Assuming $R = 0.5$, 19 (24%) of the 80 effects changed credibility by >20% in 2010 vs. 2005 (six decreased, 13 increased). Assuming $R = 0.1$, 31 (39%) of the 80 interventions changed credibility by >20% in 2010 vs. 2005 (nine decreased, 22 increased). In sensitivity analyses assuming different expected effects, 15–31 (19–39%) of the 80 effects changed credibility by >20% in 2010 vs. 2005.

4. Discussion

We evaluated 461 meta-analyses of clinical trials on diverse interventions, 80 of which had also been updated over a period of 5 years. We estimated under different assumptions that 63–84% of the 461 meta-analyses

Table 2
Summary effect sizes and between-study heterogeneity (τ^2) for the period 2005–2010 by outcome type

		Meta-analyses with updating—2005		Updated meta-analyses—2010		Update 2005–2010	
Type of outcome	*N* (%)	OR[a]	τ^2	OR[a]	τ^2	OR[a]	τ^2
Efficacy—clinical	55 (68.8)	2.23 (1.15–9.51)	0.08 (0.03–0.35)	2.12 (1.16–7.72)	0.11 (0.03–0.26)	1.89 (0.74–10.6)	0.10 (0.00–0.37)
Efficacy—laboratory	3 (3.8)	2.21 (1.60–3.04)	0.47 (0.01–0.59)	2.06 (1.55–3.06)	0.38 (0.01–0.44)	1.78 (1.11–3.89)	0.19 (0.19–0.19)
Harms	7 (8.8)	4.14 (1.99–11.4)	0.26 (0.02–0.39)	4.21 (1.95–10.3)	0.14 (0.00–0.30)	4.42 (1.70–10.5)	0.17 (0.00–0.18)
Pain response	5 (6.3)	4.68 (1.76–9.50)	0.20 (0.00–0.24)	3.81 (1.61–6.50)	0.19 (0.16–0.30)	2.53 (1.28–5.01)	0.14 (0.08–0.30)
Withdrawals	6 (7.5)	1.60 (1.24–2.12)	0.01 (0.00–0.12)	1.50 (1.27–1.85)	0.01 (0.00–0.13)	1.12 (0.77–1.35)	0.00 (0.00–0.93)
Mortality	4 (5)	1.40 (1.08–1.66)	0.00 (0.00–0.01)	1.24 (1.08–1.54)	0.00 (0.00–0.02)	1.16 (0.89–1.39)	0.05 (0.02–0.16)

Abbreviations: OR, odds ratio; IQR, interquartile range.

[a] OR estimates were coined to be greater than one in the meta-analyses in 2005 and are given as geometric mean (range). $\tau^2$is summarized as median (IQR).

Table 3
Median (IQR) Bayes factor and credibility estimates from the investigated meta-analyses according to different scenarios of alternative effect sizes (θ_A) and prestudy probabilities (R)

Bayes factor and credibility	Significant meta-analyses—2005 ($N = 461$)	Meta-analyses with updating—2005 ($N = 80$)	Meta-analyses without updating—2005 ($N = 381$)	Updated meta-analyses—2010 ($N = 80$)
Bayes factor				
Main analysis	0.04 (2.8×10^{-4}–0.28)	0.03 (3.4×10^{-7}–0.20)	0.04 (3.8×10^{-4}–0.29)	0.01 (2.2×10^{-9}–0.15)
Assuming outcome-specific θ_A based on 2005 estimates	0.02 (3.1×10^{-5}–0.23)	0.009 (8.7×10^{-9}–0.29)	0.02 (5×10^{-5}–0.22)	0.004 (2.2×10^{-9}–0.12)
Assuming outcome-specific θ_A based on 2005–2010 estimates	0.02 (5×10^{-5}–0.24)	0.009 (1.7×10^{-8}–0.22)	0.03 (1×10^{-4}–0.24)	0.004 (2.6×10^{-9}–0.10)
Credibility				
Based on prior odds $R = 0.5$				
Main analysis	0.92 (0.64–1.00)	0.95 (0.71–1.00)	0.92 (0.64–1.00)	0.98 (0.77–1.00)
Assuming outcome-specific θ_A based on 2005 estimates	0.96 (0.69–1.00)	0.98 (0.63–1.00)	0.96 (0.69–1.00)	0.99 (0.80–1.00)
Assuming outcome-specific effects based on 2005–2010 estimates	0.96 (0.68–1.00)	0.98 (0.69–1.00)	0.95 (0.68–0.99)	0.99 (0.83–1.00)
Based on prior odds $R = 0.1$				
Main analysis	0.71 (0.26–1.0)	0.78 (0.33–1.00)	0.70 (0.26–0.99)	0.91 (0.40–1.00)
Assuming outcome-specific θ_A based on 2005 estimates	0.84 (0.31–1.00)	0.91 (0.25–1.00)	0.83 (0.31–1.00)	0.96 (0.45–1.00)
Assuming outcome-specific θ_A based on 2005–2010 estimates	0.82 (0.30–1.00)	0.91 (0.31–1.00)	0.80 (0.30–1.00)	0.96 (0.50–1.00)

