European Heart Journal (2014) **35**, 3336–3345
doi:10.1093/eurheartj/ehu424

REVIEW

*Statistical tutorials*

# Systematic reviews and meta-analyses of randomized trials: principles and pitfalls

**Bruno R. da Costa[1,2,3] and Peter Jüni[1,3]***

[1]Institute of Social and Preventive Medicine, University of Bern, Finkenhubelweg 11, 3012 Bern, Switzerland; [2]Department of Physical Therapy, Florida International University, Miami, USA; and [3]Institute of Family Medicine (BIHAM), University of Bern, Bern, Switzerland

*Received 8 November 2013; revised 14 September 2014; accepted 6 October 2014; online publish-ahead-of-print 21 November 2014*

Systematic reviews and meta-analyses allow for a more transparent and objective appraisal of the evidence. They may decrease the number of false-negative results and prevent delays in the introduction of effective interventions into clinical practice. However, as for any other tool, their misuse can result in severely misleading results. In this article, we discuss the main steps that should be taken when conducting systematic reviews and meta-analyses, namely the preparation of a review protocol, identification of eligible trials, and data extraction, pooling of treatment effects across trials, investigation of potential reasons for differences in treatment effects across trials, and complete reporting of the review methods and findings. We also discuss common pitfalls that should be avoided, including the use of quality assessment tools to derive summary quality scores, pooling of data across trials as if they belonged to a single large trial, and inappropriate uses of meta-regression that could result in misleading estimates of treatment effects because of regression to the mean or the ecological fallacy. If conducted and reported properly, systematic reviews and meta-analyses will increase our understanding of the strengths and weaknesses of the available evidence, which may eventually facilitate clinical decision making.

**Keywords** Systematic review • Meta-analysis • Research synthesis • Random effects • Fixed effect • Heterogeneity

## Introduction

With the ever growing accumulation of evidence (*Figure 1*), it is impossible to ongoingly identify, summarize, and properly interpret the available evidence to support clinical decision making. Unsurprisingly, the number of systematic reviews and meta-analyses as a means of providing a clinically useful and reliable synthesis of the evidence shows a near exponential growth since the beginning of the 1990s (*Figure 1*). Although few would question the general usefulness of systematic reviews and meta-analyses, inappropriate conduct may produce misleading results. In principle, systematic reviews are studies of studies, with explicit methods to identify, select, critically appraise, and summarize the results of all studies that are relevant to a clearly defined question. The definition of a meta-analysis is much narrower, referring to the statistical methods that are used to combine the results from different studies.[1] The objective of this article is to present the main steps which should be taken (see Box 1 for summary), and to discuss common pitfalls that should be avoided (see Box 2 for summary), when conducting systematic reviews and meta-analyses. We will give focus to systematic reviews and meta-analyses of randomized clinical trials (RCTs). However, most principles discussed in this article also apply to systematic reviews and meta-analyses of other study types.

## Review protocol

Just like for any other study type, a study protocol is considered essential when conducting a systematic review. Reviewers should fully report in a protocol the eligibility criteria, outcomes of interest, and a strategy for data analysis, much in analogy to the protocol of a randomized trial.[2,3] However, it is important to stress that the conduct of a systematic review is an iterative process and reviewers may need to adapt their protocol accordingly. Such modifications are considered acceptable, as long as the reviewers document them and provide a clear rationale for them.[3] Such documentation helps reviewers to keep track of important decisions when conducting or updating reviews. A growing appeal has been seen in recent years for the prospective registration of review protocols, and we encourage reviewers to do so.[3–5] Such prospective registration may avoid unplanned duplication of systematic reviews, and minimize

* Corresponding author. Tel: +41 (0)31 631 57 93; Fax: +41 (0)31 631 35 20, Email: juni@ispm.unibe.ch

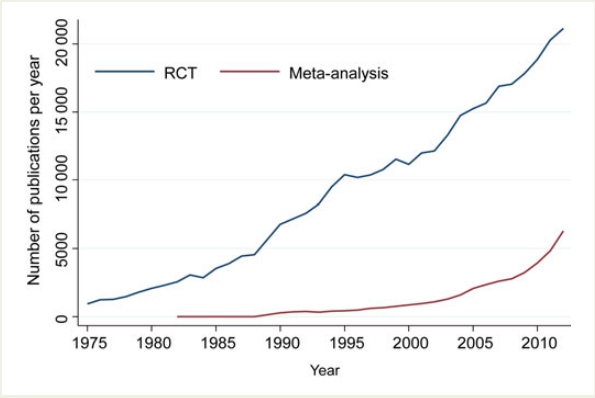


**Figure 1** Annual number of publications of randomized clinical trials and meta-analyses indexed in PubMed. RCT: randomized clinical trial.

**Box 1 Things to do in systematic reviews and meta-analysis**

- Write-up a review protocol
- Do trial selection and data extraction in duplicate and independently by two or more reviewers
- Assess the methodological quality of trials included in the systematic review
- Use appropriate methods to pool effect estimates from different trials to preserve within-trial comparisons
- Estimate statistical heterogeneity
- Use a forest plot to display results
- Conduct stratified analyses to investigate whether treatment effect estimates depend on specific trial characteristics
- Build funnel plots and conduct asymmetry tests to investigate small-study effects
- Write-up the manuscript following recommendations of the PRISMA statement

**Box 2 Things not to do in systematic reviews and meta-analysis**

- Do not use quality assessment tools to derive summary quality scores
- Do not use tests of heterogeneity to decide whether fixed- or random-effects models should be used in analysis
- Do not simply sum up across trials the number of events and the number of patients within experimental and control groups as if they belonged to a single large trial
- Do not pool risk differences without a strong rationale
- Do not use meta-regression to investigate the association between baseline risk and treatment effect
- Do not investigate the association between treatment effect and patient characteristics aggregated at trial level, such as mean age or the percentage of females, in meta-regression

biases, for instance, due to selective reporting of outcomes, data dredging, or non-publication. Different opportunities that allow registration of protocols of systematic reviews, such as PROSPERO, are currently available or underway.[3–5]

# Literature search

Reviewers should aim at identifying as many eligible trials as possible. Searching a single electronic database will not be enough in most cases.[6] For a systematic review of RCTs, reviewers should at a minimum conduct their search in Medline, Embase, and the Cochrane Collaboration's Central Register of Controlled Trials. Subject-specific databases could also be used to increase the sensitivity of the search.[7] Whenever possible, reviewers should also use alternative methods, including screening the reference list of eligible trials, searching clinical trial registers, searching conference proceedings, and contacting experts in the field. Chapter 6 of the Cochrane Handbook for Systematic Reviews of Interventions (freely available on http://handbook.cochrane.org/) provides guidance on how to design and conduct a proper literature search.

