

available at www.sciencedirect.com
journal homepage: www.europeanurology.com



Platinum Opinion

Guidelines for Meta-analyses and Systematic Reviews in Urology[☆]

Andrew J. Vickers^{a,*}, Melissa Assel^a, Rodney L. Dunn^b, Graeme MacLennan^c, Betsy Jane Becker^d, Richard D. Riley^{e,f}

^a Memorial Sloan Kettering Cancer Center, New York, NY, USA; ^b University of Michigan, Ann Arbor, MI, USA; ^c University of Aberdeen, Aberdeen, UK; ^d Synthesis Research Group, Roswell, GA, USA; ^e Department of Applied Health Sciences, School of Health Sciences, College of Medicine and Health, University of Birmingham, Birmingham, UK; ^f National Institute for Health and Care Research Birmingham Biomedical Research Centre, Birmingham, UK

1. Introduction

Systematic review: The application of strategies that limit bias in the assembly, critical appraisal, and synthesis of all relevant studies on a specific topic. Meta-analysis may be, but is not necessarily, used as part of this process.

Meta-analysis: The statistical synthesis of the data from separate but ... comparable studies, leading to a quantitative summary of the pooled results.

Keith O'Rourke [1]

Systematic reviews are fast becoming the dominant type of paper in the clinical research literature. The number of such papers added to PubMed increased nearly fivefold between 2010 and 2020, whereas the number of trials was relatively constant. Today, more than twice as many systematic reviews are published each year (~60 000) in comparison to randomized controlled trials (~25 000). Similar trends have been seen in urology journals.

It is widely agreed that systematic reviews should be the basis of practice guidelines and medical policy, and thus they have an essential place in the research literature. However, the sheer volume of systematic review articles should be a red flag. It is hard to escape the conclusion that many contemporary systematic reviews are more about padding the authors' curriculum vitae than adding to the scientific record.

Two aspects of systematic reviews make them particularly problematic. First, relative to most research studies, systematic reviews are extremely cheap to conduct, requiring

little or no additional equipment, assistant research staff, or medical costs; even the time commitment for the investigators is relatively moderate. This means that systematic reviews can be conducted without the typical safeguards of funder and institutional peer review. Second, the methodology for systematic reviews is extremely formulaic. The protocol for a randomized trial of, say, a behavioral intervention for patients with dementia should be quite unlike that for a trial of, say, a chemotherapy agent for advanced prostate cancer. Conversely, the protocols for corresponding systematic reviews of such trials could in theory largely overlap, and whole sections on study quality, data extraction, statistical analysis, heterogeneity, and publication bias could essentially be copied and pasted from one to another. It has been documented that the methods sections of different systematic reviews often include identical language [2]. This issue has been exacerbated by the availability of software programs such as RevMan that organize text, tables, and statistical analyses using a common structure. Although such structure is meant to be helpful, it can also lead to less-than-thoughtful application of review methods: the recipe-book nature of systematic reviews can mislead researchers into believing they are easy to perform.

These twin features of systematic reviews have led to a large number of poor-quality systematic review papers being submitted to specialty journals. To address this, we offer the following guidelines to improve the quality of such submissions and the medical research literature as a whole.

[☆] This paper was jointly developed by European Urology, Urology, The Journal of Urology, BJU International and jointly published by Elsevier BV, Elsevier Inc, Wolters Kluwer Health Japan Co., Ltd., and John Wiley and Sons Inc.. The articles are identical except for minor stylistic and spelling differences in keeping with each journals style. Either citation can be used when citing this article

* Corresponding author. Memorial Sloan Kettering Cancer Center, 1275 York Avenue, NY 11215, USA.
E-mail address: vickersa@mskcc.org (A.J. Vickers).

2. Overall considerations

2.1. Follow the PRISMA and MOOSE guidelines

The basics on how to report systematic reviews and meta-analyses have been carefully laid out in excellent guidelines written by methodologists and statisticians: the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines for reviews of studies evaluating the effects of interventions [3] and the Meta-analysis of Observational Studies in Epidemiology (MOOSE) guidelines for observational studies [4]. Much of this work is also relevant for the conduct of reviews, alongside textbooks such as the *Cochrane Handbook* (<https://training.cochrane.org/handbook>). Introductory papers for urologists are a further resource [5].

