BJA

British Journal of Anaesthesia, 126 (4): 903–911 (2021)

doi: 10.1016/j.bja.2021.01.004
Advance Access Publication Date: 6 February 2021
Special Article

STATISTICS IN ANAESTHESIA

Methodologies for systematic reviews with meta-analysis of randomised clinical trials in pain, anaesthesia, and perioperative medicine

Brett Doleman[1,*], Ole Mathiesen[2,3], Janus C. Jakobsen[4,5], Alex J. Sutton[6], Suzanne Freeman[6], Jonathan N. Lund[1] and John P. Williams[1]

[1]Department of Anaesthesia and Surgery, Graduate Entry Medicine, University of Nottingham, Nottingham, UK, [2]Department of Medicine, University of Copenhagen, Copenhagen, Denmark, [3]Department of Anaesthesia, Zealand University Hospital, Køge, Denmark, [4]Copenhagen Trial Unit, Copenhagen, Denmark, [5]Department of Regional Health Research, Faculty of Heath Sciences, University of Southern Denmark, Odense, Denmark and [6]Department of Health Sciences, University of Leicester, Leicester, UK

*Corresponding author. E-mail: dr.doleman@gmail.com

Summary

Systematic reviews and meta-analyses (SRMAs) are increasing in popularity, but should they be used to inform clinical decision-making in anaesthesia? We present evidence that the certainty of evidence from SRMAs in anaesthesia (and in general) may be unacceptably low because of risks of bias exaggerating treatment effects, unexplained heterogeneity reducing certainty in estimates, random errors, and widespread prevalence of publication bias. We also present the latest methodological advances to help improve the certainty of evidence from SRMAs. The target audience includes both review authors and practising clinicians to help with SRMA appraisal. Issues discussed include minimising risks of bias from included trials, trial sequential analysis to reduce random error, updated methods for presenting effect estimates, and novel publication bias tests for commonly used outcome measures. These methods can help to reduce spurious conclusions on clinical significance, explain statistical heterogeneity, and reduce false positives when evaluating small-study effects. By reducing concerns in these domains of Grading of Recommendations, Assessment, Development and Evaluation, it should help improve the certainty of evidence from SRMAs used for decision-making in anaesthesia, pain, and perioperative medicine.

Keywords: meta-analysis; methodology; perioperative medicine; postoperative pain; publication bias; systematic review

Editor's key points

- Systematic reviews and meta-analyses (SRMAs) may be poor predictors of results from large, well-conducted RCTs.
- The number of high-certainty SRMAs published on anaesthesia topics may be unacceptably low because of risks of bias, random error, unexplained heterogeneity, and publication bias.
- This review outlines methodological strategies that can help authors and researchers improve the certainty of evidence in the anaesthesia literature, and help clinicians appraise the evidence.
- Application of these strategies should help improve the certainty of evidence from SRMAs used for decision-making in anaesthesia, pain, and perioperative medicine.

Received: 9 November 2020; **Accepted:** 7 January 2021

© 2021 British Journal of Anaesthesia. Published by Elsevier Ltd. All rights reserved.
For Permissions, please email: permissions@elsevier.com

Systematic reviews and meta-analyses (SRMAs) have increased in popularity over recent years with annual publications increasing by more than 2000% since 1991.[1] The widespread availability of free analysis software and guidance, in addition to their prominence in the hierarchy of evidence, has presumably led to their increased popularity. Whilst SRMAs of RCTs have many advantages over individual trials, such as improved power and lower error rates,[2] significant criticisms of SRMAs persist.[3] For example, the accuracy of SRMAs in predicting the results from later large RCTs has been widely questioned. To illustrate issues with SRMA accuracy, diagnostic test statistics can be used, such as positive predictive values (PPVs: SRMA positive and large RCT positive). Using these statistics, previous studies found a PPV of only 68%[4] and another <67%[5] in general and perinatal medicine, respectively. More recently and specific to anaesthesia, a study of perioperative interventions found PPV of only 23% and the discrimination no better than a coin toss.[6] Although there are limitations with this approach, in view of their apparent poor predictive ability, should SRMAs be used to inform clinical decision-making in anaesthesia, pain, and perioperative medicine?

To ensure the validity of SRMAs, a concerted effort is required from those performing such reviews to adhere to the highest methodological standards rather than the pursuit of significant or interesting results. This article aimed to increase the certainty of evidence (COE) as per Grading of Recommendations, Assessment, Development and Evaluation (GRADE) by presentation of novel methodological developments. These evaluations present findings as a range from high certainty to very low certainty depending on deficiencies in five areas. When evaluating Cochrane reviews in anaesthesia using GRADE, only 10% of primary outcomes were of high certainty.[7]

This article is structured in the format of GRADE. First, we present a background on each domain using simple examples for the concepts discussed. This is supplemented with specific research findings in the field of anaesthesia highlighting widespread deficiencies in each domain. Second, we supplement these with examples of methodology that can be used to improve the COE in that domain. To illustrate specific issues, we reanalyse a data set from a recent SRMA. This is a Cochrane review published in 2018, which evaluated perioperative administration of ketamine compared with placebo in adult participants undergoing general anaesthesia. It was conducted in a variety of surgeries, and outcomes included postoperative pain, opioid consumption, and opioid adverse events (such as nausea and vomiting).[8]

General issues

The starting point for any systematic review, once a suitable population, intervention, control, and outcome (PICO) question has been developed, is to register the review on a database, such as International Prospective Register of Systematic Reviews (PROSPERO)[9] or publish a protocol.[10] There are now more than 25 000 systematic reviews registered on PROSPERO.[11] Such registration helps reduce selective outcome reporting, prevents duplicate reviews, and helps ensure methods are not amended during review of the obtained studies. Although registration is associated with higher-quality SRMA,[12] 30% still alter outcomes,[13] often without explanation.[14] Also essential is the Preferred Reporting Items for Systematic Review and Meta-Analysis (PRISMA) reporting standard checklist, which is often a requirement for publication and ensures that important aspects of methodology are reported.[15] Indeed, use of PRISMA has led to both improvements in reporting and methodological standards for SRMAs.[16] In anaesthesia SRMAs, improvements in PRISMA reporting may be associated with improvements in study quality, whilst involvement of statisticians may help improve this further.[17]