Abbreviation: IQR, interquartile range.

probably represent true effects, whereas the remaining 16–37% of the statistically significant meta-analyses are false positives. Moreover, based on the updated sample, the point estimates of the nominally statistically significant effects are, on average, inflated. The inflation is greater for meta-analyses that have more limited data and thus greater uncertainty about their estimates.

The estimated proportion of true-positive meta-analyses suggests that this design, the most highly valued in evidence-based medicine, does detect true effects rather than noise—usually, but false positives are probably not uncommon. The estimated proportion of false positives (16–37%) depends on assumptions that are unavoidably subjective to some extent. We did not consider here significant meta-analyses with only two to three trials. Single trials with statistically significant results may have credibility ranging from <20% (when small, underpowered, biased studies find some nominally statistically significant result) to 95% (when a very large, well-conducted trial finds a significant effect) [5]. Empirical data are also commensurate with these estimates [27–29]. Our findings suggest that meta-analyses are indeed a useful way for improving substantially the

Table 4
Characteristics of the five meta-analyses that lost nominal statistical significance in 2010 compared with 2005

		Summary estimates					
		Fixed effects		Random effects			
Meta-analysis ID	Trials (participants)	OR (95% CI)	P	OR (95% CI)	P	P (Q)	I^2
Evidence in 2005							
CD001865	10 (7,465)	1.32 (1.17–1.50)	0.0001	1.50 (1.11–2.03)	0.008	<0.001	81
CD003288	23 (2,549)	0.69 (0.52–0.90)	0.007	0.69 (0.52–0.90)	0.007	0.736	0
CD003295	4 (276)	0.39 (0.17–0.89)	0.025	0.39 (0.16–0.94)	0.035	0.370	5
CD003388	14 (815)	1.55 (1.02–2.34)	0.038	1.55 (1.02–2.36)	0.038	0.442	1
CD003634	5 (731)	0.71 (0.51–0.98)	0.039	0.71 (0.51–0.98)	0.039	0.675	0
Evidence in 2010							
CD001865	14 (7,341)	1.13 (1.02–1.24)	0.020	1.31 (0.98–1.77)	0.073	<0.001	88
CD003288	28 (7,450)	0.91 (0.78–1.05)	0.18	0.91 (0.78–1.05)	0.18	0.630	0
CD003295	5 (340)	0.48 (0.22–1.02)	0.055	0.49 (0.21–1.16)	0.11	0.331	13
CD003388	15 (862)	1.48 (0.99–2.21)	0.055	1.48 (0.99–2.21)	0.055	0.465	0
CD003634	7 (1,039)	0.79 (0.59–1.06)	0.11	0.79 (0.59–1.06)	0.11	0.518	0

Abbreviations: OR (95% CI), odds ratio (95% confidence intervals); P (Q), P-value for the Cochran's Q-test of homogeneity.

Detailed information on topic/comparison/outcome for each of these systematic reviews is found in the Appendix. Note that for CD001865, the evidence in 2010 includes more trials but the total sample size is less than that in 2005. This is because, compared with 2005, the 2010 version included six more trials, but also excluded two trials (one of which had a large sample size) that the reviewers considered that they were addressing different comparisons/outcomes.