2.2. Follow the urology guidelines

Systematic reviews and meta-analysis should follow the guidelines on statistical reporting [6] and on presentation of figures and tables [7] and, where appropriate, reporting of causality in observational research [8], just the same as for any other paper in clinical urology. Some of the “standard” or default settings in software such as RevMan can lead to output that is inconsistent with our guidelines, such as those on precision [6]. Our view is that decisions on how we present scientific evidence should be based on best practice, not software.

2.3. Describe the review methodology in sufficient detail to allow an independent team to replicate the results exactly

A systematic review or meta-analysis should be reproducible. To ensure reproducibility, investigators must be specific and detailed about their methodology choices and document these in a prespecified protocol. For instance, a review evaluating a diagnostic test for the presence of high-grade cancer would need to be 100% clear about the definition of high grade and about how data were handled from papers that used alternative definitions. For reviews of studies evaluating the effects of interventions, the PICO (Population, Intervention, Comparator, Outcome) question should be clearly stated.

3. Systematic reviews

3.1. Justify why the systematic review is of value

Any systematic review must include a compelling rationale for its publication. The main conclusion of all too many systematic reviews is that although preliminary data are promising, the quality of studies is moderate or poor, so more research is required. Such conclusions would be readily apparent to any investigator casually reading the relevant literature and thus the systematic review does not make a scientific contribution. A persuasive rationale should be framed first in terms of how the review will provide more information than a reading of the original literature. Second, given that so many systematic reviews already exist, the rationale should state how the new review is an

important advance on prior reviews. Inclusion of a newly published primary study not included in a prior review is not, by itself, a sufficient rationale for a new systematic review. A good rationale will give convincing reasons to believe that the new data or analyses would importantly influence the quality, strength, or direction of the findings, or provide a critical methodologic perspective.

3.2. Recommendations for further research must follow as specific conclusions that are based on the data

Systematic reviews appear to be particularly prone to vague conclusions about the need for further research. If such calls are not informed by the results of the review, then the review has not been of value. Reviews must go beyond linking a finding such as “research was of poor quality” to a conclusion that “better research is needed”, as this is almost always true in clinical research and is usually readily apparent to any attentive reader. The logical link between findings and conclusions with respect to implications for research must identify specific and idiosyncratic aspects of the subject matter for improvement. For instance, a systematic review might document that follow-up periods were too short and so make recommendations that subsequent research should follow patients for longer-term outcomes.

3.3. Any risk-of-bias assessments must be incorporated in the results or conclusions

In many systematic reviews, investigators conduct a risk-of-bias assessment evaluating the quality of primary studies using a published set of criteria, but then do little more than give the results of that assessment in a table, with no further consideration of how this impacts the results or conclusions. This can lead to situations where, for instance, a biased study showing a strong effect is combined with a high-quality study showing no difference between groups, leading to an invalid conclusion that a treatment has an effect somewhere in between the two.

4. Meta-analysis

4.1. Do not ignore heterogeneity when providing central estimates

Heterogeneity is a fundamental aspect of meta-analysis that always needs careful consideration. We expect the results observed (e.g., treatment effect estimates) in different studies to vary, but need to assess whether the degree of variation is greater than that expected by chance alone (sampling variability). If so, this is referred to as between-studies heterogeneity. Such variation can be caused by substantive differences between the studies, whether clinical—for example, the case mix of patients, the nature of interventions (eg, dose, duration) or comparators, or the length of follow up—or methodologic, such as how the outcome was assessed or the data were analyzed. The presence of heterogeneity raises important questions about whether it is reasonable to combine the results of studies and report an overall summary estimate from meta-analysis. For instance, Figure 1A provides an example in which three tri-