As part of PICO, selection of important outcomes is essential,[18] and *minimal important differences* should always be prespecified.[19] For perioperative medicine, a move towards patient-important, standardised outcomes should be advocated, which helps with the synthesis of outcomes and helps reduce the use of *surrogate outcomes*.[20–24] Surrogate outcomes are outcomes that are thought to be representative of the actual outcome of interest (e.g. bispectral index values instead of awareness).[25] Another outcome-related issue is the inclusion of adverse events and serious adverse events with relevant follow-up periods, which are essential when making balanced clinical decisions on the use of a particular intervention.[26,27] For example, although use of multimodal analgesics may reduce opioid consumption, if they cause other adverse events (e.g. visual disturbance with pregabalin[28]) in selected patients, then the risks may outweigh any benefits.

When different interventions are used to treat a particular condition, *network meta-analysis* may be considered, which can compare interventions both directly and indirectly.[29] For example, in addition to including data from trials, where drug A and drug B are compared in the same trial, indirect comparison can be made, where both drugs are compared with a common comparator using the relative effects. This analysis can be used to identify the most effective treatment from a range of options, such as non-opioid analgesics for postoperative pain.[30] However, limitations of such analyses must be considered, including common violations to assumptions and consistency (direct and indirect effects should be similar).[31]

Risk of bias

Issues surrounding the conduct of individual RCTs included in an SRMA are fundamental, as inclusion of high-risk-of-bias trials in an SRMA will inevitably bias the results of a review (garbage in=garbage out). Review authors (and readers) require detailed knowledge of what issues in RCTs cause bias and what methods constitute a low risk of bias in any domain.

Risk of bias describes elements of systematic errors in the individual trials that may cause results to deviate from their 'true' value. *Selection bias* occurs when intervention and control groups differ in a particular way, which can occur because of inadequate randomisation or allocation concealment. For example, a high risk of bias from quasi-randomisation occurs if participants are randomised based on time of day seen in a preoperative assessment clinic. Selection bias occurs because participants attending later in the day (possibly because of employment) may differ from those attending earlier. With regard to allocation concealment, if researchers know which group the next participant will be allocated, they may subconsciously (or consciously) not approach that participant in an attempt to skew results. Research conducted more than a decade ago in anaesthesia found allocation concealment was inadequately described in 39–65% of trials, although it may be improving.[32] More recently, in our example ketamine review, 83/130 (64%) of trials did not describe allocation concealment adequately.[8] Patient selection bias has been shown to exaggerate effects in trials generally[33,34] and in pain studies.[35] Using transcutaneous electrical nerve stimulation as an example, 88% of randomised studies found no benefit, whilst in non-RCT studies 89% found a benefit.[36]

Blinding involves participants, staff, and researchers being unaware which group participants have been allocated to in a study. Inadequate blinding can occur when no placebo is used, or when the placebo differs from the intervention in either appearance or effects (gabapentin causing sedation). Another example occurs in trials, where blinding is difficult or impossible, such as perioperative exercise interventions. Blinding reporting in trial publications may be poor,[37] and assessments are frequently inadequately performed by review authors.[38] Because of subjective outcomes in pain studies, placebo effects may be large, and therefore, inadequate blinding can contribute to a large overestimation of effects.[39] Inadequate blinding of outcome assessors, especially with subjective outcomes, can also overestimate treatment effects.[40]

Other domains of bias that can exaggerate effects include patients lost to follow-up (*attrition bias*)[41] and their subsequent exclusion from the analysis (no longer *intention to treat*). This can cause a form of selection bias. For example, in studies of gabapentin for postoperative pain, excluding participants with uncontrolled pain from the gabapentin group will potentially bias results.[42] *Selective outcome reporting* occurs when trial authors fail to report results or change primary outcomes based on statistical significance, which leads to a form of reporting bias. In the anaesthesia literature, one study found discrepancies in 48% of registrations when compared with published trials,[43] and another found 92% of registered trials had an outcome discrepancy when compared with the final publication.[44]

Because of the aforementioned issues, undertaking risk-of-bias assessments (in duplicate) is an essential part of the review process. For RCTs, the Cochrane risk-of-bias tool is recommended.[45] It was used in only 20% of anaesthesia SRMAs in 2015, although widespread adoption since means this proportion is likely much higher for current SRMAs.[46] However, to improve COE, sensitivity (or primary) analysis should aim to analyse only those trials at low risk of bias for all domains,[47] or alternatively, include risk of bias in subgroup or meta-regression analysis to see how this influences effect estimates. Only then can this domain of GRADE have no concerns, as this is the only circumstance when the effects of risk of bias can be minimised. In the anaesthesia literature, such reanalysis was shown to change effect estimates around 50% of the time.[46]

Inconsistency

In SRMAs of RCTs, *clinical heterogeneity* can occur, which describes differences between study characteristics, including population, interventions, or methods. It may only be valid to perform meta-analysis if the clinical heterogeneity is low. *Statistical heterogeneity* concerns the differences between studies; it should be assessed by visual inspection of forest plots and can be quantified using the I^2 statistic.[48] Statistical tests have inherent issues, such as low power with small numbers of studies (common) and should therefore be avoided. The I^2 statistic gives a value between 0 and 100%, and describes how much of the variability is attributable to differences between studies rather than sampling variance (chance). Therefore, a value of 95% means nearly all the variability is attributable to differences between studies and was frequently observed for continuous outcomes in the example ketamine review.[8] This could be caused by many factors, some of which were discussed previously. It can be difficult to suggest absolute cut-offs for I^2, as its value should be regarded as a continuum, although values >50%, which are unexplained by study factors, may indicate at least substantial heterogeneity (as per the *Cochrane Handbook for Systematic Reviews of Interventions*). Note that, although I^2 does not depend on the number of studies in a meta-analysis, it does depend on the precision of studies (which is proportional to study size). Therefore, if study sizes are large, confidence intervals become smaller and the heterogeneity measured using I^2 increases. The between-study variance, τ^2, ranges from 0 to infinity, and avoids this problem, but can be harder to interpret.[49]

