Poor agreement in significant findings between meta-analyses and subsequent large randomized trials in perioperative medicine

H. Sivakumar¹ and P. J. Peyton²,³,⁴,*

¹Peter MacCallum Cancer Centre, East Melbourne, Victoria 3002, Australia, ²Department of Anaesthesia, Austin Health, Heidelberg, Victoria 3084, Australia, ³Dept of Surgery, Austin Hospital and University of Melbourne, Heidelberg, Victoria 3084, Australia, and ⁴Institute for Breathing and Sleep (IBAS), Victoria, Australia

*Corresponding author. E-mail: phil.peyton@austin.org.au

Abstract

Background: The reliability of meta-analysis (MA) in predicting the findings of subsequent large randomized controlled trials (RCTs) has not been assessed in perioperative medicine and anaesthesia.

Methods: Using Medline and PubMed, large RCTs (n≥1000) published since 2000 in the anaesthesia and perioperative medicine/critical care literature were identified. All previous MAs of RCTs investigating the same intervention and population were sourced. For all reported major morbid endpoints common to each, results (significant/non-significant P<0.05) were compared.

Results: 18 large RCTs and 44 prior MAs investigating the effects of 16 interventions were identified. Where endpoint results in the large RCTs were each compared with the single largest recent preceding MA, 35 of a total of 57 outcomes were predicted correctly by the MAs (61.4%). The odds ratio for a significant result from MA compared with the subsequent large RCT was 3.6, P=0.033 Bonferroni corrected. The positive predictive value of MAs was 22.7%; the negative predictive value was 85.7%, Kappa was 0.094 indicating slight agreement. The estimated power for each endpoint for large RCTs and MAs were similar, but the median study size for large RCTs was larger than that of the MAs, n=4,482 vs 1,389, P<0.0001.

Conclusions: There was a strong tendency towards positive findings in MA not substantiated by subsequent large RCTs, which was not attributable to differences in study power. This finding suggests caution in clinical decision-making in anaesthesia and perioperative medicine based on findings of meta-analysis.

Key words: anaesthesia; clinical trials; meta-analysis

Because of the logistic difficulties and expense of large clinical trials, most published trials in perioperative medicine and anaesthesia are small, single centre studies, underpowered to examine major morbid endpoints or mortality. This increases the risk of Type 2 error, the failure to detect a real treatment effect, and promotes reliance on surrogate endpoints of doubtful significance.¹⁻³ When attempting to determine treatment effects on clinically important outcomes, meta-analysis is often used. Pooling of study data in meta-analysis reduces Type 2 error. Quality of evidence is often ranked according to a hierarchical structure, in which meta-analysis is ranked at or near the top, and randomized trials lower down.⁴

However, there are good reasons to question the reliability of meta-analysis. The potential weaknesses of meta-analysis have been pointed out by a number of commentators, and include heterogeneity of studies and positive publication bias, which increases the risk of Type 1 error, the finding of a treatment effect which is not real.⁵⁻¹² This question has been explored by previous
authors by examining the diagnostic performance of meta-analyses in predicting the findings of subsequent large RCTs. Le Lorier and colleagues5 in 1997 compared statistically significant findings of large RCTs in the medical literature, with those of previously published meta-analyses, and found limited predictive value of meta-analysis for findings of the subsequent large RCT. Other authors have found better agreement, depending on the method of comparison used.6,7

There has been a substantial increase in the number of large RCTs and meta-analyses conducted in perioperative medicine and critical care over the last two decades. We therefore sought to determine the predictive ability of meta-analysis relative to subsequent large RCTs in this field. In similar fashion to previous authors in other fields,5,9,7 we assessed agreement between large RCTs and prior meta-analyses of RCTs in finding of statistically significant treatment effects on major morbid endpoints. In addition, as an alternative measure of agreement, we determined how often differences in the point estimate for treatment effect (risk ratio) on these endpoints obtained by large RCTs and prior meta-analyses were statistically significant. Study quality in meta-analysis and study power may be important factors determining findings and agreement between meta-analysis and large RCTs. We therefore also examined the relationship of significant findings to study power, by comparing study size and estimated study power for each endpoint between meta-analysis and subsequent large RCTs. Heterogeneity and risk of publication bias are commonly reported as indicators of quality in meta-analysis. The influence of these on agreement with subsequent large RCTs was also examined in secondary analyses, where MAs were excluded where evidence of absence of heterogeneity or publication bias was not provided.