Search strategies for the identification of clinical studies of interventions commonly address three concepts: study design, interventions, and patient populations. Box 3 shows an example of a search strategy used to identify RCTs (study design, lines 1–20) comparing early generation drug-eluting stents with bare-metal stents (intervention, lines 21–28) in patients with ST-segment elevation myocardial infarction (population, lines 29–36).[8] Developing a search strategy is iterative in nature. Good starting points for a search are the controlled indexing terms found in the thesaurus of major databases, such as the Medical Subject Headings tree in Medline and the Emtree in Embase, complemented by terms used for indexing articles already known to be eligible. Controlled terms will usually be complemented by free text words, which can be identified by screening title, abstract, keywords, and main body of text of already identified articles and by using lists of synonyms provided by major databases. As a rule of thumb based on the authors' own experience, the proportion of trials included in the review should be roughly 1–5% of all references screened for inclusion. For example, in a recent systematic review on drug-eluting vs. bare-metal stents in patients with ST-elevation myocardial infarction, we expected ~10–15 trials and therefore anticipated to screen 300–600 references.[8]

In the first step, reviewers will typically screen titles and abstracts to exclude only clearly ineligible references. In the second step, the full text of remaining references will be examined to determine eligibility. To minimize bias and error, two reviewers should independently screen all references in duplicate, with disagreements resolved by discussion or by having a third reviewer making the final decision. If restricted resources do not allow for reference screening in duplicate, one reviewer may screen all the references, while the other reviewer screens a random sample.

# Data extraction

Data extraction should always be performed in duplicate, with disagreements resolved by consensus or involvement of the third reviewer. To minimize potential errors, data should be extracted on a standardized and piloted extraction form accompanied by clear instructions on how each of the variables should be extracted.

# Quality assessment

If the 'raw material' is flawed then the conclusions of systematic reviews and meta-analyses cannot be trusted. Therefore, the

**Box 3**

**Medline**

| Step | Search strategy | |
|---|---|---|
| 1 | randomized controlled trial.pt. | Search terms for 'study design' |
| 2 | controlled clinical trial.pt. | |
| 3 | randomized controlled trial.sh. | |
| 4 | random allocation.sh. | |
| 5 | double blind method.sh. | |
| 6 | single blind method.sh. | |
| 7 | clinical trial.pt. | |
| 8 | exp clinical trial/ | |
| 9 | (clin$ adj25 trial$).ti,ab. | |
| 10 | ((singl$ OR doubl$ OR trebl$ OR tripl$) adj25 (blind$ OR mask$)).ti,ab. | |
| 11 | placebos.sh. | |
| 12 | placebo$.ti,ab. | |
| 13 | random$.ti,ab. | |
| 14 | research design.sh. | |
| 15 | comparative study.sh. | |
| 16 | exp evaluation studies/ | |
| 17 | follow up studies.sh. | |
| 18 | prospective studies.sh. | |
| 19 | (control$ OR prospectiv$ OR volunteer$).ti,ab. | |
| 20 | OR/1-19 | |
| 21 | (sirolimus OR rapamycin OR I-2190A OR I 2190A OR I2190A OR AY 22-989 OR AY 22 989 OR AY 22989 OR rapamune OR SES OR DES).tw. | Search terms for 'intervention' |
| 22 | (paclitaxel OR anzatax OR NSC-125973 OR NSC 125973 OR NSC125973 OR taxol OR taxol A OR paxene OR praxel OR 7-epi-Taxol OR 7 epi Taxol OR onxol OR PES OR DES).tw | |
| 23 | (bare-metal OR bare metal OR BMS).tw. | |
| 24 | 21 AND 23 | |
| 25 | 22 AND 23 | |
| 26 | 24 OR 25 | |
| 27 | exp drug-eluting stents/ | |
| 28 | 26 AND 27 | |
| 29 | exp acute coronary syndrome/ | Search terms for 'patient population' |
| 30 | exp myocardial Infarction/ | |
| 31 | (myocardial adj2 infarct*).tw. | |
| 32 | (st-segment OR st segment OR st-elevation OR st elevation).tw. | |
| 33 | (STEMI OR AMI).tw. | |
| 34 | Primary angioplasty.tw. | |
| 35 | Primary percutaneous coronary intervention.tw. | |
| 36 | OR/29-35 | |
| 37 | 20 AND 28 AND 36 | |

Search strategy adapted from Kalesan *et al.* (8). pt, publication type; sh, subject heading; ti, title; ab, abstract; tw, text word; exp, explode.

methodological quality of eligible trials should always be assessed. There is evidence indicating that inappropriate concealment of allocation, lack of blinding of patient, therapists, or outcome assessors, or analysis not according to the intention to treat principle may bias trial results.[9–12] Box 4 provides definitions of components

**Box 4 Items for methodological assessment**

**Generation of allocation sequences**
Adequate in preventing selection bias if sequences are unpredictable: random numbers generated by computer, table of random numbers, drawing of lots or envelopes, tossing a coin, shuffling cards, throwing dice, etc.

**Concealment of allocation sequences**
Adequate in preventing selection bias if patients and investigators enrolling patients cannot foresee assignment: a priori numbered or coded drug containers of identical appearance prepared by an independent pharmacy; central randomization (performed at a remote site); sequentially numbered, sealed, opaque envelopes; etc.

**Blind adjudication of events**
Adequate in preventing detection bias if the adjudication of events used in the analysis is performed by an independent external clinical events committee that is not aware of which treatment patients were allocated to. Blind adjudication of events is not necessary for overall mortality as an outcome.

**Intention to treat analysis**
Adequate in preventing attrition bias if all patients randomized are analysed in the group they were originally allocated to. In time-to-event analyses, up to 10% loss to follow-up may be acceptable, provided that the percentage of patients lost to follow-up is similar between groups, and all randomized patients are initially included in the analysis and only censored at the time they were lost to follow-up.

of methodological quality most relevant for cardiovascular trials. Summary scores derived from quality assessment scales, such as the frequently used Jadad scale,[13] should not be used, neither as a criterion for inclusion in the meta-analysis nor for stratifying analyses, since results may depend considerably on the scale used to assess quality of trials.[14,15] In a study of 25 different scales used to assess 17 trials of low-molecular-weight heparin vs. standard heparin for thromboprophylaxis, with some scales the relative risks of high-quality trials were close to one and not statistically significant, indicating that low-molecular-weight heparins were not superior to standard heparin, whereas low-quality trials showed significantly better protection with low-molecular-weight heparins. With other scales the opposite was the case: high-quality trials suggested that low-molecular-weight heparin were significantly superior to standard heparin, whereas low-quality trials found no significant difference.[14] In addition, potentially important associations between components of methodological quality and estimates of treatment effects might be missed if associations cancel each other out because of opposite directions, or if they are diluted due to a large number of irrelevant components assessed.[14,16,17] Rather, analyses should be stratified by individual components of methodological quality, such as concealment of allocation (see Stratified analyses and meta-regression). Meta-analyses should always be interpreted in the light of the methodological quality of included trials and the results of analyses stratified by components of methodological quality. The frequently recommended Cochrane Risk of Bias tool includes the most important components of methodological quality that should be addressed.[18]