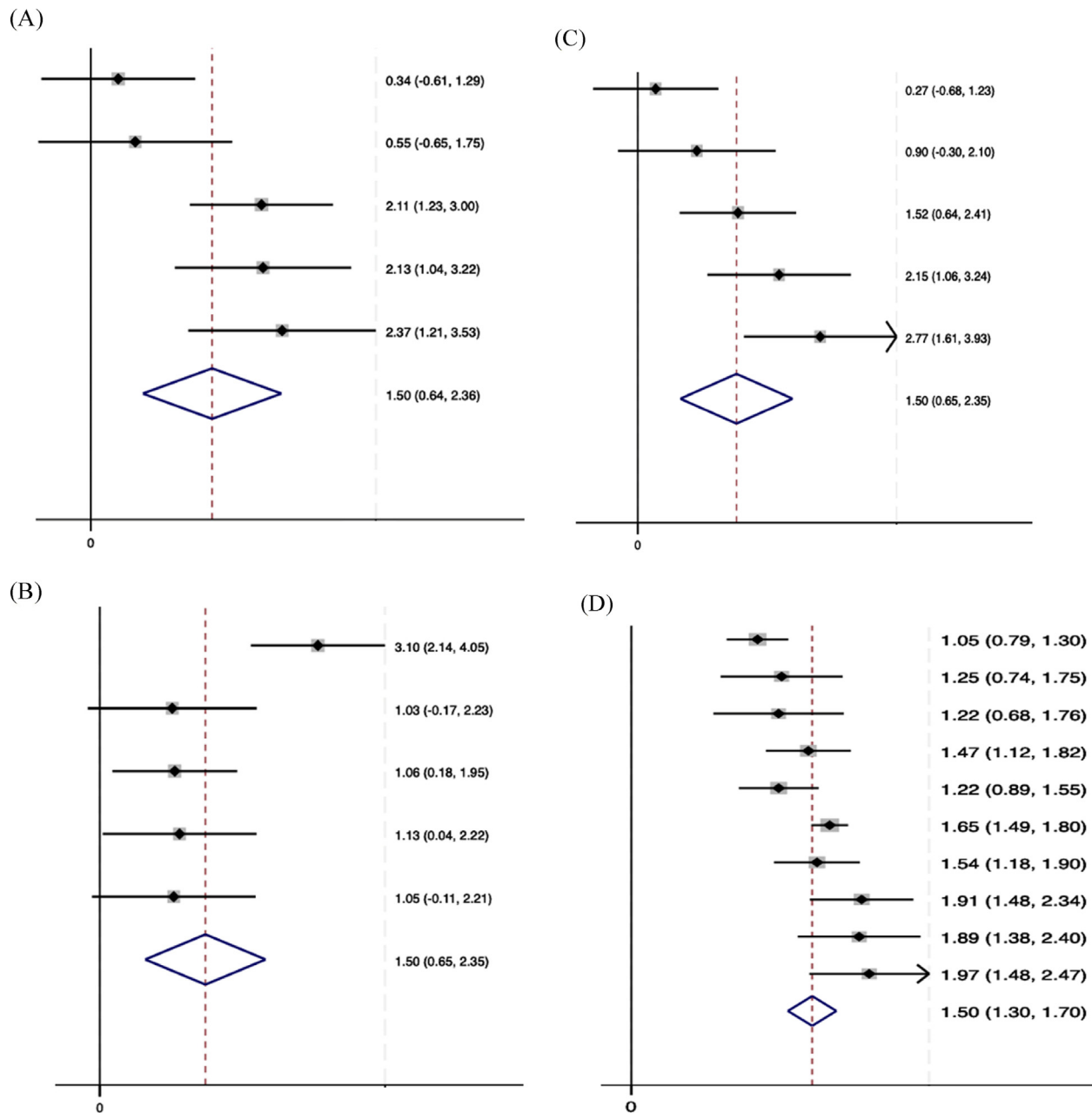


Fig. 1 – (A–D) Examples of meta-analyses showing heterogeneity and its measurement. Random-effects results are shown using an arbitrary scale, with a value of 1.5 being considered clinically significant. All trials in A–C are of similar size; the trials in D are on average much larger (16-fold). All four meta-analyses have statistically significant heterogeneity, with an I^2 value of approximately 70% for each. The τ^2 value is ~ 0.65 for A–C, but ~ 0.065 for D. Prediction intervals are -1.9 to 4.9 for A–C, and 0.1–2.9 for D.

als show a similar, large treatment effect, and two have small and nonsignificant effects. It would not be appropriate to enter all five trials into a meta-analysis and conclude that the treatment has an effect size somewhere between the two sets of trials. However, this is exactly the sort of mistake that many authors make when they ignore evidence of heterogeneity and report a meta-analytic average that is essentially meaningless.