Methods

Search strategy and study endpoint selection

Large RCTs were identified via computer search of Medline/PubMed, with the initial search terms of ‘anaesthesia’ and ‘perioperative medicine’, and limits to search of ‘randomized controlled trial’, ‘multicentre’ and ‘published 2000–2014’. Included studies met all of the following criteria: a) randomized controlled trial study design, b) published between 2000 and 2014 (inclusive), c) ≥1000 study subjects, and d) investigating the effect of a clinical intervention on one or more morbid clinical endpoints including mortality, with dichotomous outcome results expressed as either risk ratio (RR), hazard ratio (HR), or odds ratio (OR). Studies performed in the critical care setting were included if a majority of patients in the study were surgical patients. The search using this protocol was last run on 5th December 2014.

Thereafter, any relevant meta-analysis (MA) that preceded each large RCT was identified using the bibliography in the RCT publication and via Medline and PubMed search. Eligible meta-analyses (MAs) met all of the following: (a) MA of randomized controlled trials, (b) investigating a similar clinical intervention, (c) in a similar patient population, and d) examining one or more similar major morbid clinical endpoints including mortality, with measurement of dichotomous outcomes expressed as either RR, HR or OR. Many RCTs use composite endpoints (for example, mortality and one or more of several major morbid endpoints such as a cardiovascular, respiratory or septic complications) as their primary endpoints, which are unsuitable for comparison with other trials or MAs, and were not used in the current study. Endpoints were considered eligible for inclusion by us if they were individual major perioperative outcomes, regardless of whether they were primary or secondary endpoints in either the large RCTs or MAs. These were mortality and major perioperative morbidity including myocardial infarction, arrhythmia, deep vein thrombosis or pulmonary embolism, stroke, surgical site infection or sepsis, postoperative haemorrhage requiring reoperation, acute kidney injury or need for renal replacement therapy, gastro-intestinal bleeding, pneumonia and intra-operative awareness under general anaesthesia. Post-operative nausea and vomiting was not included a priori because it represents a substantially lower level of serious patient harm.

Clinical interventions were included where they were generally similar between the large RCTs and prior MAs. For example randomized drug trials, being prospective studies, usually study a specific drug. In contrast, to maximize statistical power and generalizability of findings, MAs will frequently include a number of trials of a generic class of drugs (e.g. beta-blockers, or steroids). We stipulated similar route of administration, generic drug class and duration and potency of dosage, for comparison to be done. For example, studies involving epidural local anaesthetics and opioids, antiplatelet agents and heparin, and different types of colloid solutions such as albumin and starch polymer solutions, were considered different interventions by us and therefore unsuitable for pooling or comparison, despite the fact that these studies may have been examining the same clinical endpoints in similar populations.

Those large RCTs with no prior published eligible MA examining similar endpoints to these criteria were not eligible for inclusion. Where multiple MAs were found by a given author (s) on the same topic with the same literature base (such as occurs, for example, where a MA is published as a Cochrane Collaboration review and also in a journal), only the latest version was included. The selection of included studies and endpoints was consistent with the four phases stipulated in QUADAS (Quality Assessment of Diagnostic Accuracy Studies) guidelines (Review Question definition and Tailoring, Flow diagram presentation and Bias and Applicability).13 To minimize bias, both authors independently reviewed each large RCT and corresponding MAs, and differences between them in either inclusion or adjudication of endpoints were then reviewed. Concordance between the two authors of the current study (HS and PP) in assessing both suitability for inclusion and agreement with the prior MAs, was calculated using the Kappa statistic.