## Fixed- and random-effects models

Meta-analysis is simply a weighted average of estimates from different trials. It makes intuitive sense that small trials estimate treatment

effects less precisely than large trials. Therefore, statistical weights used in a meta-analysis take into account the statistical precision of each trial and give more weight to larger trials. Fixed-effect models assume that there is only one common treatment effect, which is estimated by each of the trials in the meta-analysis. The only source of variation to take into account under this assumption is the statistical imprecision of estimates of treatment effects from individual trials and the weights assigned to each trial correspond to the inverse of the variance of these estimates. Therefore, for trial *i*, the assigned weight will be $\omega = (1/\text{var}_i)$, with $\text{var}_i = \text{se}_i^2$, where $\text{var}_i$ equals the observed within-trial variance and $\text{se}_i$ the standard error of the estimated treatment effect in trial *i*. Random-effects models do not assume that there is one common treatment effect, but rather a series of different treatment effects, and each of the included trials may estimate a different treatment effect. Accordingly, two sources of variation need to be taken into account under this assumption: the already discussed statistical imprecision of individual trials as expressed by the variance within trials and the variance between trials, typically referred to as $\tau^2$ (see Statistical heterogeneity between trials). For trial *i*, the assigned weight will then be $\omega = (1/(var_i + \tau^2))$. The less variation between trials, the lower the between-trial variance $\tau^2$, and the closer pooled estimate and corresponding confidence interval from random-effects models will correspond to those from fixed-effect models. In the extreme case of no variance between trials over and above of what would be expected by chance, $\tau^2$ will be 0 and results from random- and fixed-effect models will be identical. Conversely, differences in pooled estimates and widths of confidence intervals will increase as between-trial variance $\tau^2$ increases. A common misconception is that random-effects models will always derive more conservative estimates than fixed-effect models. Whenever $\tau^2$ is different than null, random-effects model will indeed derive more conservative (i.e. wider) confidence intervals. However, since weights are more similar across trials of different sizes in random effects when compared with fixed-effect models,[19] the pooled estimates from random-effects models are more affected by small-study effects, defined as biases due to publication bias or other methodological problems commonly associated to small studies (see Funnel plots).[19] Accordingly, pooled estimates from fixed-effect models will be more conservative, i.e. closer to the line of no difference, in the presence of small-study effects. *Figure 2* presents an extreme example of a random-effects meta-analysis comparing intravenous magnesium with placebo in patients with acute myocardial infarction, which indicated a large effect of magnesium on overall mortality (relative risk 0.53, 95% confidence interval 0.38–0.75, left) despite the inclusion of the null result found in the mega-trial ISIS-4, which included more than four times as many patients as all previous trials combined.[20] These results could be entirely explained by large benefits erroneously found in small trials in the presence of moderate-to-large heterogeneity (see Statistical heterogeneity). An inspection of the size of the squares and the quantification of statistical weights indicates that the small and moderately sized trials all received unduly large weights, while the weight of ISIS-4 was a mere 17.7%. Conversely, a fixed-effect model, which gave very little weight to the small and moderately sized trials, but 92.8% to ISIS-4, yielded a clear-cut null-result (relative risk 1.01, 95% confidence interval 0.95–1.06).[21] This example shows that reviewers should not mechanistically decide to give preference to a random-effects model if moderate-to-large

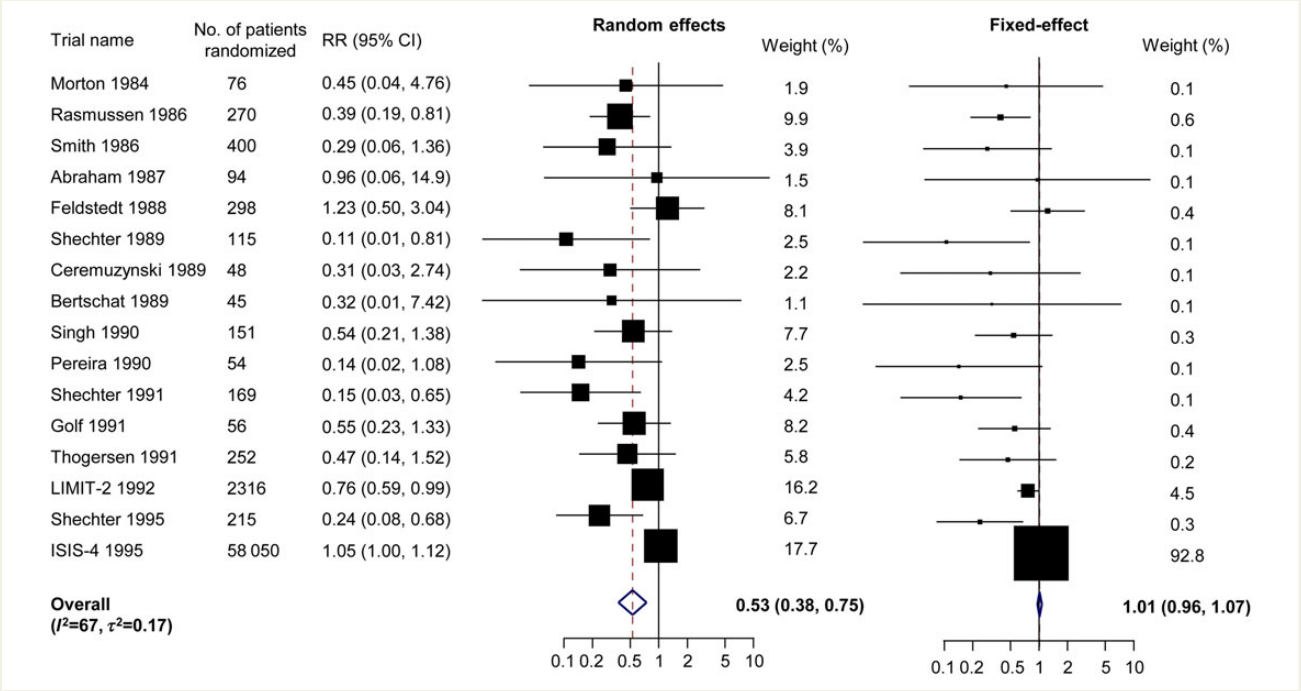


**Figure 2** Random- and fixed-effect meta-analyses comparing the effect of intravenous magnesium with placebo on overall mortality in patients with acute myocardial infarction. RR: risk ratio; CI: confidence interval. A risk ratio below 1 indicates that intravenous magnesium is better than placebo.

statistical heterogeneity is found, as is unfortunately the case in many meta-analyses. Such an approach can yield completely misleading results, particularly if reviewers do not carefully explore sources of variation between trials in estimates of treatment effects (see Stratified analyses and funnel plots). Rather, reviewers should decide a priori which model to use in view of the considerations above, with the understanding that random-effects models will only be truly more conservative than fixed-effect models if statistical heterogeneity is present and small-study effects absent. On a related note, different methods used to conduct random-effects meta-analysis may yield pooled estimates of different magnitude and precision. The DerSimonian & Laird estimator,[22] the most commonly used method for conducting random-effects meta-analysis, does not take into consideration the uncertainty around $\tau^2$ estimation and may yield biased estimates with spuriously high precision. A recently published article discusses other approaches that could be used instead.[23]