Review consumers can, as mentioned, also informally assess statistical heterogeneity visually. A forest plot is used to display the results of a meta-analysis and should be familiar to most readers (Fig. 1). Results in all the studies that differ from each other more than would be expected by chance (95% confidence intervals more dispersed) could indicate high statistical heterogeneity (blue plot in Fig. 1). Conversely, similar results between studies (95% confidence intervals less dispersed) could indicate low statistical heterogeneity (purple plot in Fig. 1). If we observe high statistical heterogeneity, intuitively, we can be less certain where the population value lies, which can cause concerns with this domain of GRADE. It must be remembered that another strength of meta-analysis is to demonstrate consistency; so, if an intervention proves

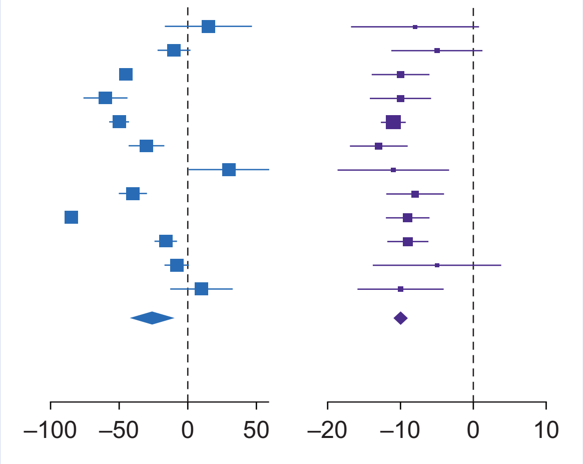

Fig 1. Forest plots of fictitious data analysed using random-effects models. Each plot shows the effect estimate scale on the x-axis (mean reductions in morphine consumption in milligrams) and the dashed vertical line a 0 mg reduction. Each square represents the effect estimate in each study with the size of the square the relative weight each study contributes to the analysis. The associated horizontal lines are the 95% confidence intervals. The diamond represents the overall effect estimate (centre of the diamond) and 95% confidence interval (width of the diamond). The blue plot shows a review of studies with high statistical heterogeneity (I^2=98%) with varying effects reflected in wider overall confidence intervals. The purple plot shows low statistical heterogeneity (I^2=0%) with studies having similar effects and a narrow overall confidence interval.

beneficial in a range of studies with low statistical heterogeneity, it helps with generalisability (purple plot in Fig. 1).

The two types of model commonly used are *fixed effect* and *random effects*. Fixed-effect models assume that there is one underlying effect to estimate, which may occur if studies have the same types of participants, same intervention, and same study design. This is rare, although may still be used by 30% of reviews in perioperative medicine.[50] However, a random-effects model assumes there is a distribution of underlying effects to estimate, which is more often the case. As fixed- and random-effects models are similar when I^2 values are small, we recommend analysing with both models in all circumstances and discussing possible discrepancies (such as small-study effects).[51,52] If the results of the two models differ, then the result of the most conservative estimate should be reported (widest confidence intervals). Alternatively, if statistical expertise is available, model fit can be assessed within a Bayesian framework using the deviance information criteria. If effect estimates from random-effects models are reported, prediction intervals can be beneficial, as they indicate the possible treatment effect in an individual setting rather than the average treatment effect.[53]

If significant statistical heterogeneity is observed, then *meta-regression*, *sensitivity*, and *subgroup analysis* should be undertaken in an attempt to understand why estimates between studies differ.[54] Meta-regression is similar to linear regression, although the data points are studies rather than participants, and they are weighted by the inverse of the standard error (generally larger studies carry more weight). It can be used to identify characteristics of each study that may improve intervention efficacy, such as dose of analgesic or weeks of exercise interventions. This can be used to inform future studies or possibly guide clinical practice, but because of the observational nature of the data (as this is a between-study analysis), they may be subject to ecological bias, where the relationship at the aggregate data level fails to reflect what is seen at the individual level.[55]

Statistical heterogeneity may also exist because of the selection of a particular effect estimate.[56] For effect estimates measured on an absolute scale (mean and risk difference), if effects vary with baseline risk (control-group risk) and baseline risk varies between trials, then statistical heterogeneity will be induced. For example, previous evidence has shown that analgesics are more effective with higher levels of pain, so in studies that have high control-group morphine consumption (higher-risk population), the effects will be greater.[57] Indeed, in the ketamine review, a sensitivity analysis was conducted in studies with higher pain levels, and found that it led to greater reductions in pain compared with the main analysis.[8] In this scenario, contemporary methods can be used that utilise meta-regression to estimate morphine reductions for different consumptions utilising local clinical data.[56] If a local average consumption is known, then the likely average effect can be calculated using a simple meta-regression equation. This principle could potentially extend to other perioperative outcomes, such as depression, chronic pain, or length of stay.

We illustrate this concept with an example from the ketamine review for acute postoperative pain.[8] The primary analysis showed a reduction in 24 h morphine consumption of 8 mg, which may not be regarded as clinically significant. However, if we perform a meta-regression with control-group morphine consumption as the predictor variable, we find that clinically significant reductions can be observed at higher average morphine consumption (50 mg). This is represented in Figure 2 by the point at which the dashed blue line intercepts the y-axis. The new method described previously avoids these incomplete conclusions about a lack of clinical significance observed in recent reviews.[58] An alternative to meta-regression may be the use of ratio measures that would also help resolve the problems with absolute scales.[59] If such statistical heterogeneity can be adequately explained as in the aforementioned example, then COE will be improved and new hypotheses generated.