Endpoint comparisons

Analysis (A): for descriptive data analysis, comparison was undertaken of the reported effect of identified interventions on all primary and secondary eligible endpoints which were common to each large RCT and each preceding MA. For each endpoint, results of prior MAs were compared with the large RCTs as to whether a significant or non-significant result was found at the conventional level of statistical significance, P<0.05. For significant results, the direction of treatment effect was also noted. For endpoints with an incidence of less than 10%, RR and OR were considered comparable. An additional 3-way analysis was made, classifying study results for each endpoint into ‘positive’ (statistically significant benefit), non-significant and ‘negative’ (statistically significant harm) treatment effect. In addition, the OR or RR and 95% confidence intervals were recorded for treatment effect on each endpoint.

Analysis (B): for comparative statistical analysis, the above process was repeated comparing endpoints from the large RCTs
with those of only the most recent MA studying that endpoint (where multiple recent MAs were published within a two year period, the largest was used). This eliminated Type 1 error intrinsic to our analysis from inclusion of multiple MAs incorporating the same studies. Residual Type 1 error from the presence of multiple endpoints within each RCT (which would be expected to be partially interdependent) was dealt with by Bonferroni correction for each endpoint comparison, with the P value adjusted accordingly. This was considered likely to provide a conservative estimate of statistical significance of differences between large RCTs and the corresponding recent MA for each endpoint compared.

**Study power and quality**

The relative study quality and power of MAs and RCTs were considered. Unlike large RCTs, MAs generally do not conduct an assessment of overall study power. For all the endpoint comparisons identified, the study size for large RCTs and MAs was compared, and the endpoint event rate in the control group of each, where reported. From these data, the power of each study for demonstrating an arbitrary 25% reduction (a typical value nominated in perioperative trials) in the incidence of the endpoint by the intervention was calculated.

Heterogeneity and publication bias in meta-analysis have been shown to contribute to discrepancies in findings between large RCTs and MAs of small studies. Assessment of the influence on our findings of heterogeneity within the included MAs was done by us, as an indicator of MA quality. This took the form of a sensitivity test of statistical significance (Analysis C), including only endpoints from those MAs where statistical heterogeneity was explicitly reported by the authors of those MAs to be non-significant (P≥0.05). A similar sensitivity test (Analysis D) of the influence of publication bias within the included MAs was also done, by including only endpoints from those MAs where data from funnel plots were explicitly reported by the authors of those MAs to be non-significant (P≥0.05).

Figure 1 shows the flow chart for the study protocol and selection process as recommended under QUADAS guidelines. Statistical calculations were made using STATA 12 (Stata Corp, TX, USA) and Microsoft Excel (Microsoft Corp, USA). Forest plots were created using GraphPad (GraphPad Software, Inc. CA, USA).

**Statistics**

**Descriptive data analysis**

(A) Where results from the large RCTs were compared with all preceding MAs of randomized trials, data for all endpoint comparisons were pooled and a two-by-two table was constructed to calculate positive and negative predictive value (NPV), sensitivity and specificity of MAs for predicting outcomes of subsequent large RCTs. This was repeated for all endpoint comparisons, but were in no instance both opposite in direction and statistically significant for both MAs and large RCT. The null hypothesis for this test was that there would be no significant difference between the most recent MAs and subsequent large RCTs, in the distribution of significant and non-significant results for the endpoints identified. This was also done for the sensitivity Analysis (C) and for Analysis (D), comparing endpoints from the large RCTs with only those from MAs where statistical heterogeneity and publication bias were absent, as described above.

In analysis (B), from the log point estimate (OR or RR) and 95% confidence intervals reported for treatment effect on each endpoint, the incidence of a statistically significant difference in treatment effect (two-way P-value<0.05) between the most recent MA (using random effects modelling where provided) and corresponding large RCT was calculated using tables of the z-statistic. The median value of the log point estimates of the MAs and subsequent large RCTs were compared using the Wilcoxon rank sum (Mann-Whitney) U-test.

The study size and power the large RCTs and MAs in the endpoint comparisons in analysis (B) were compared using the Student’s t-test for normally distributed data or the Wilcoxon rank sum test for non-normal data.