## Pooling of binary data

Binary outcomes, as frequently reported in cardiology, should be expressed as estimates of the relative risk, such as risk ratios, rate ratios, hazard ratios, or odds ratios. To be amenable for pooling, these estimates need to be log-transformed using the natural logarithm so that their behaviour is additive and approximately follows a normal distribution. An unfortunate error easily made when using current software packages is to use untransformed estimates for meta-analysis, for example, the risk ratio rather than the natural logarithm of the risk ratio, in combination with appropriately calculated standard errors. Reviewers should be careful with combining risk differences in meta-analysis, since these are sensitive to variations of the baseline risk.[24,25] In most clinical situations, it is reasonable, for example, to assume that an intervention, which approximately halves the risk of myocardial infarction in a trial, which included a high-risk population with an annual baseline risk of myocardial infarction of 10% in the control group, also halves the risk in a trial, which included an average-risk population with a baseline risk of 1%. In both trials, the relative risk will be ~0.5. Conversely, the risk difference will be 5% in the first, but only 0.5% in the second trial. Obviously, pooling risk differences of these two trials will introduce statistical heterogeneity, whereas pooling relative risks will not. Accordingly, numbers needed to treat, or numbers needed to harm, cannot be calculated directly in a meta-analysis, but need to be derived indirectly by applying the pooled relative risk reduction found in the meta-analysis to the baseline risk relevant to specific groups of patients.[24]

## Example of inappropriate method for pooling of trials

It may be tempting for some reviewers to simply sum up across trials the number of events and the number of patients within experimental and control groups as if they belonged to a single large trial, and thereafter calculate a treatment effect estimate with 95% confidence interval. Results of such an exercise will only be approximately correct and correspond to results from a fixed-effect meta-analysis if all included trials used 1 : 1 randomization so that the numbers of patients allocated to experimental and control group in each trial were near identical. We advise against using this approach, since it can yield seriously misleading results if some of the included trials had unequal group sizes due to randomization ratios other than 1 : 1.[26,27] Reviewers should instead first calculate treatment effect estimates and respective standard errors for each single trial and subsequently conduct a meta-analysis to derive an overall treatment effect estimate and its 95% confidence interval as described above.

# Meta-analysis using individual patient data

Individual patient data (IPD) meta-analysis refers to the meta-analysis of raw data of each individual patient from all trials included in the review. The observations in a dataset used in an IPD meta-analysis consists of individuals (or patients), whereas the dataset used in an aggregate level meta-analysis consists of averaged treatment effect estimates, such as odds ratios or differences in means. One of the main advantages is the possibility to stratify analyses according to patient characteristics, which cannot be properly done on an aggregate level because of the ecological fallacy (see Stratified analyses and meta-regression).[28] A detailed discussion of IPD meta-analysis is beyond the scope of the present tutorial; further information can be found elsewhere.[29,30]

# Statistical heterogeneity

Statistical heterogeneity is defined as the variation of treatment effect estimates between trials over and above of the variation expected by chance alone. Heterogeneity will occur if there are characteristics of patients, co-interventions or trials that act as effect modifiers and influence treatment effects measured on a relative risk scale (see Pooling of trials), and if these characteristics are unequally distributed across trials. The larger the degree of unexplained heterogeneity, the less confidence reviewers and readers should have into a meta-analysis, irrespective of the model used (see Stratified analyses, meta-regression, and funnel plots for ways of exploring sources of heterogeneity).[31]

The *Q*-statistic traditionally used to quantify heterogeneity is difficult to interpret as it depends on the number of trials included in the meta-analysis and on the precision of these trials.[32] The currently most frequently used metrics are $I^2$ and $\tau^2$. $I^2$ ranges from 0 to 100% and quantifies the percentage of the variation of treatment effect estimates between trials due to heterogeneity rather than the play of chance. An $I^2$ of 40% indicates, for example, that 40% of the observed variation between estimated treatment effects is due to real heterogeneity, while 60% is due to chance. $I^2$ should be interpreted with care since it will be influenced by the precision of the trials included in the meta-analysis: as the precision of trials increases so does the $I^2$, irrespective of variation in estimates of treatment effects.[32] $\tau^2$ is an estimate of between-trial variance measured on the same scale as the within-trial variance $\text{var}_i = \text{se}_i^2$ referred to above. Its interpretation will therefore depend on the type of estimate used. *Figure 3* presents guidance for the interpretation of $I^2$ and $\tau^2$.

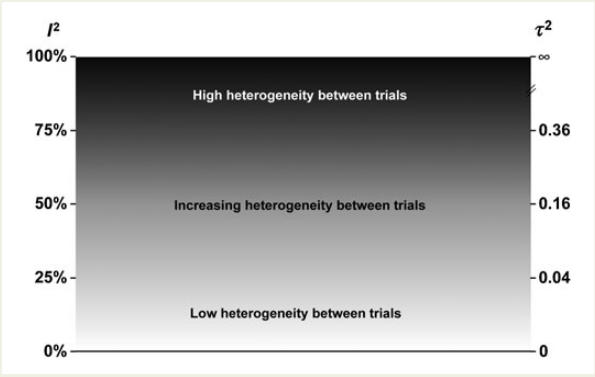


**Figure 3** Interpretation of statistical heterogeneity. The interpretation of $\tau^2$ only holds approximately for estimates of the relative risk (risk ratios, rate ratios, hazard ratios, or odds ratios), while interpretation will be different for risk differences and for continuous outcomes measured on various scales.

## Forest plots

As in any other area of quantitative research, visual inspection of the data used for the analysis is paramount. A forest plot provides at a glance a complete visual summary of results from individual trials included in the meta-analysis. *Figure 4A* and *B* gives examples of two forest plots with 11 trials each. The squares in the plots represent the risk ratios estimated in each of the 11 trials, with the area of each square proportional to the trial's weight in the meta-analysis. The vertical solid line at 1 represents the line of no difference between experimental and control group. Squares to the left of the line of no difference indicate that the experimental intervention is better than the control intervention, squares to the right the opposite. The horizontal lines intersecting the squares represent the 95% confidence intervals of the point estimate of individual trials. The pooled estimate is plotted as a diamond, with the midpoints and the dashed vertical line representing the point estimate and the lateral points the confidence intervals of the pooled estimate.

