BJA

British Journal of Anaesthesia, 117 (4): 431–41 (2016)

doi: 10.1093/bja/aew170
Review Article

REVIEW ARTICLE

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

H. Sivakumar[1] and P. J. Peyton[2,3,4,*]

[1]Peter MacCallum Cancer Centre, East Melbourne, Victoria 3002, Australia, [2]Department of Anaesthesia, Austin Health, Heidelberg, Victoria 3084, Australia, [3]Dept of Surgery, Austin Hospital and University of Melbourne, Heidelberg, Victoria 3084, Australia, and [4]Institute 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 ($n \geq 1000$) 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. LeLorier and colleagues[5] 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,[5–7] 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) $n \geq 1000$ study subjects, and (d) investigating the effect of a clinical intervention on one or more major 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. Postoperative 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 generically 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 according 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 each endpoint in the RCTs, regarding 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