Imprecision (random error)

A Type I error is a false positive, the risk of which increases when multiple SRMAs are conducted after the publication of each trial (multiple comparisons). Type II error (false negative), as in primary research studies, can occur if insufficient numbers of participants are included in a review. Although one of the benefits of SRMAs is increased power, their conduct does not ensure such power. Recent developments in methodology have allowed the conduct of *trial sequential analysis* (TSA) to counteract this increase in the Type I error rate; the approach is broadly analogous to the methodology used for interim analyses when monitoring the results of ongoing RCTs.[60] Trial sequential analysis allows authors to adjust for multiple comparisons (reduce Type I errors), identify when sufficient numbers of participants have been recruited to trials (reduce Type II errors), and identify when future trials are unlikely to result in a beneficial intervention effect (futility). When examining Cochrane reviews, TSA helped prevent 93% of false positives.[61] Specific to anaesthesia research, in a

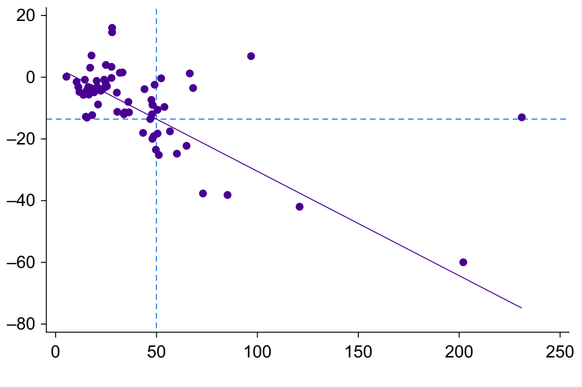

Fig 2. Reanalysis of a Cochrane review of ketamine for acute postoperative pain that showed an I^2 value of 96%, which could lead to downgrading of evidence with Grading of Recommendations, Assessment, Development and Evaluation. The x-axis is baseline morphine consumption (mean milligram of morphine consumed in control group) and the y-axis is mean reduction in morphine with ketamine. The diagonal purple line is linear prediction from the meta-regression equation and the dashed blue line represents the morphine reduction with ketamine at an average consumption of 50 mg in the control group. It can be seen that at higher consumptions of morphine, greater reductions in morphine are observed, which allow more accurate interpretation of clinical significance (approximately 30 mg reduction at consumptions of 100 mg). Explanation of heterogeneity will improve the certainty of evidence.

random sample of 50 SRMAs, only 12% had a power >80% and only 32% preserved Type I error rates.

However, TSA may be too conservative and could potentially encourage further research when little clinical equipoise exists to conduct further (resource-intensive) trials in pursuit of reaching the required information size. Other arguments against the use of TSA are that previous reviews should not influence the current review, that review authors should be able to judge uncertainty *via* confidence intervals (rather than rejecting a null hypothesis), the risk of incorrectly stopping the review process once statistical significance is reached, and limitations when between-study heterogeneity is present in the included studies.

With all this in mind, conduct of TSA (in low-risk-of-bias trials) can be considered or evaluation of the width of confidence intervals can be used to maintain COE in this domain, and both are recommended in the latest *Cochrane Handbook for Systematic Reviews of Interventions*. If 95% confidence intervals include a wide range of benefit and harm (e.g. for morphine consumption, −20 to 10 mg), this would lead to downgrading. However, it should be kept in mind that it is not possible to exclude a difference by observing confidence intervals alone. Also, another advantage of TSA is that its use can help review authors improve adherence to downgrading of imprecision as per the *Cochrane Handbook for Systematic Reviews of Interventions*.[62]

To illustrate how to assess for imprecision, Figure 3 shows an annotated TSA plot when reanalysing nausea and vomiting from the ketamine review.[8] It confirms a beneficial effect of the intervention when adjusted for multiple comparisons (crosses line B) whilst also showing the review has achieved the required number of participants (crosses line C). Alternatively, the 95% confidence interval was 0.81–0.96, which demonstrates precision in the estimates as defined previously. Therefore, no downgrading of evidence in this domain is indicated based on GRADE.

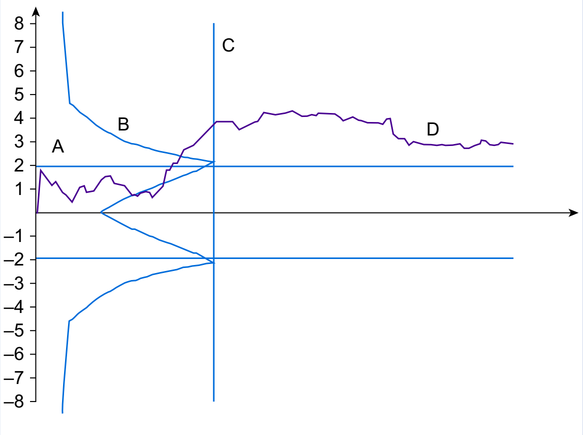

Fig 3. Annotated trial sequential analysis plot for nausea and vomiting from a published Cochrane review on ketamine. A, the conventional boundary for statistical significance (P<0.05); B, the O'Brien–Fleming boundary for statistical significance (adjusted for multiple comparisons), which requires greater Z scores (y-axis) earlier in evidence accrual; C, the required number of participants (information size on x-axis, which equals around 2000 participants) to reduce Type II errors; D, the cumulative Z score, which changes as each study is added to the meta-analysis. It crosses lines A, B, and C. The plot shows both a significant reduction in nausea and vomiting with enough participants recruited to reduce Type II errors (Z curve crossing line C). This result would result in no concerns for the imprecision domain, improving certainty of evidence.

Indirectness

As SRMAs collate results from a variety of primary studies, it is important to ensure that evidence can be directly applied to the population in question (*external validity*). For example, in reviews on postoperative pain, including trials from only dental surgery may not be applicable to more invasive surgeries,[63] or excluding participants in individual trials to whom results may be applied.[64] Commonly, participants with chronic pain are excluded from acute pain trials, so applying evidence to this group is problematic. To help with external validity, the Procedure-Specific Postoperative Pain Management group advocates procedure-specific evidence,[65] although direct evidence for such individual responses to different surgical procedures is often limited or even lacking.[56,57] However, it was found in the Cochrane review of ketamine that opioid reductions in subgroups of surgical procedures varied, with greater reductions in major orthopaedic procedures.[8]

Authors of SRMAs should use clinical judgement and consider subgroup or sensitivity analysis of similar trials to make evidence directly applicable to the population of interest. Any concerns over the applicability of evidence should result in downgrading of this GRADE domain, unless it can be demonstrated in a subgroup of the population of interest.