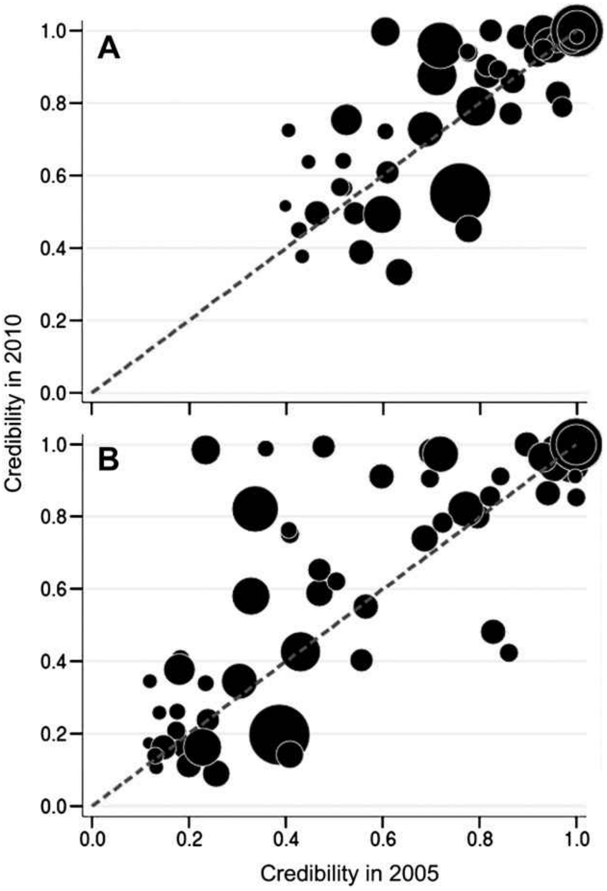

Fig. 2. Credibility estimates obtained by combined data up to 2010 vs. credibility observed in 2005 ($N = 80$). Each comparison is represented by a circle, whose area is proportional to the inverse of the standard error (larger areas are given for comparisons with more precision) in 2005. Panels A and B correspond to random-effects analyses considering R equal to 50% and 10%, respectively. The discontinuous line diagonal corresponds to the points where the credibility in 2010 is the same as in 2005.

credibility of the evidence. On average, the evidence from these meta-analyses changed by 25- to 50-fold the prior odds of an effect being present. However, many clinicians and statisticians may still find a 16–37% false-positive rate alarmingly high.

For clinical practice, detecting a true treatment effect is useful, but decision making also typically requires the magnitude of the treatment effect [30]. The ability to arrive at more accurate treatment effects with reduced uncertainty (tight CIs) is a commonly stated advantage of meta-analysis [31,32]. This is probably true, if one considers all meta-analyses regardless of their results. However, when one focuses on meta-analyses with nominally statistically significant results, the effect sizes are, on average, inflated. The winner's curse phenomenon that has been pinpointed in very diverse disciplines [13,14,24,33–49], including also in single randomized trials, stopped early for perceived effectiveness [11,35,37,49]. The winner's curse is a form of regression to the mean. Its basic principle is that one cannot select and estimate accurately at the same time [12]. On average, when significant meta-analyses find ORs of 2, the true ORs may be around $2 \times 0.85 = 1.7$. As expected from theory [13,14], we observed a greater inflation when the amount of the evidence was more modest. Our analyses suggest that, on average, when significant meta-analyses find ORs of 2, but have fewer than 300 accrued events, the true ORs may be around $2 \times 0.67 = 1.34$. Thus, large statistically significant effects that arise from meta-analyses with limited evidence should be met with caution and get deflated for practice and health policy considerations. These lessons are very much in line with what has been proposed with sequential testing approaches to meta-analyses [7,10,11]. These approaches adjust statistical significance and effect size taking into account the sequential nature of accumulation of trial results and have the advantage of allowing continuous monitoring of the meta-analysis results as new trials are added. However, sequential testing approaches also need to make assumptions about the plausible effect sizes and account for the extent of heterogeneity, but the estimate of heterogeneity usually has substantial uncertainty in the average meta-analysis.