A visual inspection of treatment effects displayed in a forest plot is complementary to the formal quantification of heterogeneity described above and often considerably more informative. *Figure 4A* (left) presents a homogeneous random-effects meta-analysis of 11 trials to determine the effect of streptokinase on overall mortality in patients with acute myocardial infarction.[33] The major feature of this forest plot is that 95% confidence intervals of all trials widely overlap, indicating that the risk ratios of all trials are compatible with each other, and that 95% confidence intervals of all trials include the pooled estimate shown as a dashed red line. The residual variation in estimated treatment effects, with three trials suggesting a reduction in the risk of death of 50%, whereas another three trials indicate only a reduction of ~20%, is entirely due to chance. *Figure 4B* (right) shows a moderately heterogeneous meta-analysis of 11 trials to determine the benefits of acetylcysteine in reducing contrast-induced nephropathy in patients undergoing angiography.[34] Here, the mutual overlap of 95% confidence intervals of some of the trials is minimal and the 95% confidence interval of one of the trials barely includes the pooled estimate. The moderate heterogeneity is also reflected by considerable discrepancies in the magnitude and direction of estimated treatment effects, with four trials suggesting an 80–90% decrease in the risk of nephropathy and five trials suggesting a 20–30% increase. Accordingly, the most beneficial risk ratios are ~10 times smaller than the least beneficial ones. Note that the mere fact that estimated treatment effects lie on opposite sides of the line of no difference is not sufficient to suggest heterogeneity. Schriger *et al*. have published recommendations regarding what information reviewers should report in forest plots so that readers can properly interpret the results of a meta-analysis.[35]

## Stratified analyses and meta-regression

Stratified analyses and meta-regression aim at determining whether estimates of treatment effects are associated with methodological or clinical characteristics of the trials included in the meta-analysis. The higher statistical heterogeneity between trials the more important these analyses become. A meta-analysis that ignores moderate-to-large extents of heterogeneity is clinically misleading and scientifically naive.[36] Even if there is no or little statistical heterogeneity between trials, stratified analyses can yield valuable insights into clinical or methodological sources of variation between trials.[14]

*Figure 5* shows the results of stratified analyses performed in a meta-analysis comparing the effect of drug-eluting stents vs. bare-metal stents on rates of target vessel revascularization in patients with ST-segment elevation myocardial infarction.[8] The analysis was stratified by four pre-specified trial characteristics: concealment of allocation, blind adjudication of events, analysis according to the intention-to-treat principle (see Box 4), and trial size. For each of the characteristics, trials of higher methodological quality, such as those with adequate concealment of allocation, were pooled separately from trials of lower methodological quality or unclear reporting of the methodological item, using the same model as for the main meta-analysis. All four stratified analyses are accompanied by a *P*-value for interaction, which determines whether the differences between strata may have occurred by chance alone or whether there is evidence for effect modification, with an interaction between-trial characteristic and estimate of treatment effect. For the analysis stratified by concealment of allocation, for example, the treatment effect estimated on a risk ratio scale was 0.58 in trials with adequate concealment and 0.47 in trials with inadequate concealment or unclear reporting. 95% confidence intervals of estimates overlap considerably and the non-significant *P*-value for interaction indicates that the probability that the observed difference between strata or an even larger difference will have occurred by chance is 18%. Conversely, in the analysis stratified by trial size, the risk ratio was 0.57 in large trials with 300 patients or more, but 0.36 in smaller trials. The overlap of 95% confidence intervals was small and the *P*-value for interaction significant ($P < 0.05$).

*P*-values for interaction are typically derived from random-effects meta-regression, which takes into account both within-trial variance of treatment effects and the residual between-trial heterogeneity, which is not explained by the covariate in the model.[37] The meta-regression models used in *Figure 5* were all univariable, including only one binary trial characteristic as independent variable, as referred to above. The model also allows the inclusion of continuous

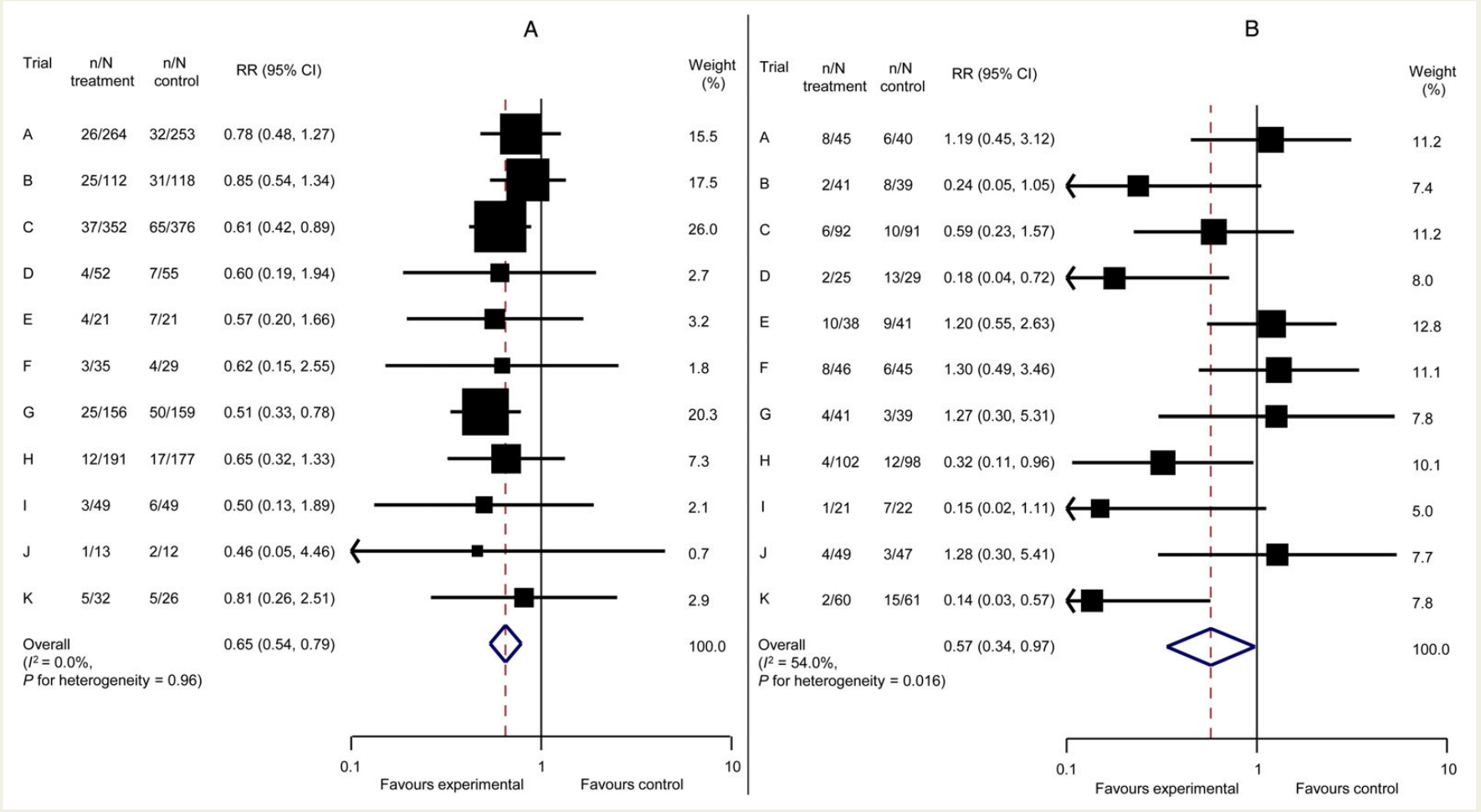


**Figure 4** Forest plots. (*A*) (left) A homogeneous random-effects meta-analysis of 11 trials to determine the effect of streptokinase on overall mortality in patients with acute myocardial infarction.[33] (*B*) (right) A moderately heterogeneous random-effects meta-analysis of 11 trials to determine the benefits of acetylcysteine in reducing contrast-induced nephropathy in patients undergoing angiography.[34] RR: risk ratio; CI: confidence interval; *n*: number or events; *N*: total number of patients.