Publication bias

Publication bias relates to the preferential publication of 'positive' studies (often defined statistically as P<0.05), which are more likely to be published, published faster,[66] cited more often,[67] and more likely to lead to duplicate publications.[68] This can create a distortion in the literature, where positive studies (or SRMAs) are over-represented. In a study on anaesthesia SRMAs, assessment for publication bias was conducted in only 43% of reviews, and publication bias had a prevalence of up to 50–80%.[69] Another study investigated subsequent publication of abstracts in the field of anaesthesia from more than 1000 conference abstracts. Only 54% proceeded to publication with positive studies being more likely to be published after adjustment for study size and quality.[70] Furthermore, positive studies may be more likely to be published in higher-impact anaesthesia journals.[71]

With evidence of publication bias and its recognition as a cause of disagreement between large RCTs and SRMAs,[72] minimisation of publication bias is vital, although often underperformed. This should involve searching clinical trial databases and reports,[73] references and citations, conference proceedings, and grey literature sources for unpublished studies. Although this helps identify unpublished studies, it may introduce trials of a lesser quality into the review.[74] If data are missing from an identified study, contacting authors for this information is vital, although data are often difficult to obtain. In postoperative pain reviews, searching clinical trial databases, conference proceedings, and grey literature sources was performed in 16%, 9%, and 4% of reviews, respectively.[75] In reviews in pain and anaesthesia, a search of clinical

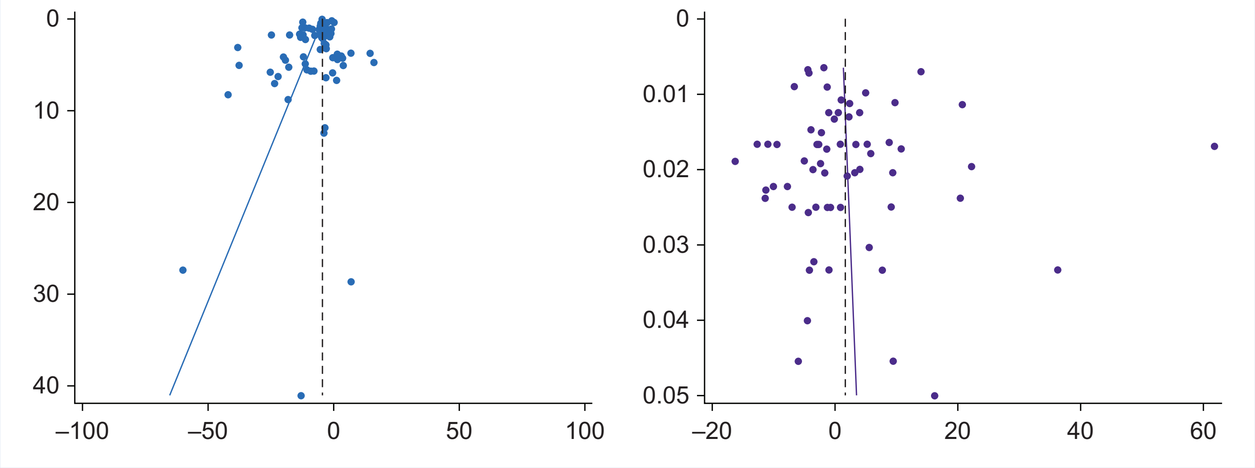

Fig 4. Reanalysis of a Cochrane review of ketamine for acute pain. Funnel plots for 24 h morphine are plotted. The left plot (blue) is the conventional funnel plot with mean difference on the x-axis and standard error (on a reverse scale) on the y-axis. The right plot (purple) is the new plot with meta-regression residuals on the x-axis and inverse sample size on the y-axis. The diagonal blue solid line is Egger's line with *P*-values in parentheses in the text below, and the diagonal purple solid line is from the meta-regression accounting for baseline risk. It can be observed that the purple plot using the updated axis guidelines is more symmetrical (*P*=0.80) than the conventional plot (*P*=0.03). Under the conventional scenario, small-study effects (possible publication bias) would be assumed when none was present, once resolving the correlation between effect estimates and standard errors. This would result in incorrect downgrading of evidence as per Grading of Recommendations, Assessment, Development and Evaluation.

trial registries was performed <50% of the time and searching for unpublished studies <10% of the time.[76] Another study found searching for unpublished studies occurred only 20% of the time.[69]

In addition to minimising publication bias, evaluation for publication bias (small-study effects) should be performed, ideally using contour-enhanced *funnel plots*.[77] Funnel plots (Fig. 4) are plots of effect estimates on the x-axis and standard errors on a reverse scale on the y-axis (so generally, larger studies towards the top of the plot). The distribution of smaller studies at the bottom should be symmetrical (random distribution). If 'negative' studies are not published, then they will be missing from one side of the plot. This could be attributable to publication bias, although other causes are described, so the term 'small-study effects' is preferred (such as more-intense interventions or lower methodological quality in smaller studies). However, as visual inspection of conventional funnel plots is unreliable,[78] this should be supplemented by statistical tests, such as Egger's linear regression, which can generate *P*-values to help identify asymmetry.[72] A higher level of statistical significance should ideally be used (*P*<0.1) because of the low power of these tests.[79]

However, recently published research has identified that for continuous outcomes dependent on baseline risk (e.g. pain and morphine consumption), Egger's test is inaccurate because of the correlation between effect estimates and standard errors, with Type I errors of 60%. An alternative test based on meta-regression residuals and inverse sample size should be considered (Fig. 4).[80] This helps reduce Type I errors to expected levels, although retains the low power of other tests. Other studies have found issues with other outcomes, where dependency between effect estimates and standard errors occurs, such as standardised mean differences,[81] odds ratios,[82] or proportion outcomes,[83] so sample-size-based tests (instead of standard errors) may perform better. Reducing these false positives by selecting the correct statistical test and minimising publication bias can help improve concerns in this domain of GRADE.