4.2. Investigate potential sources of heterogeneity to determine the factors that lead to differences in study results

Assessment of heterogeneity might involve trying to identify common features of studies with similar findings

(eg, perhaps short treatment durations led to the null results for the two trials in Fig. 1A) or idiosyncratic aspects of studies with outlying results (such as the first study in Fig. 1B). This is admittedly a difficult process, especially if the review has only a few studies. Such questions can be challenging to address via post hoc statistical analyses alone, such as meta-regression to evaluate factors associated with the study results. The process will instead often depend on background knowledge. For instance, investigators might notice that trials in which the effect size was smaller tended to have a more aggressive treatment in the control arms. This would make scientific sense as an explanation for heterogeneity, even if were difficult to

statistically test a hypothesis because of low power. The preferred approach would be to include in the review protocol the features (such as the comparator or duration of follow-up) that are expected to lead to differences in study results.

4.3. Avoid the I^2 statistic

Heterogeneity can be assessed using Cochran's Q test, which gives a p value for the null hypothesis of homogeneity. It is purported that the I^2 statistic complements this inference by providing an estimate: the percentage of the total variance of study effect estimates that is attributable to genuine differences between studies, such as patients or treatments, rather than to chance (sampling error within studies). The major problem with I^2 is that it is sample size-dependent. Unless there is truly no heterogeneity, I^2 will tend towards 100% as the number and size of studies increase. For example, the I^2 estimate for the meta-analysis shown in Figure 1D, which includes a greater number of large trials, all with statistically and clinically significant benefits, is the same as for the meta-analyses in Figure 1A–C, which include lower numbers of small trials with extremely different results. I^2 has been widely promoted and is hard-wired in most meta-analysis software, but it is not actually an estimate of heterogeneity and so should be avoided. Some authors have recommended reporting τ^2 [9] in place of I^2 on the grounds that τ^2 is an actual estimate of between-study variance and does not vary with sample size. However, τ^2 can be difficult to interpret, as it is on the scale of the meta-analysis model, which might be a log odds ratio or log hazard. Others have advocated for a prediction interval, which is the range of likely values of the effect size were a new study to be conducted [10]. However, the most important step is a critical evaluation of the forest plot. The plots in Figure 1A–C show that the same τ^2 value and the same prediction interval can result from very different forest plots: one plot has a single outlying study; another has two groups of studies, one with null and one with clinically relevant findings; and the third plot has effect sizes that vary along a continuum. The inferences we would draw from these three forest plots will be very different despite similar estimates of variation.

4.4. Do not assume that random-effects approaches dispense with the problem of heterogeneity

Results from multiple studies are combined in a meta-analysis using either fixed- or random-effects methods. An overview of the difference between the two approaches is given in Text Box 1. We will not debate the approaches here, other than to say that (1) random-effects models are not a solution to the problem of study heterogeneity, and (2) observation of heterogeneity cannot be used as a justification for choosing a random-effects approach. With respect to the latter, authors of meta-analyses often include statements such as “the test for heterogeneity was statistically significant ($p = 0.002$) so a random-effects model was used”. However, this is not a justifiable position because, as mentioned, a meta-analytic estimate is difficult to interpret in the face of heterogeneity. Rather, as described in Section 2.1,

Text box 1 Fixed versus random effects.