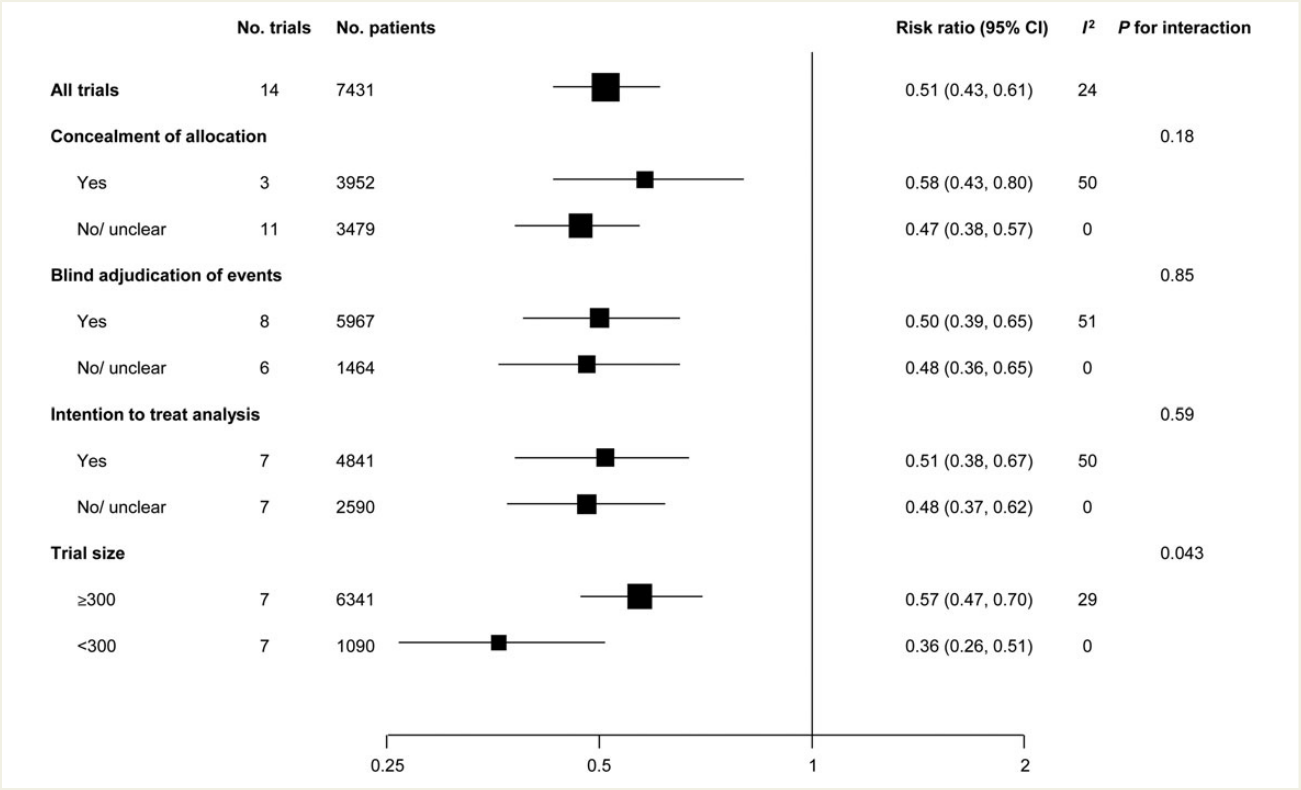


**Figure 5** Stratified analysis by characteristics of trials (adapted from Kalesan *et al.*[8]). CI: confidence interval. A risk ratio below 1 indicates that drug-eluting stents are better than bare-metal stents.

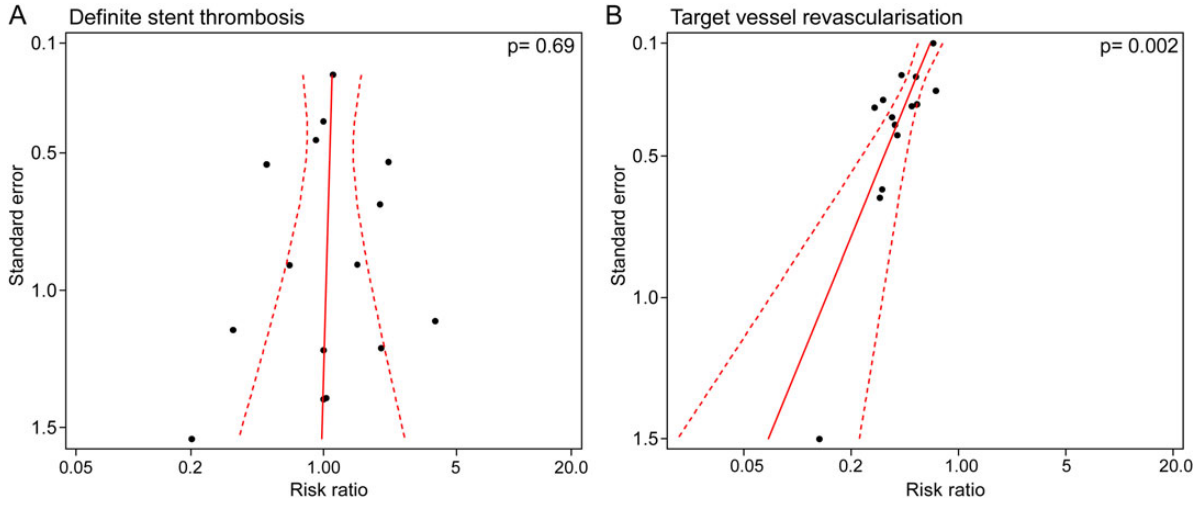


**Figure 6** Funnel plots for definite stent thrombosis (*A*) and target vessel revascularization (*B*) with log of the risk ratio of individual trials on the *x*-axis scattered against the corresponding standard error on the *y*-axis. The larger a trial, the more events accumulated, the smaller the standard error as a measure of statistical precision. In the absence of bias, the scatter of trials should have the shape of an inverted funnel, with large trials scattering little at the top and small trials scattering considerably at the bottom. If the funnel plot is asymmetrical, this suggests the presence of small-study effects, suggesting that methodological problems and selective reporting of outcomes in small trials, and publication bias may have resulted in an overestimation of effects. Red solid lines are prediction lines from univariable meta-regression models with standard error as explanatory variable and red dashed lines are corresponding 95% prediction intervals. The more the prediction line deviates from the vertical line, the more pronounced is asymmetry. *P*-values are from the Harbord test (adapted from Kalesan *et al.*[8]).

covariates. However, it is problematic to include patient characteristics aggregated at trial level, such as mean age or the percentage of females, in the model, since this may produce wrong results due to the ecological fallacy:[28,38] real associations observed at patient level may disappear or even be inversed at trial level, or spurious associations may be found that cannot be verified when analysing IPD.