Conclusions

There are concerns with SRMAs as poor predictors of results from large, well-conducted RCTs. The number of high-certainty SRMAs in anaesthesia (and in general) may be unacceptably low because of problems with risks of bias, random error, unexplained heterogeneity, and publication bias. We have outlined methodological strategies that review authors and primary researchers can use to improve the COE in the anaesthesia literature.

Authors' contributions

Conception: BD
Data analysis: BD
Data interpretation: OM, JCJ, AJS, SF, JNL, JPW
Writing paper: all authors
Approving final version and agreement to be accountable for data: all authors

Declarations of interest

The authors declare that they have no conflicts of interest.

References

1. Ioannidis JP. The mass production of redundant, misleading, and conflicted systematic reviews and meta-analyses. *Milbank Q* 2016; **94**: 485–514

2. IntHout J, Ioannidis JP, Borm GF. Obtaining evidence by a single well-powered trial or several modestly powered trials. *Stat Methods Med Res* 2016; **25**: 538–52
3. Doleman B, Williams JP, Lund J. Why most published meta-analysis findings are false. *Tech Coloproctol* 2019; **23**: 925–8
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. Villar J, Carroli G, Belizan JM. Predictive ability of meta-analyses of randomised controlled trials. *Lancet* 1995; **345**: 772–6
6. Sivakumar H, Peyton PJ. Poor agreement in significant findings between meta-analyses and subsequent large randomized trials in perioperative medicine. *Br J Anaesth* 2016; **117**: 431–41
7. Conway A, Conway Z, Soalheira K, Sutherland J. High quality of evidence is uncommon in Cochrane systematic reviews in anaesthesia, critical care and emergency medicine. *Eur J Anaesthesiol* 2017; **34**: 808–13
8. Brinck EC, Tiippana E, Heesen M, et al. Perioperative intravenous ketamine for acute postoperative pain in adults. *Cochrane Database Syst Rev* 2018; **12**: CD012033
9. Booth A, Clarke M, Dooley G, et al. The nuts and bolts of PROSPERO: an international prospective register of systematic reviews. *Syst Rev* 2012; **1**: 1–9
10. Allers K, Hoffmann F, Mathes T, Pieper D. Systematic reviews with published protocols compared to those without: more effort, older search. *J Clin Epidemiol* 2018; **95**: 102–10
11. Page MJ, Shamseer L, Tricco AC. Registration of systematic reviews in PROSPERO: 30,000 records and counting. *Syst Rev* 2018; **7**: 32
12. Sideri S, Papageorgiou SN, Eliades T. Registration in the international prospective register of systematic reviews (PROSPERO) of systematic review protocols was associated with increased review quality. *J Clin Epidemiol* 2018; **100**: 103–10
13. Tricco AC, Cogo E, Page MJ, et al. A third of systematic reviews changed or did not specify the primary outcome: a PROSPERO register study. *J Clin Epidemiol* 2016; **79**: 46–54
14. Koensgen N, Rombey T, Allers K, Mathes T, Hoffmann F, Pieper D. Comparison of non-Cochrane systematic reviews and their published protocols: differences occurred frequently but were seldom explained. *J Clin Epidemiol* 2019; **110**: 34–41
15. Moher D, Liberati A, Tetzlaff J, Altman DG, Prisma Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *PLoS Med* 2009; **6**, e1000097
16. Panic N, Leoncini E, De Belvis G, Ricciardi W, Boccia S. Evaluation of the endorsement of the preferred reporting items for systematic reviews and meta-analysis (PRISMA) statement on the quality of published systematic review and meta-analyses. *PLoS One* 2013; **8**, e83138
17. Oh JH, Shin WJ, Park S, Chung JS. Reporting and methodologic evaluation of meta-analyses published in the anesthesia literature according to AMSTAR and PRISMA checklists: a preliminary study. *Korean J Anesthesiol* 2017; **70**: 446–55
18. Møller MH. Patient-important outcomes and core outcome sets: increased attention needed! *Br J Anaesth* 2019; **122**: 408–10
19. Myles PS, Myles DB, Galagher W, et al. Measuring acute postoperative pain using the visual analog scale: the minimal clinically important difference and patient acceptable symptom state. *Br J Anaesth* 2017; **118**: 424–9
20. Myles PS, Boney O, Botti M, et al. Systematic review and consensus definitions for the Standardised Endpoints in Perioperative Medicine (StEP) initiative: patient comfort. *Br J Anaesth* 2018; **120**: 705–11
21. Buggy DJ, Freeman J, Johnson MZ, et al. Systematic review and consensus definitions for standardised endpoints in perioperative medicine: postoperative cancer outcomes. *Br J Anaesth* 2018; **121**: 38–44
22. McIlroy DR, Bellomo R, Billings 4th FT, et al. Systematic review and consensus definitions for the Standardised Endpoints in Perioperative Medicine (StEP) initiative: renal endpoints. *Br J Anaesth* 2018; **121**: 1013–24
23. Barnes J, Hunter J, Harris S, et al. Systematic review and consensus definitions for the Standardised Endpoints in Perioperative Medicine (StEP) initiative: infection and sepsis. *Br J Anaesth* 2019; **122**: 500–8
24. Moonesinghe SR, Jackson AI, Boney O, et al. Systematic review and consensus definitions for the Standardised Endpoints in Perioperative Medicine initiative: patient-centred outcomes. *Br J Anaesth* 2019; **123**: 664–70
25. Schuster Bruce C, Brhlikova P, Heath J, McGettigan P. The use of validated and nonvalidated surrogate endpoints in two European Medicines Agency expedited approval pathways: a cross-sectional study of products authorised 2011–2018. *PLoS Med* 2019; **16**, e1002873
26. Fabritius ML, Mathiesen O, Wetterslev J, Dahl JB. Post-operative analgesia: focus has been on benefit—are we forgetting the harm? *Acta Anaesthesiol Scand* 2016; **60**: 839–41
27. Edwards JE, McQuay HJ, Moore RA, Collins SL. Reporting of adverse effects in clinical trials should be improved: lessons from acute postoperative pain. *J Pain Symptom Manage* 1999; **18**: 427–37
28. Fabritius ML, Strøm C, Koyuncu S, et al. Benefit and harm of pregabalin in acute pain treatment: a systematic review with meta-analyses and trial sequential analyses. *Br J Anaesth* 2017; **119**: 775–91
29. Mills EJ, Thorlund K, Ioannidis JP. Demystifying trial networks and network meta-analysis. *BMJ* 2013; **346**: f2914
30. Martinez V, Beloeil H, Marret E, Fletcher D, Ravaud P, Trinquart L. Non-opioid analgesics in adults after major surgery: systematic review with network meta-analysis of randomized trials. *Br J Anaesth* 2017; **118**: 22–31
31. Faltinsen EG, Storebø OJ, Jakobsen JC, Boesen K, Lange T, Gluud C. Network meta-analysis: the highest level of medical evidence? *BMJ Evid Based Med* 2018; **23**: 56–9
32. Greenfield ML, Mhyre JM, Mashour GA, Blum JM, Yen EC, Rosenberg AL. Improvement in the quality of randomized controlled trials among general anesthesiology journals 2000 to 2006: a 6-year follow-up. *Anesth Analg* 2009; **108**: 1916–21
33. Kunz R, Vist GE, Oxman AD. Randomisation to protect against selection bias in healthcare trials. *Cochrane Database Syst Rev* 2007; **2**: MR000012
34. 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