Inflated treatment effects also may be because of biases other than winner's curse. There may be differential biases in early vs. late studies. For example, time lag bias may result in the more promising "positive" results being published earlier than "negative" results on the same intervention [50]. Alternatively, early trials may have more flaws in study design and conduct than later trials. These biases are difficult to decipher and dissociate from each other. Evaluation of time lag bias requires registry information [51] on when each trial was launched. Differential biases in the design and conduct of early vs. late trials also require in-depth knowledge of each field and insider views of the trials. However, typical design features, such as randomization, blinding, and allocation concealment, have, on average, more modest impact on the treatment effects (in the range of $<10\%$) [8] than the inflations that we have documented here. Early trials sometimes may target higher risk patients with more prominent treatment effects, whereas later studies may target wider populations where the benefit is more questionable. Nonetheless, such effect modification based on baseline risk is difficult to document unless very extensive data are available. Finally, statistically significant results from small trials and their meta-analyses that are considered to be too good to be true may sometimes encourage more, larger trials to verify them, and these may turn out to show no effect.

Some limitations should be discussed. Updates between 2005 and 2010 were performed for selected topics. Most meta-analyses did not have updates including more trials in this time window. Regardless of whether new trials were available or not available, the reviewers did not update their review. Both the decision to perform new trials and to update a review may be related to the status of the evidence in 2005. One might expect that debated topics with more uncertainty would be more likely to have additional trials performed and reviews updated. With strong evidence, it is unethical to run more trials. Moreover, reviewers may be inclined to update systematic reviews when they feel that new evidence changes the big picture [52−55]. If so, updated meta-analyses may be prone to show greater changes in the effect size than the average meta-analysis, had more trials and updates been performed on all meta-analyses. Nevertheless, we did not observe any major differences between meta-analyses that had updates and those that did not in terms of types of outcomes, distribution of *P*-values and effect sizes, except for a nonsignificant trend for smaller effect sizes in updated meta-analyses. If anything, the updated meta-analyses had by 2005 more trials and more participants than those that were not updated. Thus it seems that, on average, these topics were simply more popular for clinical experimentation. Moreover, a clinical trial takes on average 5 or more years from its design to its publication [56], thus most of the trials added in the 2005−2010 updates had probably been launched before the 2005 meta-analysis.

We used the CDSR, which is a well-established, inclusive database with wide coverage of diverse medical fields. Cochrane reviews differ from non-Cochrane reviews in several aspects, and on average their quality is better [57,58]. Also, non-Cochrane meta-analyses tend to have more trials and make more conclusive statements [59]. Nonetheless, it is unlikely that the credibility and accuracy of significant effect estimates are better in Cochrane reviews than in the average non-Cochrane meta-analysis with similar amount of evidence. However, exceptions may exist, for example, in meta-analyses sponsored by the industry with spuriously inflated results [60]. Similar evaluations of Bayes factors and credibility would be useful to perform also in non-Cochrane meta-analyses.

We focused on meta-analyses with statistically significant results by random-effects calculations. These comprise slightly less than half of all meta-analyses. One may also ask how nonsignificant meta-analyses should be interpreted. Some of these meta-analyses have statistically significant results based on fixed effects only, whereas others are nonsignificant with either fixed or random effects. In the sample of 1,011 meta-analyses, these two groups included 78 and 472 meta-analyses, respectively, and 48 (62%) of the first group had *P*-values by fixed effects that ranged from 0.01 to 0.05; thus their Bayes factors would not be impressive. Bayes factors can be calculated regardless of what the *P*-value is. Bayes factors for studies with $P > 0.05$ are generally very modest [6,61,62]; that is, the evidence does not improve much the chances that an effect is present compared with prior beliefs. Thus, nonstatistically significant meta-analyses should probably be interpreted based on prior beliefs and also examining the range of the CIs about the remaining uncertainty. Not surprisingly, most systematic reviews conclude that the evidence is inconclusive [58,59].

Occasionally, CIs are so tight that no more trials are indicated, and the topic can be laid to rest, knowing that either no effect exists or, if it exists, it is too small. This is particularly useful for questions of noninferiority [63]. In most nonsignificant meta-analyses, additional trials are warranted. Of note, the winner's curse acts in the opposite direction when meta-analyses are selected based on *not* crossing a threshold of statistical significance. The true effect sizes, on average, should be larger than the observed point estimates of nonsignificant meta-analyses which is another reason why obtaining additional evidence is indicated in most of them.