For the meta-analysis of the trials in Figure 1D, a fixed-effect meta-analysis (also known as a common-effect meta-analysis) combines results from the ten studies and assumes that the true effect is the same in each study [17]. This gives the same result as if all patients were part of one study. By contrast, a random-effects meta-analysis assumes that the true effect varies across studies because of unexplained heterogeneity [18]; it often also assumes that the true effects are sampled from a hypothetical distribution (eg, a normal distribution). The key impact is that the weighting of each study in the meta-analysis summary changes depending on whether a fixed effect or random effects are assumed, with the latter giving more equal weight to small and large studies as heterogeneity increases. This occurs because the estimand is changing: in a fixed-effect meta-analysis, the summary result is the best estimate of an assumed single true effect; in a random-effects meta-analysis, the summary result is the best estimate of the average of the assumed distribution of true effects.

The debate about which approach is superior is essentially philosophical. Proponents of the fixed-effect approach argue that the job of meta-analysis is to summarize the results of studies in the literature, not to speculate about hypothetical distributions of trials that have never been conducted. Advocates of the random-effects approach counter that we should expect study results to differ and that ignoring a distribution of effects does not make it go away. An argument for the random-effects approach is that assumption of a fixed effect gives so little weight to small studies that they might as well not have been conducted; the counterargument is that it seems odd that the random-effects approach upweights small trials, given that we know that such trials are generally unreliable. Supporters of the random-effects approach claim that it is more conservative (because it results in wider confidence intervals), although that is not always true [19]; in any case, it is not clear why that is an advantage.

That said, there is general agreement on two points. First, fixed-effect analysis gives a valid test of the null hypothesis of no effect in any trial. Second, random-effects analysis requires an estimate of the between-study variance, and this estimate is unreliable when the number of studies is small. In typical simulation studies of meta-analytic variance estimators, the minimum number of studies used to evaluate different estimators of between-study variance is ten, and authors often conclude that at least 20–30 studies are required [20]. In practice, most medical meta-analyses include fewer trials than this: one study reported a median for the number of trials included in meta-analyses of medical treatments as just four [21].

if heterogeneity exists, the reasons for this heterogeneity need to be explored and the relevance of any meta-analytic estimate of the “average effect” must be justified when pooling dissimilar studies. Merely giving the results from a random-effects meta-analysis is not an explanation of heterogeneity or a justification for a meta-analytic average. For instance, in Figure 1A the random-effects estimate is the same as the fixed-effects estimate (1.50) but the 95% confidence interval (CI) is wider: 0.64–2.36 versus 1.04–1.97. Thus, we would be less sure about the results of any

future study because we would not know whether it would be more like the null trials or more like the trials showing a treatment effect. But such a conclusion is fundamentally uninteresting given that the meta-analytic estimate is difficult to interpret in the context of such heterogeneity. To re-emphasize, the appropriate response to heterogeneity is not statistical, it is to evaluate the patients, treatments, and other aspects of the substance and methodology of the different studies to understand the mechanisms causing the heterogeneity observed.

4.5. Use appropriate methods to extract time-to-event data from primary studies

Studies with a time-to-event endpoint, such as overall or recurrence-free survival, report summary statistics in a variety of different ways. This is problematic because meta-analysis requires data in a single form, an estimate such as a log hazard ratio and its standard error. Extracting a log hazard ratio and standard error from a trial report can be relatively straightforward, for instance, if the authors have reported a hazard ratio and 95% CI. However, more involved calculations may be required, for example, if the authors have reported only a *p* value and the number of events in each arm [11]. In some cases, meta-analysts need to use the survival curve, draw lines to the *x* and *y* axes (perhaps using a data extraction app such as PlotDigitizer to do so), and combine the extracted data with information on the number of patients at risk over time [12]. The approaches used to obtain time-to-event data should be described in detail. Some methods, however, are not valid, such as treating survival as a continuous variable, or dividing two Kaplan-Meier probabilities at a specific time point to derive a relative risk.