**Results**

18 large RCTs with at least one preceding MA for an eligible endpoint comparison were identified, examining the effects of 16 clinical interventions on major morbid endpoints; 49 preceding MAs of examining these interventions and endpoint outcomes were found (Fig. 1). Among the remaining total of 44 eligible MAs, the Kappa for initial agreement between the authors for inclusion or exclusion of endpoints common to the large RCTs was 0.96, and achieved unity after joint review.

Analysis (A): 137 primary and secondary endpoint comparisons with the corresponding large RCTs were able to be made (for six endpoints in three large RCTs, hazard ratios were reported but ORs were calculated from data in tables for the purposes of this analysis). Supplementary Table S1 shows these data and Fig. 2 displays these comparisons in the form of a forest plot for each intervention. The positive predictive value (PPV) of MAs was 18.6%; the NPV was 88.3%. 91 of 137 results for endpoints in large RCTs were predicted correctly by preceding MAs (66.4%), but this agreement consisted heavily of prediction of non-significant treatment effects. These data are shown in Table 1. Point estimates for the effect of interventions in MAs and subsequent large RCTs were in the same direction in 83 out of 137 endpoint comparisons, but were in no instance both opposite in direction and statistically significant for both MAs and large RCT.

Analysis (B): where results of endpoints in the 18 large RCTs were each compared with the single largest recent preceding MA, 35 of a total of 57 endpoint outcomes were predicted correctly by the MAs (61.4%). These are indicated in Supplementary Table S1. The PPV of MAs was 22.7%; the NPV was 85.7% (Table 2). The Kappa statistic for agreement between MAs and subsequent large RCTs was 0.094, indicating slight agreement according to the criteria of Landis and Koch. However, interpretation of the Kappa is difficult because they included data from non-randomized studies. McNemar’s test statistic was 6.55, P=0.010. This was statistically significant after Bonferroni correction for multiple endpoint comparisons (P=0.033).

Analysis (C): in this sensitivity analysis, with exclusion of endpoints from MAs that did not report lack of heterogeneity,
31 of a total of 46 endpoint outcomes were predicted correctly by the MAs (67.4%). These are indicated in Supplementary Table S1. The PPV of MAs was 27.8%; the NPV was 92.9%. The Kappa statistic was 0.232, indicating slight agreement. McNemar’s test statistic was 8.07, with an exact significance $P=0.007$. This was statistically significant after Bonferroni correction for multiple endpoint comparisons ($P=0.021$).

Analysis (D): in this sensitivity analysis, with exclusion of endpoints from MAs that did not provide data showing lack of evidence of publication bias, 19 of a total of 31 endpoint outcomes were predicted correctly by the MAs (61.3%). These are indicated in Supplementary Table S1. The PPV of MAs was 0%; the NPV was 90.5%. The Kappa statistic was -0.12, indicating no agreement. McNemar’s test statistic was 5.33, with an exact significance $P=0.021$. This was not statistically significant after Bonferroni correction for multiple endpoint comparisons ($P=0.066$).

In Analysis (B), the median value for the point estimate of the log treatment effect was further from zero in the MAs than the large RCTs ($\text{OR natural log median} [\text{IQR}] = -0.25 [-0.51 to 0.049]$).

Fig 1 Study method flow diagram. RCT, Randomized controlled trial; MA, Meta-analysis; MAs, Meta-analyses.
**Fig 2** Forest plots comparing the findings of large randomized controlled trials and those of all prior meta-analyses. Precise data are given in Supplementary Table S1. The findings of large randomized controlled trials (RCTs) and those of all prior meta-analyses for the effect of each clinical intervention vs control on major morbid endpoints identified in the study. These endpoints are listed on the vertical axis along with the corresponding study references. The horizontal axis is odds ratio (less than one favours the intervention effect). Square boxes represent point estimates for treatment effect in large RCTs; circle points represent control; error bars represent confidence intervals. The most recent meta-analyses (as defined in Methods), which were included in the comparative statistical analysis (Analysis B), are indicated with an asterisk *. Those endpoints where a statistically significant difference in the point estimate was found between large RCTs and the most recent MA, are indicated by a cross. GA, general anaesthesia; BIS, bispectral index; HES, hydroxyethyl starch; PPF, plasma protein fraction; FIO2, inspired oxygen concentration; PAC, pulmonary artery catheter; DVT, deep venous thrombosis; PE, pulmonary embolism; ORTHO, orthopaedic surgery.
vs $-0.04 [-0.25$ to 0.11], $z=2.276$, $P=0.023$ on the Wilcoxon rank sum test). However, the position of the point estimate was significantly different between MAs and large RCTs in only eight out of 57 (14.0%) of the endpoint comparisons.