35. Nüesch E, Reichenbach S, Trelle S, et al. The importance of allocation concealment and patient blinding in osteoarthritis trials: a meta-epidemiologic study. *Arthritis Rheum* 2009; **61**: 1633–41
36. Carroll D, Tramer M, McQuay H, Nye B, Moore A. Randomization is important in studies with pain outcomes: systematic review of transcutaneous electrical nerve stimulation in acute postoperative pain. *Br J Anaesth* 1996; **77**: 798–803
37. Montori VM, Bhandari M, Devereaux PJ, Manns BJ, Ghali WA, Guyatt GH. In the dark: the reporting of blinding status in randomized controlled trials. *J Clin Epidemiol* 2002; **55**: 787–90
38. Barcot O, Boric M, Dosenovic S, Pericic TP, Cavar M, Puljak L. Risk of bias assessments for blinding of participants and personnel in Cochrane reviews were frequently inadequate. *J Clin Epidemiol* 2019; **113**: 104–13
39. Vase L, Petersen GL, Riley 3rd JL, Price DD. Factors contributing to large analgesic effects in placebo mechanism studies conducted between 2002 and 2007. *Pain* 2009; **145**: 36–44
40. Hróbjartsson A, Thomsen AS, Emanuelsson F, et al. Observer bias in randomized clinical trials with measurement scale outcomes: a systematic review of trials with both blinded and nonblinded assessors. *CMAJ* 2013; **185**: E201–11
41. Jüni P, Egger M. Commentary: empirical evidence of attrition bias in clinical trials. *Int J Epidemiol* 2005; **34**: 87–8
42. Doleman B, Heinink TP, Read DJ, Faleiro RJ, Lund JN, Williams JP. A systematic review and meta-regression analysis of prophylactic gabapentin for postoperative pain. *Anaesthesia* 2015; **70**: 1186–204
43. De Oliveira Jr GS, Jung MJ, McCarthy RJ. Discrepancies between randomized controlled trial registry entries and content of corresponding manuscripts reported in anesthesiology journals. *Anesth Analg* 2015; **121**: 1030–3
44. Jones PM, Chow JT, Arango MF, et al. Comparison of registered and reported outcomes in randomized clinical trials published in anesthesiology journals. *Anesth Analg* 2017; **125**: 1292–300
45. Sterne JA, Savović J, Page MJ, et al. RoB 2: a revised tool for assessing risk of bias in randomised trials. *BMJ* 2019; **366**: l4898
46. Detweiler BN, Kollmorgen LE, Umberham BA, Hedin RJ, Vassar BM. Risk of bias and methodological appraisal practices in systematic reviews published in anaesthetic journals: a meta-epidemiological study. *Anaesthesia* 2016; **71**: 955–68
47. Heesen M, Klimek M, Imberger G, Hoeks SE, Rossaint R, Straube S. Co-administration of dexamethasone with peripheral nerve block: intravenous vs perineural application: systematic review, meta-analysis, meta-regression and trial-sequential analysis. *Br J Anaesth* 2018; **120**: 212–27
48. Borenstein M, Higgins JP, Hedges LV, Rothstein HR. Basics of meta-analysis: I^2 is not an absolute measure of heterogeneity. *Res Synth Methods* 2017; **8**: 5–18
49. Rücker G, Schwarzer G, Carpenter JR, Schumacher M. Undue reliance on I^2 in assessing heterogeneity may mislead. *BMC Med Res Methodol* 2008; **8**: 79
50. Choi PT, Halpern SH, Malik N, Jadad AR, Tramèr MR, Walder B. Examining the evidence in anesthesia literature: a critical appraisal of systematic reviews. *Anesth Analg* 2001; **92**: 700–9
51. Borenstein M, Hedges LV, Higgins JP, Rothstein HR. A basic introduction to fixed-effect and random-effects models for meta-analysis. *Res Synth Methods* 2010; **1**: 97–111
52. Jakobsen JC, Wetterslev J, Winkel P, Lange T, Gluud C. Thresholds for statistical and clinical significance in systematic reviews with meta-analytic methods. *BMC Med Res Methodol* 2014; **14**: 120
53. Riley RD, Higgins JP, Deeks JJ. Interpretation of random effects meta-analyses. *BMJ* 2011; **342**: d549
54. Thompson SG, Higgins JP. How should meta-regression analyses be undertaken and interpreted? *Stat Med* 2002; **21**: 1559–73
55. 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**: 371–87
56. Doleman B, Sutton AJ, Sherwin M, Lund JN, Williams JP. Baseline morphine consumption may explain between-study heterogeneity in meta-analyses of adjuvant analgesics and improve precision and accuracy of effect estimates. *Anesth Analg* 2018; **126**: 648–60
57. Doleman B, Lund JN, Williams JP. Clinically significant reductions in morphine consumption need to take account of baseline risk: presentation of a novel meta-analysis methodology. *Br J Anaesth* 2018; **120**: 414–5
58. Verret M, Lauzier F, Zarychanski R, et al. Perioperative use of gabapentinoids for the management of postoperative acute pain: a systematic review and meta-analysis. *Anesthesiology* 2020; **133**: 265–79
59. Friedrich JO, Adhikari NK, Beyene J. The ratio of means method as an alternative to mean differences for analyzing continuous outcome variables in meta-analysis: a simulation study. *BMC Med Res Methodol* 2008; **8**: 32
60. Wetterslev J, Jakobsen JC, Gluud C. Trial sequential analysis in systematic reviews with meta-analysis. *BMC Med Res Methodol* 2017; **17**: 1–8
61. Imberger G, Thorlund K, Gluud C, Wetterslev J. False-positive findings in Cochrane meta-analyses with and without application of trial sequential analysis: an empirical review. *BMJ Open* 2016; **6**, e011890
62. Castellini G, Bruschettini M, Gianola S, Gluud C, Moja L. Assessing imprecision in Cochrane systematic reviews: a comparison of GRADE and trial sequential analysis. *Syst Rev* 2018; **7**: 110
63. Doleman B, Leonardi-Bee J, Heinink TP, Bhattacharjee D, Lund JN, Williams JP. Pre-emptive and preventive opioids for postoperative pain in adults undergoing all types of surgery. *Cochrane Database Syst Rev* 2018; **12**: CD012624
64. Pedersen C, Troensegaard H, Laigaard J, et al. Differences in patient characteristics and external validity of randomized clinical trials on pain management following total hip and knee arthroplasty: a systematic review. *Reg Anesth Pain Med* 2020; **45**: 709–15
65. Barazanchi AW, MacFater WS, Rahiri JL, et al. Evidence-based management of pain after laparoscopic cholecystectomy: a PROSPECT review update. *Br J Anaesth* 2018; **121**: 787–803
66. Hopewell S, Loudon K, Clarke MJ, Oxman AD, Dickersin K. Publication bias in clinical trials due to statistical significance or direction of trial results. *Cochrane Database Syst Rev* 2009; **1**: MR000006