Meta-regression is frequently used to determine whether estimated treatment effects on a relative risk scale are associated with the underlying baseline risk as measured by the event rate observed in the control group. From a clinical viewpoint, baseline risk is an appealing summary measure of the spectrum of disease severity, comorbid conditions, and risk factors observed in the different patient populations and/or clinical settings of included trials. Unfortunately, this approach is flawed and likely to produce misleading results in most situations.[39,40] When calculating an estimate of the relative risk, the control group event rate is included in the denominator of the estimate. As a special case of regression towards the mean,[41] the relative risk must therefore be associated with the control group event rate: if random variation results in a high control group event rate, then a treatment benefit will become more pronounced, if chance results in a low control group event rate then the treatment benefit will become less pronounced.[40] In the absence of any true association with the predicted risk of a future event in individual patients, meta-regression models, which determine the association of treatment effects with control group event rates, will therefore always find that an observed benefit will be more pronounced in trials with high control group event rates when compared with trials with low control group event rates—a self-fulfilling prophecy. There are no straightforward solutions to determine whether treatment effects depend on the patients' baseline risk using aggregate data at a trial level. Only an analysis of IPD with interaction terms between treatment effect and risk scores, such as the logistic Euroscore,[42] to predict the risk of future events in individual patients will provide clinically meaningful results.[40]

## Funnel plots

The funnel plot is a scatter plot of treatment effects against standard error as a measure of statistical precision. It is expected that treatment effect estimates from smaller trials will scatter more widely in this graph than those of larger trials due to chance. Thus, in the absence of biases, we expect these plots to have the symmetrical shape of an inverted funnel.[43] Deviations from this shape are indicative of small-study effects: larger treatment benefits observed in smaller trials are most likely due to publication bias, selective reporting of outcomes, and other biases commonly seen in small studies with methodological limitations, or—rather exceptionally—due to real clinical heterogeneity with more targeted patient selection or better implementation of interventions in small when compared to large trials.[43] Meta-regression models can be used to test the funnel plot asymmetry.[44] *Figure 6* shows examples of symmetrical and asymmetrical funnel plots with *P* values of meta-regression tests for asymmetry: in a meta-analysis comparing drug-eluting stents with bare-metal stents in patients with ST-segment elevation myocardial infarction the funnel plot was symmetrical for definite stent thrombosis (*Figure 6*A, left), but clearly asymmetrical for the effectiveness outcome of target vessel revascularization (*Figure 6*B, right).[8] In this case, small-study effects such as detection and attrition bias (see Box 4)[12] likely distorted results for revascularization as a major clinical outcome, but not for stent thrombosis as an infrequently occurring safety outcome of secondary importance.

## Complete reporting

Obviously, authors should report all relevant steps taken in their systematic review and meta-analysis.[45] Transparent and complete reporting is crucial so that readers can understand the overall quality of the review, properly interpret results, and replicate or update the review. The Preferred Reporting Items for Systematic Reviews and Meta-analyses (PRISMA) statement is a reporting guideline for systematic reviews and meta-analyses evaluating health-care interventions.[3] Review authors should try to adhere as closely as possible to this guideline when preparing their manuscript.

## Conclusions

Systematic reviews and meta-analyses allow for a more transparent and objective appraisal of the evidence, which may eventually facilitate clinical decision making. They may decrease the number of false-negative results and prevent delays in the introduction of effective interventions into clinical practice. As for any other tool, their misuse can result in severely misleading results. If conducted and reported properly in accordance with the guidance provided in this tutorial, systematic reviews and meta-analyses will increase our understanding of the strengths and weaknesses of the available evidence.

## Funding

This article was supported by intramural funds from the Institute of Social and Preventive Medicine, University of Bern, Switzerland.

**Conflict of interest:** CTU Bern, which is part of the University of Bern, has a staff policy of not accepting honoraria or consultancy fees. However, CTU Bern is involved in design, conduct, or analysis of clinical studies funded by Abbott Vascular, Ablynx, Amgen, AstraZeneca, Biosensors, Biotronik, Boehringer Ingelheim, Eisai, Eli Lilly, Exelixis, Geron, Gilead Sciences, Nestlé, Novartis, Novo Nordisc, Padma, Roche, Schering-Plough, St Jude Medical, and Swiss Cardio Technologies. PJ is an unpaid steering committee or statistical executive committee member of trials funded by Abbott Vascular, Biosensors, Medtronic and Johnson & Johnson. BRDC has no conflicts of interest to declare.

## References

1. Juni P, Egger M. PRISMAtic reporting of systematic reviews and meta-analyses. *Lancet* 2009;**374**:1221–1223.
2. Egger M, Smith GD, Phillips AN. Meta-analysis: principles and procedures. *BMJ* 1997;**315**:1533–1537.
3. Liberati A, Altman DG, Tetzlaff J, Mulrow C, Gotzsche PC, Ioannidis JP, Clarke M, Devereaux PJ, Kleijnen J, Moher D. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate healthcare interventions: explanation and elaboration. *BMJ* 2009;**339**:b2700.
4. Van der Wees P, Qaseem A, Kaila M, Ollenschlaeger G, Rosenfeld R. Prospective systematic review registration: perspective from the Guidelines International Network (G-I-N). *Syst Rev* 2012;**1**:3.
5. Booth A, Clarke M, Dooley G, Ghersi D, Moher D, Petticrew M, Stewart L. PROSPERO at one year: an evaluation of its utility. *Syst Rev* 2013;**2**:4.
6. Smith BJ, Darzins PJ, Quinn M, Heller RF. Modern methods of searching the medical literature. *Med J Aust* 1992;**157**:603–611.