For the three-way analysis, classifying study results for each endpoint into 'positive', non-significant and 'negative' treatment effects, as the result of the presence of several empty cells, no statistical analysis was possible.
Study size and power

The median study size (interquartile range) for the endpoints in the large RCTs was n=4,482 (1,994–5,228). This was more than three times larger than that of the MAs, which was n=1389 (641–2851) (\(P<0.0001\) on the Wilcoxon rank sum test for non-normally distributed data). Median (IQR) study power to demonstrate a 25% reduction from the endpoint event rate in the control group by the intervention was 0.190 (0.128–0.437) in the large RCTs, and 0.222 (0.087–0.422) in the prior MAs, (\(P=0.75\) on the Wilcoxon rank sum test). This is summarized in Table 3.

Discussion

This study found that meta-analysis was several times more likely to find a significant treatment effect than subsequent large

<table>
<thead>
<tr>
<th>Meta-analyses</th>
<th>Large RCTs</th>
<th>Non-significant</th>
</tr>
</thead>
<tbody>
<tr>
<td>Significant</td>
<td>8</td>
<td>35</td>
</tr>
<tr>
<td>Non-significant</td>
<td>11</td>
<td>83</td>
</tr>
<tr>
<td>PPV=18.6%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NPV=88.3%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sensitivity=42.1%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Specificity=70.3%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>LR (+)=1.42</td>
<td></td>
<td></td>
</tr>
<tr>
<td>LR (-)=0.82</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Odds ratio=1.72</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ROC area=0.56</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2 Statistical analysis (Analysis B) of endpoint comparisons between large randomized controlled trials (RCTs) and the most recent preceding meta-analysis of randomized trials examining those endpoints, where study findings were classified as ‘significant’ or ‘non-significant’. PPV, Positive predictive value; NPV, Negative predictive value; LR, Likelihood ratio; ROC, Receiver Operator Characteristic

<table>
<thead>
<tr>
<th>Meta-analyses</th>
<th>Large RCTs</th>
<th>Non-significant</th>
</tr>
</thead>
<tbody>
<tr>
<td>Significant</td>
<td>5</td>
<td>17</td>
</tr>
<tr>
<td>Non-significant</td>
<td>5</td>
<td>30</td>
</tr>
<tr>
<td>PPV=22.7%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NPV=85.7%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sensitivity=50.0%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Specificity=63.8%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>LR (+)=1.38</td>
<td></td>
<td></td>
</tr>
<tr>
<td>LR (-)=0.78</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Odds ratio=1.76</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ROC area=0.57</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kappa=0.094</td>
<td></td>
<td></td>
</tr>
<tr>
<td>McNemar’s (\chi^2) test statistic 6.55, (P=0.033) (Bonferroni corrected)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 3 Study size and event rates. Study size (intervention and control groups combined) and study power of the large randomized controlled trials (RCTs) vs preceding meta-analyses of randomized trials in the 57 endpoint comparisons in Analysis (B). IQR, interquartile range; Wilcoxon, Wilcoxon rank sum test

<table>
<thead>
<tr>
<th>Study size: median (IQR)</th>
<th>(P) value (Wilcoxon)</th>
<th>Study power: median (IQR)</th>
<th>(P) value (Wilcoxon)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Large RCTs</td>
<td>4842 (1994–5228)</td>
<td>0.0001</td>
<td>0.190 (0.128–0.437)</td>
</tr>
<tr>
<td>Meta-analyses</td>
<td>1389 (641–2851)</td>
<td>0.222 (0.087–0.422)</td>
<td>0.75</td>
</tr>
</tbody>
</table>
MAs. However, Type 2 error in the large RCTs does not explain the greater tendency of MAs to find significant treatment effects. In contrast, Type 2 error is a likely explanation for the high NPV we demonstrated, because both the large RCTs and the MAs were underpowered to find a treatment effect for all but a few endpoints.