67. Duyx B, Urlings MJ, Swaen GM, Bouter LM, Zeegers MP. Scientific citations favor positive results: a systematic review and meta-analysis. *J Clin Epidemiol* 2017; **88**: 92−101
68. Tramèr MR, Reynolds DJ, Moore RA, McQuay HJ. Impact of covert duplicate publication on meta-analysis: a case study. *BMJ* 1997; **315**: 635−40
69. Hedin RJ, Umberham BA, Detweiler BN, Kollmorgen L, Vassar M. Publication bias and nonreporting found in majority of systematic reviews and meta-analyses in anesthesiology journals. *Anesth Analg* 2016; **123**: 1018−25
70. Chong SW, Collins NF, Wu CY, Liskaser GM, Peyton PJ. The relationship between study findings and publication outcome in anesthesia research: a retrospective observational study examining publication bias. *Can J Anaesth* 2016; **63**: 682−90
71. De Oliveira Jr GS, Chang R, Kendall MC, Fitzgerald PC, McCarthy RJ. Publication bias in the anesthesiology literature. *Anesth Analg* 2012; **114**: 1042−8
72. Egger M, Smith GD, Schneider M, Minder C. Bias in meta-analysis detected by a simple, graphical test. *BMJ* 1997; **315**: 629−34
73. Mayo-Wilson E, Li T, Fusco N, et al. Cherry-picking by trialists and meta-analysts can drive conclusions about intervention efficacy. *J Clin Epidemiol* 2017; **91**: 95−110
74. Egger M, Juni P, Bartlett C, Holenstein F, Sterne J. How important are comprehensive literature searches and the assessment of trial quality in systematic reviews? Empirical study. *Health Technol Assess* 2003; **7**: 1−76
75. Doleman B, Lund JN, Williams JP. Detection and prevention of publication bias in meta-analyses of postoperative analgesics: a meta-epidemiological study. *Anaesthesia* 2017; **72**(S3): 20. https://doi.org/10.1111/anae.13974
76. Biocic M, Fidahic M, Cikes K, Puljak L. Comparison of information sources used in Cochrane and non-Cochrane systematic reviews: a case study in the field of anesthesiology and pain. *Res Synth Methods* 2019; **10**: 597−605
77. Peters JL, Sutton AJ, Jones DR, Abrams KR, Rushton L. Contour-enhanced meta-analysis funnel plots help distinguish publication bias from other causes of asymmetry. *J Clin Epidemiol* 2008; **61**: 991−6
78. Terrin N, Schmid CH, Lau J. In an empirical evaluation of the funnel plot, researchers could not visually identify publication bias. *J Clin Epidemiol* 2005; **58**: 894−901
79. Sterne JA, Sutton AJ, Ioannidis JP, et al. Recommendations for examining and interpreting funnel plot asymmetry in meta-analyses of randomised controlled trials. *BMJ* 2011; **343**: d4002
80. Doleman B, Freeman SC, Lund JN, Williams JP, Sutton AJ. Funnel plots may show asymmetry in the absence of publication bias with continuous outcomes dependent on baseline risk: presentation of a new publication bias test. *Res Synth Methods* 2020; **11**: 522−34
81. Zwetsloot PP, Van Der Naald M, Sena ES, et al. Standardized mean differences cause funnel plot distortion in publication bias assessments. *eLife* 2017; **6**: 1−20
82. Peters JL, Sutton AJ, Jones DR, Abrams KR, Rushton L. Comparison of two methods to detect publication bias in meta-analysis. *JAMA* 2006; **295**: 676−80
83. Hunter JP, Saratzis A, Sutton AJ, Boucher RH, Sayers RD, Bown MJ. In meta-analyses of proportion studies, funnel plots were found to be an inaccurate method of assessing publication bias. *J Clin Epidemiol* 2014; **67**: 897−903

Handling editor: Hugh C Hemmings Jr