7. Lefebvre C, Manheimer E, Glanville J. Searching for studies. 2011. In: *Cochrane Handbook for Systematic Reviews of Interventions* [Internet]. The Cochrane Collaboration. **www.cochrane-handbook.org.**
8. Kalesan B, Pilgrim T, Heinimann K, Raber L, Stefanini GG, Valgimigli M, da Costa BR, Mach F, Luscher TF, Meier B, Windecker S, Juni P. Comparison of drug-eluting stents with bare metal stents in patients with ST-segment elevation myocardial infarction. *Eur Heart J* 2012;**33**:977–987.
9. Wood L, Egger M, Gluud LL, Schulz KF, Juni P, Altman DG, Gluud C, Martin RM, Wood AJ, Sterne JA. Empirical evidence of bias in treatment effect estimates in controlled trials with different interventions and outcomes: meta-epidemiological study. *BMJ* 2008;**336**:601–605.
10. Nuesch E, Reichenbach S, Trelle S, Rutjes AW, Liewald K, Sterchi R, Altman DG, Juni P. The importance of allocation concealment and patient blinding in osteoarthritis trials: a meta-epidemiologic study. *Arthritis Rheum* 2009;**61**:1633–1641.
11. Schulz KF, Chalmers I, Hayes RJ, Altman DG. Empirical evidence of bias. Dimensions of methodological quality associated with estimates of treatment effects in controlled trials. *JAMA* 1995;**273**:408–412.
12. Juni P, Altman DG, Egger M. Systematic reviews in health care: Assessing the quality of controlled clinical trials. *BMJ* 2001;**323**:42–46.
13. Jadad AR, Moore RA, Carroll D, Jenkinson C, Reynolds DJ, Gavaghan DJ, McQuay HJ. Assessing the quality of reports of randomized clinical trials: is blinding necessary? *Control Clin Trials* 1996;**17**:1–12.
14. Juni P, Witschi A, Bloch R, Egger M. The hazards of scoring the quality of clinical trials for meta-analysis. *JAMA* 1999;**282**:1054–1060.
15. da Costa BR, Hilfiker R, Egger M. PEDro's bias: summary quality scores should not be used in meta-analysis. *J Clin Epidemiol* 2013;**66**:75–77.
16. Greenland S, O'Rourke K. On the bias produced by quality scores in meta-analysis, and a hierarchical view of proposed solutions. *Biostatistics* 2001;**2**:463–471.
17. Greenland S. Quality scores are useless and potentially misleading. *Am J Epidemiol* 1994;**140**:300–301.
18. Higgins JP, Altman DG, Gotzsche PC, Juni P, Moher D, Oxman AD, Savovic J, Schulz KF, Weeks L, Sterne JA. The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. *BMJ* 2011;**343**:d5928.
19. Nuesch E, Trelle S, Reichenbach S, Rutjes AW, Tschannen B, Altman DG, Egger M, Juni P. Small study effects in meta-analyses of osteoarthritis trials: meta-epidemiological study. *BMJ* 2010;**341**:c3515.
20. ISIS-4: a randomised factorial trial assessing early oral captopril, oral mononitrate, and intravenous magnesium sulphate in 58,050 patients with suspected acute myocardial infarction. ISIS-4 (Fourth International Study of Infarct Survival) Collaborative Group. *Lancet* 1995;**345**:669–685.
21. Nuesch E, Juni P. Commentary: which meta-analyses are conclusive? *Int J Epidemiol* 2009;**38**:298–303.
22. DerSimonian R, Laird N. Meta-analysis in clinical trials. *Control Clin Trials* 1986;**7**:177–188.
23. Cornell JE, Mulrow CD, Localio R, Stack CB, Meibohm AR, Guallar E, Goodman SN. Random-effects meta-analysis of inconsistent effects: a time for change. *Ann Intern Med* 2014;**160**:267–270.
24. Smeeth L, Haines A, Ebrahim S. Numbers needed to treat derived from meta-analyses--sometimes informative, usually misleading. *BMJ* 1999;**318**:1548–1551.
25. Deeks JJ. Issues in the selection of a summary statistic for meta-analysis of clinical trials with binary outcomes. *Stat Med* 2002;**21**:1575–1600.
26. O'Farrell N, Egger M. Circumcision in men and the prevention of HIV infection: a 'meta-analysis' revisited. *Int J STD AIDS* 2000;**11**:137–142.
27. Altman DG, Deeks JJ. Meta-analysis, Simpson's paradox, and the number needed to treat. *BMC Med Res Methodol* 2002;**2**:3.
28. Berlin JA, Santanna J, Schmid CH, Szczech LA, Feldman HI. Individual patient- versus group-level data meta-regressions for the investigation of treatment effect modifiers: ecological bias rears its ugly head. *Stat Med* 2002;**21**:371–387.
29. Riley RD, Lambert PC, Abo-Zaid G. Meta-analysis of individual participant data: rationale, conduct, and reporting. *BMJ* 2010;**340**:c221.
30. Stewart LA, Tierney JF, Clarke M. Chapter 18: reviews of individual patient data. 2011. In: *Cochrane Handbook for Systematic Reviews of Interventions* [Internet]. The Cochrane Collaboration. Version 5.1.0. **www.cochrane-handbook.org**.
31. Guyatt GH, Oxman AD, Vist G, Kunz R, Brozek J, Alonso-Coello P, Montori V, Akl EA, Djulbegovic B, Falck-Ytter Y, Norris SL, Williams JW Jr, Atkins D, Meerpohl J, Schunemann HJ. GRADE guidelines: 4. Rating the quality of evidence-study limitations (risk of bias). *J Clin Epidemiol* 2011;**64**:407–415.
32. Rucker G, Schwarzer G, Carpenter JR, Schumacher M. Undue reliance on I(2) in assessing heterogeneity may mislead. *BMC Med Res Methodol* 2008;**8**:79.
33. Egger M, Smith GD. Meta-analysis. Potentials and promise. *BMJ* 1997;**315**:1371–1374.
34. Bagshaw SM, Ghali WA. Acetylcysteine for prevention of contrast-induced nephropathy after intravascular angiography: a systematic review and meta-analysis. *BMC Med* 2004;**2**:38.
35. Schriger DL, Altman DG, Vetter JA, Heafner T, Moher D. Forest plots in reports of systematic reviews: a cross-sectional study reviewing current practice. *Int J Epidemiol* 2010;**39**:421–429.
36. Thompson SG. Why sources of heterogeneity in meta-analysis should be investigated. *BMJ* 1994;**309**:1351–1355.
37. Thompson SG, Higgins JP. How should meta-regression analyses be undertaken and interpreted? *Stat Med* 2002;**21**:1559–1573.
38. Deeks JJ, Higgins JPT, Altman DG. Analysing data and undertaking meta-analyses. 2011. In: *Cochrane Handbook for Systematic Reviews of Interventions* [Internet]. The Cochrane Collaboration. Version 5.1.0. **www.cochrane-handbook.org**.
39. Senn S. Importance of trends in the interpretation of an overall odds ratio in the meta-analysis of clinical trials. *Stat Med* 1994;**13**:293–296.
40. Sharp SJ, Thompson SG, Altman DG. The relation between treatment benefit and underlying risk in meta-analysis. *BMJ* 1996;**313**:735–738.
41. Bland JM, Altman DG. Some examples of regression towards the mean. *BMJ* 1994;**309**:780.
42. Roques F, Michel P, Goldstone AR, Nashef SA. The logistic EuroSCORE. *Eur Heart J* 2003;**24**:881–882.
43. Egger M, Davey Smith G, Schneider M, Minder C. Bias in meta-analysis detected by a simple, graphical test. *BMJ* 1997;**315**:629–634.
44. Harbord RM, Egger M, Sterne JA. A modified test for small-study effects in meta-analyses of controlled trials with binary endpoints. *Stat Med* 2006;**25**:3443–3457.
45. EQUATOR (Enhancing the Quality and Transparency of Health Research) Network. **http://www.equator-network.org/** (27 October 2014).