Bayesian approaches offer advantages to frequentist, *P*-based interpretations of research [61], but the biomedical literature has been entrenched in *P*-values [62]. Commonly listed disadvantages for Bayesian methods are computational difficulty and the need to make assumptions that affect the conclusions [64]. Here we have used a simple method that can be implemented routinely. Other full-Bayesian or false discovery rate methods may be considered, and usually they give comparable inferences [65]. We have tried different assumptions in sensitivity analyses to provide a plausible range for our main inferences. One could also examine in everyday practice for each meta-analysis how inferences change under different assumptions.

Appendix

Supplementary material

Supplementary material can be found, in the online version, at doi:10.1016/j.jclinepi.2010.12.012.

References

[1] Olkin I. Meta-analysis: current issues in research synthesis. Stat Med 1996;15:1253−7.

[2] Lyman GH, Kuderer NM. The strengths and limitations of meta-analyses based on aggregate data. BMC Med Res Methodol 2005;5:14.

[3] Patsopoulos NA, Analatos AA, Ioannidis JP. Relative citation impact of various study designs in the health sciences. JAMA 2005;293:2362−6.

[4] LeLorier J, Gregoire G, Benhaddad A, Lapierre J, Derderian F. Discrepancies between meta-analyses and subsequent large randomized, controlled trials. N Engl J Med 1997;337:536−42.

[5] Ioannidis JP. Why most published research findings are false. PLoS Med 2005;2:e124.

[6] Ioannidis JP. Effect of formal statistical significance on the credibility of observational associations. Am J Epidemiol 2008;168:374−83.

[7] Pogue J, Yusuf S. Overcoming the limitations of current meta-analysis of randomised controlled trials. Lancet 1998;351:47−52.

[8] Wood L, Egger M, Gluud LL, Schulz KF, Juni P, Altman DG, et al. Empirical evidence of bias in treatment effect estimates in controlled trials with different interventions and outcomes: meta-epidemiological study. BMJ 2008;336:601−5.

[9] Kjaergard LL, Villumsen J, Gluud C. Reported methodologic quality and discrepancies between large and small randomized trials in meta-analyses. Ann Intern Med 2001;135:982−9.

[10] Wetterslev J, Thorlund K, Brok J, Gluud C. Trial sequential analysis may establish when firm evidence is reached in cumulative meta-analysis. J Clin Epidemiol 2008;61:64−75.

[11] Thorlund K, Devereaux PJ, Wetterslev J, Guyatt G, Ioannidis JP, Thabane L, et al. Can trial sequential monitoring boundaries reduce spurious inferences from meta-analyses? Int J Epidemiol 2009;38: 276−86.

[12] Ioannidis JP. Why most discovered true associations are inflated. Epidemiology 2008;19:640–8.

[13] Zollner S, Pritchard JK. Overcoming the winner's curse: estimating penetrance parameters from case–control data. Am J Hum Genet 2007;80:605–15.

[14] Pereira TV, Patsopoulos NA, Salanti G, Ioannidis JP. Discovery properties of genome-wide association signals from cumulatively combined data sets. Am J Epidemiol 2009;170:1197–206.

[15] Patsopoulos NA, Evangelou E, Ioannidis JP. Sensitivity of between-study heterogeneity in meta-analysis: proposed metrics and empirical evaluation. Int J Epidemiol 2008;37:1148–57.

[16] Patsopoulos NA, Ioannidis JP. The use of older studies in meta-analyses of medical interventions: a survey. Open Med 2009;3:e62–8.

[17] Ioannidis JP, Patsopoulos NA, Rothstein HR. Reasons or excuses for avoiding meta-analysis in forest plots. BMJ 2008;336:1413–5.

[18] DerSimonian R, Laird N. Meta-analysis in clinical trials. Control Clin Trials 1986;7:177–88.

[19] Young NS, Ioannidis JP, Al-Ubaydli O. Why current publication practices may distort science. PLoS Med 2008;5:e201.

[20] Ioannidis JP. Calibration of credibility of agnostic genome-wide associations. Am J Med Genet B Neuropsychiatr Genet 2008;147B: 964–72.

[21] Spiegelhalter DJ, Abrams KR, Myles JP. Bayesian approaches to clinical trials and health-care evaluation. Chichester, UK: John Wiley & Sons; 2004.