4.6. Use care in the interpretation of funnel plots and tests for publication bias

Publication bias occurs when the propensity of a finding to be published depends on the strength and direction of its findings. Suppose that a surgeon has developed a new procedure and conducted a small trial. If the findings are positive, the surgeon will be keen to publish; if the findings do not favor the new procedure, the surgeon may lose interest and leave the trial report to languish in a file drawer. If trials with nonsignificant differences between groups are less likely to be published, then any meta-analysis of the published literature will give an inflated estimate of the treatment effect. Statistical methods to assess publication bias are available, but application of these methods is far from straightforward. Statistical tests for publication bias have very low power and generally require at least 10–20 studies [13]. Moreover, the tests are unreliable in the presence of heterogeneity [14]. The same is true of funnel plots, which relate effect sizes to measures of their uncertainty. For instance, if the results from small trials differ from those from large trials because of differences in patient populations or treatments (ie, factors that cause heterogeneity), then asymmetry may be evident on a funnel plot even when publication bias does not exist. This is why funnel plots and tests of bias evaluate not only publication bias but also

small-study effects. Given that most meta-analyses involve modest numbers of trials and often show evidence of heterogeneity, funnel plots and tests for publication bias will rarely be of value.

4.7. Avoid flat or unspecified prior distributions in Bayesian meta-analysis

Bayesian methods are becoming increasingly popular, particularly in network meta-analysis. Such methods require the use of prior distributions on the basis of external or user-specific knowledge. However, it is not at all uncommon for authors to use flat prior distributions. This is irrational. Take, for instance, a meta-analysis on the effect of third-line chemotherapy versus best supportive care for metastatic cancer. Use of a flat prior distribution would suggest that it is equally possible that chemotherapy cures every patient as that it leads to immediate death, and both are no less likely than a clinically reasonable outcome such as a moderate increase in time to progression. Other authors fail to specify their prior distribution altogether, something which makes the Bayesian result uninterpretable. A good example of sensible prior distributions in meta-analysis is for τ^2 , whereby empirical distributions of heterogeneity estimates from previous meta-analyses are used as the prior distribution [15].

4.8. Avoid ranking methods such as SUCRA in isolation

In network meta-analyses, authors sometimes report ranking scores for different interventions that are based on a metric such as the surface under the cumulative ranking (SUCRA), P-score, or mean rank, and then make conclusions about relative effectiveness. For example, authors might report SUCRA scores of, say, 85%, 75%, and 45% for interventions A, B, and C, respectively, and conclude that intervention A is superior to B, and that B is superior to C. Alternatively, they may order treatments by the probability of being ranked first. This type of conclusion is generally unsound because it does not consider study quality or statistical or clinical significance. The SUCRA scores would be unaffected if the trials on intervention A were of poor quality, or if no statistical differences were found between the three treatments, or the differences between treatments were very small. In a published example, the SUCRA scores above—suggesting the superiority of intervention A—were derived from nonsignificant risk ratios of 0.95 (95% CI 0.65–1.38) for A versus B, 0.87 (95% CI 0.55–1.36) for B versus C, and 0.75 (95% CI 0.35–1.65) for A versus C [16]. These data would not justify favoring intervention A in comparison to B and C, as would be suggested by the SUCRA rankings, especially if it was more toxic, expensive, or inconvenient, because there is little evidence that any of these interventions is superior to the others.

5. Conclusions

The guideline points above address both the interpretation and the conduct of meta-analyses. They reflect our view that meta-analysis cannot be reduced to “cookie cutter science” and that statistical analysis is meant to inform,

not replace critical thinking. Application of the guidelines would lead to a more considered interpretation of a smaller number of systematic reviews and meta-analyses, and could thus help to translate evidence into better decision-making for doctors and patients.

Conflicts of interest: The authors have nothing to disclose.

Funding/Support and role of the sponsor: This work was supported in part by the National Institutes of Health/National Cancer Institute with a cancer center support grant to Memorial Sloan Kettering Cancer Center (P30 CA008748), a SPORC grant in prostate cancer to Dr. H. Scher (P50-CA92629), the Sidney Kimmel Center for Prostate and Urologic Cancers, and David H. Koch through the Prostate Cancer Foundation. Funding was also received from the National Institute for Health and Care Research (NIHR) Birmingham Biomedical Research Centre at the University Hospitals Birmingham NHS Foundation Trust and the University of Birmingham, and an MRC-NIHR methodology grant (MR/V038168/1). Richard D. Riley is an NIHR Senior Investigator. The views expressed are those of the authors and not necessarily those of the NHS, the NIHR, or the Department of Health and Social Care.