There are a number of possible contributing factors to the greater tendency we found of MAs to find significant treatment effects. The study of multiple secondary endpoints of component trials in meta-analysis increases the risk of Type 1 error, the chance finding of a treatment effect that is not real. Statistical correction for multiple endpoint testing is rarely done in meta-analysis, possibly because of the heterogeneity and complexity of the raw data being collated. More sophisticated techniques such as trial sequential analysis and Bayesian methods may strengthen the robustness of meta-analysis in avoiding these sources of statistical error in future research.12

We also found it was not possible in our analysis to focus on comparison of primary endpoints between large RCTs and MAs, because many large RCTs used a composite morbidity endpoint (typically mortality and several major morbidities combined) as their primary outcome measure, which is rarely used in meta-analysis. The need to include secondary study endpoints increased the risk of Type 1 error in our review, because of some interdependence between the outcomes represented by these endpoints. However the Bonferroni correction we applied for these multiple endpoint comparisons, is likely to provide a highly conservative estimate of statistical significance for the findings in our primary analysis.31 82

The relative strengths and weaknesses of large RCTs and MAs are a complex background to this study. Large RCTs, often referred to as pragmatic or effectiveness trials, are usually multicentre in design, often with a broad surgical population cross section as the study sample. Bias and Type 1 and 2 error are minimized by appropriate power, blinding and research governance, including publication expectations, and reliance on stochastic principles to ensure comparability between study arms. In contrast, MAs consist of smaller single centre studies or ‘efficacy’ studies, with the availability of tighter control of confounding factors in the study sample and protocol, but greater vulnerability to Type 2 error, and also to publication bias that promotes Type 1 error in subsequent MA. Both effectiveness and efficacy trials may test a narrow research question and intervention which limits generalizability of their findings to clinical practice.

Other factors may contribute to potential bias in the conduct and publication of the component trials in MA. Interventions in perioperative patient care are often unable to be delivered fully double blinded, which increases the risk of observer bias. Positive publication bias has been demonstrated in the medical literature by many authors, including in perioperative medicine. De Oliveira and colleagues83 showed that the odds ratio was 2.3 for publication of studies with positive vs negative findings in the four highest impact anaesthesiology journals compared with the remainder. While this may reflect bias during submission and review for publication, positive publication bias has origins at many points in the research process, with researchers less likely to complete and submit studies without positive findings.84 85

Testing for evidence of this bias is now routine best practice in meta-analysis, such as the provision of ‘funnel plots’ which look for a relationship between the magnitude of the treatment effect and the power of included studies.86 87 However, because of a lack of power in these tests, bias (including publication bias) cannot be excluded, even where the results of these tests are not significant.88 89 In analyses of the Cochrane Database of Systematic Reviews, Sutton and colleagues90 and more recently Kicinski and colleagues92 have shown evidence of significant residual publication bias. Egger93 asserts that failure to recognize publication bias has resulted in acceptance of treatments subsequently shown to be ineffective or harmful. A recent review of publication outcomes of research presented over four years, at a major international anaesthesiaology meeting, found a risk ratio of 1.42 for eventual full journal publication of studies with positive findings as opposed to those with non-significant findings.93 Our data could not exclude publication bias in meta-analysis as a contributing factor to our findings. Indeed, the difference in incidence of positive findings we found between large RCTs and prior MAs in our Analysis (B), failed to achieve statistical significance in our Analysis (D), with exclusion of endpoints from those MAs where funnel plot data was not provided showing a lack of Type 1 error, consistent with publication bias.

There are limitations to the comparison of results from MA and RCTs, which have been previously explored by Ioannidis and colleagues90 and which should be considered when evaluating our findings. These include heterogeneity of study protocols, treatment effects, study power and quality in both RCTs and studies included in MAs, and the use of fixed vs random effects statistical models in MAs. In an alternative approach which avoids this, Pereira and Ioannidis94 compared the results of published MAs in medicine with those of updated versions from 2005–10. They found the earlier MAs were smaller but had inflated treatment effects and a false positive rate for statistically significant findings of 16–37%, when tested against the later versions. These findings are qualitatively similar to those found by us and others comparing MAs with RCTs. However, our data demonstrate more dramatic difference in perioperative medicine.