[22] Gelman A, Carlin J, Stern H, Rubin D. Bayesian data analysis. 2nd ed. New York: Chapman & Hall/CRC; 2004.

[23] Djulbegovic B, Kumar A, Soares HP, Hozo I, Bepler G, Clarke M, et al. Treatment success in cancer: new cancer treatment successes identified in phase 3 randomized controlled trials conducted by the National Cancer Institute-sponsored cooperative oncology groups, 1955 to 2006. Arch Intern Med 2008;168:632–42.

[24] Bassler D, Briel M, Montori VM, Lane M, Glasziou P, Zhou Q, et al. Stopping randomized trials early for benefit and estimation of treatment effects: systematic review and meta-regression analysis. JAMA 2010;303:1180–7.

[25] Kumar A, Soares H, Wells R, Clarke M, Hozo I, Bleyer A, et al. Are experimental treatments for cancer in children superior to established treatments? Observational study of randomised controlled trials by the Children's Oncology Group. BMJ 2005;331:1295.

[26] Soares HP, Kumar A, Daniels S, Swann S, Cantor A, Hozo I, et al. Evaluation of new treatments in radiation oncology: are they better than standard treatments? JAMA 2005;293:970–8.

[27] Ioannidis JP. Contradicted and initially stronger effects in highly cited clinical research. JAMA 2005;294:218–28.

[28] Ioannidis JP, Lau J. The impact of high-risk patients on the results of clinical trials. J Clin Epidemiol 1997;50:1089–98.

[29] Trikalinos TA, Churchill R, Ferri M, Leucht S, Tuunainen A, Wahlbeck K, et al. Effect sizes in cumulative meta-analyses of mental health randomized trials evolved over time. J Clin Epidemiol 2004;57: 1124–30.

[30] Guyatt G, Rennie D. The users' guides to the medical literature: a manual for evidence-based clinical practice. 2nd ed. New York, NY: McGraw-Hill; 2008.

[31] Sacks HS, Berrier J, Reitman D, Ancona-Berk VA, Chalmers TC. Meta-analyses of randomized controlled trials. N Engl J Med 1987;316:450–5.

[32] Lau J, Ioannidis JP, Schmid CH. Quantitative synthesis in systematic reviews. Ann Intern Med 1997;127:820–6.

[33] Jeffries NO. Ranking bias in association studies. Hum Hered 2009;67:267–75.

[34] Montori VM, Devereaux PJ, Adhikari NK, Burns KE, Eggert CH, Briel M, et al. Randomized trials stopped early for benefit: a systematic review. JAMA 2005;294:2203–9.

[35] Gehr BT, Weiss C, Porzsolt F. The fading of reported effectiveness. A meta-analysis of randomised controlled trials. BMC Med Res Methodol 2006;6:25.

[36] Krum H, Tonkin A. Why do phase III trials of promising heart failure drugs often fail? The contribution of "regression to the truth". J Card Fail 2003;9:364–7.

[37] Pocock SJ, Hughes MD. Practical problems in interim analyses, with particular regard to estimation. Control Clin Trials 1989;10: 209S–21S.

[38] Bagshaw SM, McAlister FA, Manns BJ, Ghali WA. Acetylcysteine in the prevention of contrast-induced nephropathy: a case study of the pitfalls in the evolution of evidence. Arch Intern Med 2006;166: 161–6.

[39] Goring HH, Terwilliger JD, Blangero J. Large upward bias in estimation of locus-specific effects from genomewide scans. Am J Hum Genet 2001;69:1357–69.

[40] Allison DB, Fernandez JR, Heo M, Zhu S, Etzel C, Beasley TM, et al. Bias in estimates of quantitative-trait-locus effect in genome scans: demonstration of the phenomenon and a method-of-moments procedure for reducing bias. Am J Hum Genet 2002;70:575–85.

[41] Siegmund D. Upward bias in estimation of genetic effects. Am J Hum Genet 2002;71:1183–8.

[42] Beavis WD. QTL analysis: power, precision, and accuracy. In: Paterson AH, editor. Molecular dissection of complex traits. Boca Raton, FL: CRC Press; 1998. p. 145–73.

[43] Garner C. Upward bias in odds ratio estimates from genome-wide association studies. Genet Epidemiol 2007;31:288–95.