References

- [1] O'Rourke K. An historical perspective on meta-analysis: dealing quantitatively with varying study results. *J R Soc Med* 2007;100:579–82.
- [2] Vickers AJ. Reducing systematic review to a cut and paste. *Forsch Komplementmed* 2010;17:303–5.
- [3] Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ* 2021;372:n71.
- [4] Stroup DF, Berlin JA, Morton SC, et al. Meta-analysis of observational studies in epidemiology: a proposal for reporting. Meta-analysis Of Observational Studies in Epidemiology (MOOSE) group. *JAMA* 2000;283:2008–12.
- [5] Tseng TY, Dahm P, Poolman RW, Preminger GM, Canales BJ, Montori VM. How to use a systematic literature review and meta-analysis. *J Urol* 2008;180:1249–56.
- [6] Assel M, Sjöberg D, Elders A, et al. Guidelines for reporting of statistics for clinical research in urology. *Eur Urol* 2019;75:358–67.
- [7] Vickers AJ, Assel MJ, Sjöberg DD, et al. Guidelines for reporting of figures and tables for clinical research in urology. *Eur Urol* 2020;78:97–109.
- [8] Vickers AJ, Assel M, Dunn RL, et al. Guidelines for reporting observational research in urology: the importance of clear reference to causality. *Eur Urol* 2023;84:147–51.
- [9] 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.
- [10] Borenstein M. Avoiding common mistakes in meta-analysis: understanding the distinct roles of Q, I-squared, tau-squared, and the prediction interval in reporting heterogeneity. *Res Synth Methods* 2024;15:354–68.
- [11] Parmar MK, Torri V, Stewart L. Extracting summary statistics to perform meta-analyses of the published literature for survival endpoints. *Stat Med* 1998;17:2815–34.
- [12] Tierney JF, Burdett S, Fisher DJ. Practical methods for incorporating summary time-to-event data into meta-analysis: updated guidance. *Syst Rev* 2025;14:84.
- [13] Almalik O, Zhan Z, van den Heuvel ER. Tests for publication bias are unreliable in case of heteroscedasticity. *Contemp Clin Trials Commun* 2021;22:100781.
- [14] Peters JL, Sutton AJ, Jones DR, Abrams KR, Rushton L, Moreno SG. Assessing publication bias in meta-analyses in the presence of between-study heterogeneity. *J R Stat Soc Ser A Stat Soc* 2010;173:575–91.
- [15] Turner RM, Jackson D, Wei Y, Thompson SG, Higgins JP. Predictive distributions for between-study heterogeneity and simple methods for their application in Bayesian meta-analysis. *Stat Med* 2015;34:984–98.
- [16] Mbuagbaw L, Rochwerg B, Jaeschke R, et al. Approaches to interpreting and choosing the best treatments in network meta-analyses. *Syst Rev* 2017;6:79.
- [17] Rice K, Higgins JPT, Lumley T. A re-evaluation of fixed effect(s) meta-analysis. *J R Stat Soc Ser A Stat Soc* 2017;181:205–27.
- [18] Higgins JP, Thompson SG, Spiegelhalter DJ. A re-evaluation of random-effects meta-analysis. *J R Stat Soc Ser A Stat Soc* 2009;172:137–59.
- [19] Poole C, Greenland S. Random-effects meta-analyses are not always conservative. *Am J Epidemiol* 1999;150:469–75.
- [20] Sidik K, Jonkman JN. A comparison of heterogeneity variance estimators in combining results of studies. *Stat Med* 2007;26:1964–81.
- [21] Int'Hout J, Ioannidis JP, Borm GF, Goeman JJ. Small studies are more heterogeneous than large ones: a meta-meta-analysis. *J Clin Epidemiol* 2015;68:860–9.