Heterogeneity was examined and interpreted variably by the authors of MAs captured in our review. For example, where heterogeneity was not shown to be present, some authors then simply used a fixed effects model for obtaining a point estimate and confidence intervals for the endpoint(s) being examined, rather than a random effects model which was less likely to find a significant difference. Heterogeneity exists within our own study. There were inevitable variations in medication dosing, or in diagnostic criteria, which complicated comparison of MAs with subsequent large RCTs. We attempted to minimize these confounders by selecting those patient populations and interventions in published MAs that best matched those of the large RCT. For instance, many MAs pooled data from placebo controlled randomized trials with randomized trials that were not double blinded. To maximize study power, most MAs pooled data from trials on multiple drugs of a given class when examining treatment effect (e.g. beta-blockers, steroids, alpha-2 adrenergic agonists, and antiplatelet agents), whereas the large RCTs tested only one drug from each class (e.g. metoprolol, dexa-methasone, clonidine and aspirin). There was some overlap in definition of endpoints reported by trials (e.g. renal failure and need for dialysis in the examination of the effects of resuscitation with hydroxy-ethyl starch solution on patient outcomes).38 40 Incremental improvements in patient care over time may bias our analysis when comparing MAs with later large RCTs. There was one instance where a negative finding in a large RCT could be attributed to lower study power compared with the corresponding MA, namely comparison of Devereaux and colleagues98 2014 with the Antiplatelet Trialists Collaboration69 1994, regarding the effect of antiplatelet agents on deep venous thrombosis (DVT). The baseline incidence of DVT was substantially lower in
Devereaux and colleagues, because of less sensitive diagnostic criteria and possibly also the more widespread routine use of heparin thromboprophylaxis in recent years, than in the time covered by the earlier MA. However, exclusion of this endpoint did not change the statistical significance of the findings of our primary sensitivity analysis.

Conclusions
Many commentators caution that the results of meta-analysis should be seen only as ‘hypothesis-forming’. However, our data suggest that significant findings of meta-analysis are poor indicators of the outcome a subsequent large RCT in the field of perioperative medicine. Our results therefore point to the continued need for large outcome trials to better inform clinical decision making on interventions influencing important patient outcomes in perioperative medicine, critical care and anaesthesia.\(^3\) 95–97

Authors’ contributions
Study design: H.S., P.J.P.
Study conduct: H.S., P.J.P.
Data analysis: H.S., P.J.P.
Writing paper: H.S., P.J.P.
Revising paper: all authors

Supplementary material
Supplementary material is available at British Journal of Anaesthesia online.

Acknowledgements
The authors are grateful to Mr. Stephen van der Hoorn and Ms. Rachel Sore of the Department of Statistics, University of Melbourne for their assistance with statistical analysis, and to Dr. Nina Parthasarathy for her help with database compilation.

Declaration of interest
None declared.

Funding
The work was supported by funding obtained from internal departmental sources.

References


57. Moller CH, Penninga L, Wetterslev J, Steinbruchel DA, Glud C. Clinical outcomes in randomized trials of off- vs. on-pump coronary artery bypass surgery: systematic review with meta-analyses and trial sequential analyses. Eur Heart J 2008; 29: 2601–16


80. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977; 33: 159–74


94. Pereira TV, Ioannidis JPA. Statistically significant meta-analyses of clinical trials have modest credibility and inflated effects. *J Clin Epidemiol* 2011; 64: 1060–9

95. Hennekens CH, Demets D. The need for large-scale randomized evidence without undue emphasis on small trials, meta-analyses, or subgroup analyses. *JAMA* 2009; 302: 2361–2

96. Devereaux PJ, Chan MT, Eisenach J, Schricker T, Sessler DI. The need for large clinical studies in perioperative medicine. *Anesthesiology* 2012; 116: 1169–75


Handling editor: J. G. Hardman