[44] Jennions MD, Moller AP. Relationships fade with time: a meta-analysis of temporal trends in publication in ecology and evolution. Proc Biol Sci 2002;269:43–8.

[45] Leimu R, Koricheva J. Cumulative meta-analysis: a new tool for detection of temporal trends and publication bias in ecology. Proc Biol Sci 2004;271:1961–6.

[46] Steyerberg EW, Eijkemans MJ, Harrell FE Jr, Habbema JD. Prognostic modelling with logistic regression analysis: a comparison of selection and estimation methods in small data sets. Stat Med 2000;19: 1059–79.

[47] Steyerberg EW, Harrell FE Jr, Borsboom GJ, Eijkemans MJ, Vergouwe Y, Habbema JD. Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. J Clin Epidemiol 2001;54:774–81.

[48] Steyerberg EW, Eijkemans MJ, Habbema JD. Stepwise selection in small data sets: a simulation study of bias in logistic regression analysis. J Clin Epidemiol 1999;52:935–42.

[49] Simon R, Altman DG. Statistical aspects of prognostic factor studies in oncology. Br J Cancer 1994;69:979–85.

[50] Hopewell S, Clarke M, Stewart L, Tierney J. Time to publication for results of clinical trials. Cochrane Database Syst Rev 2007; MR000011.

[51] Rennie D. Trial registration: a great idea switches from ignored to irresistible. JAMA 2004;292:1359–62.

[52] Moher D, Cook DJ, Eastwood S, Olkin I, Rennie D, Stroup DF. Improving the quality of reports of meta-analyses of randomised controlled trials: the QUOROM statement. Quality of Reporting of Meta-analyses. Lancet 1999;354:1896–900.

[53] Moher D, Tsertsvadze A. Systematic reviews: when is an update an update? Lancet 2006;367:881–3.

[54] Moher D, Tsertsvadze A, Tricco AC, Eccles M, Grimshaw J, Sampson M, et al. A systematic review identified few methods and strategies describing when and how to update systematic reviews. J Clin Epidemiol 2007;60:1095–104.

[55] Moher D, Tsertsvadze A, Tricco AC, Eccles M, Grimshaw J, Sampson M, et al. When and how to update systematic reviews. Cochrane Database Syst Rev 2008; MR000023.

[56] Ioannidis JP. Effect of the statistical significance of results on the time to completion and publication of randomized efficacy trials. JAMA 1998;279:281–6.

[57] Moseley AM, Elkins MR, Herbert RD, Maher CG, Sherrington C. Cochrane reviews used more rigorous methods than non-Cochrane reviews: survey of systematic reviews in physiotherapy. J Clin Epidemiol 2009;62:1021–30.

[58] Moher D, Tetzlaff J, Tricco AC, Sampson M, Altman DG. Epidemiology and reporting characteristics of systematic reviews. PLoS Med 2007;4:e78.

[59] Tricco AC, Tetzlaff J, Pham B, Brehaut J, Moher D. Non-Cochrane vs. Cochrane reviews were twice as likely to have positive conclusion statements: cross-sectional study. J Clin Epidemiol 2009;62:380–6.

[60] Jorgensen AW, Hilden J, Gotzsche PC. Cochrane reviews compared with industry supported meta-analyses and other meta-analyses of the same drugs: systematic review. BMJ 2006;333:782.

[61] Goodman SN. Toward evidence-based medical statistics. 2: the Bayes factor. Ann Intern Med 1999;130:1005–13.

[62] Goodman SN. Toward evidence-based medical statistics. 1: the P value fallacy. Ann Intern Med 1999;130:995–1004.

[63] Piaggio G, Elbourne DR, Altman DG, Pocock SJ, Evans SJ. Reporting of noninferiority and equivalence randomized trials: an extension of the CONSORT statement. JAMA 2006;295:1152–60.

[64] Johnson SR, Tomlinson GA, Hawker GA, Granton JT, Feldman BM. Methods to elicit beliefs for Bayesian priors: a systematic review. J Clin Epidemiol 2010;63:355–69.

[65] Katki H. Invited commentary: evidence-based evaluation of p-values and Bayes factors. Am J Epidemiol 2008;168:384–8.