Techniques in Coloproctology (2019) 23:925–928
https://doi.org/10.1007/s10151-019-02020-y

CONTROVERSIES IN COLORECTAL SURGERY

Why most published meta-analysis findings are false

B. Doleman[1] · J. P. Williams[1] · J. Lund[1]

Received: 27 May 2019 / Accepted: 17 June 2019 / Published online: 25 June 2019

Introduction

Almost a decade after the controversial paper 'why most published research findings are false' was written [1], we re-visit this concern with regard to published meta-analyses. Although reading the title of this article may make some pause for thought (or vehemently disagree), if you simply subscribe to the findings of Ioannidis's paper then the logical conclusion is that the meta-analyses of primary studies are equally susceptible to being false. Indeed, it was asserted in the paper that meta-analyses of small, inconclusive studies (very common) are probably false [1]. But how can this be when meta-analyses sit unchallenged at the top of the hierarchy of evidence (despite being retrospective and observational in nature)? We will first examine the evidence for our assertion, and then provide reasons why meta-analyses are poor predictors of results from large trials gained from the authors' experience in perioperative meta-analyses.

Meta-analyses as predictors of findings from large randomised controlled trials

For reasons described later, large (adequately powered and multiplicity adjusted) and low risk of bias randomised controlled trials (RCTs) can be considered the gold standard to evaluate the effect of interventions in medicine. We can compare the outcomes of these high-quality trials with prior meta-analyses on the same subject and use conventional statistics such as positive (PPV) and negative predictive values (NPV) to quantify how good meta-analyses are at predicting the outcome of the gold standard RCTs.

Using these statistics, one study compared 12 large RCTs of medical interventions with 19 meta-analyses addressing the same question and found a PPV of 68% and NPV of 67% [2]. Another study found PPVs of < 67%. More specific to surgery, a study including perioperative interventions which included 18 RCTs (compared with the most recent meta-analysis) and 57 endpoints (22 significant in the meta-analysis with 5 of these significant in the RCT) found a PPV of only 23% although the NPV was higher at 86% [3]. The corresponding area under the receiver-operating characteristic curve was 0.57 (similar to a coin toss) [3].

Therefore, even if there is a 'positive' meta-analysis, the probability of a 'positive' finding in a subsequent large RCT is unacceptably low, particularly in the field of perioperative interventions, where a positive result in a subsequent RCT becomes unlikely (depending on prevalence). Despite the faith we have in meta-analysis as the last word in truth, they would predict the same outcome in a subsequent, definitive RCT (the 'true' answer) similar to tossing a coin. Some may be perplexed at these findings. We will now examine possible reasons for these discrepancies.

Heterogeneity

Heterogeneity describes the differences in characteristics between the included studies such as differences in inclusion/exclusion criteria and study conduct. Statistical heterogeneity occurs when estimates from each study vary by more than would be expected by chance and can be quantified by statistics such as I^2 (the percentage of variability between studies rather than chance variability). To understand this concept, consider the following example evaluating mortality after surgery. One meta-analysis (Fig. 1) includes two trials which were conducted in the same population of surgical participants, with the same surgeon and the same methodology. Consequently, the results are similar, so we can be more confident the 'true' result is similar to that obtained in the meta-analysis. Now consider a second example (Fig. 2), the two trials in this meta-analysis both look at the same intervention but are conducted in different patients (Doleman 2019 includes emergency patients) and have different

✉ B. Doleman
dr.doleman@gmail.com

[1] Department of Surgery and Anaesthesia, University of Nottingham, Graduate Entry Medicine, Royal Derby Hospital, Derby, UK

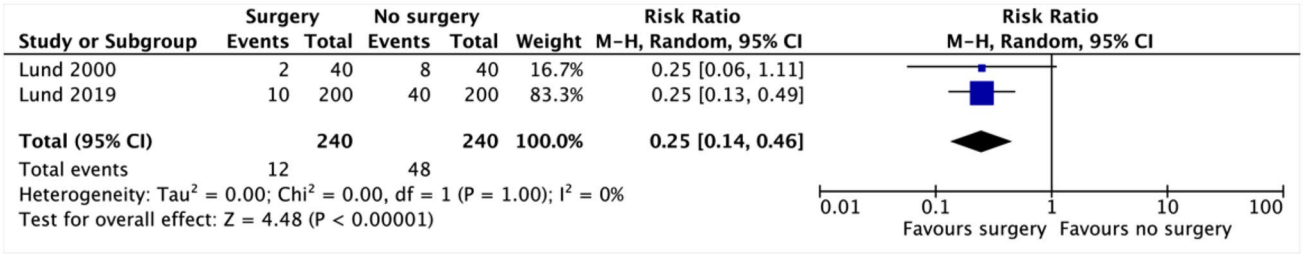

Fig. 1 Forest plot showing results from two trials demonstrating low statistical heterogeneity. This can be observed from overlap of confidence intervals and, therefore, a narrow confidence interval overall on random effects analysis (black diamond)

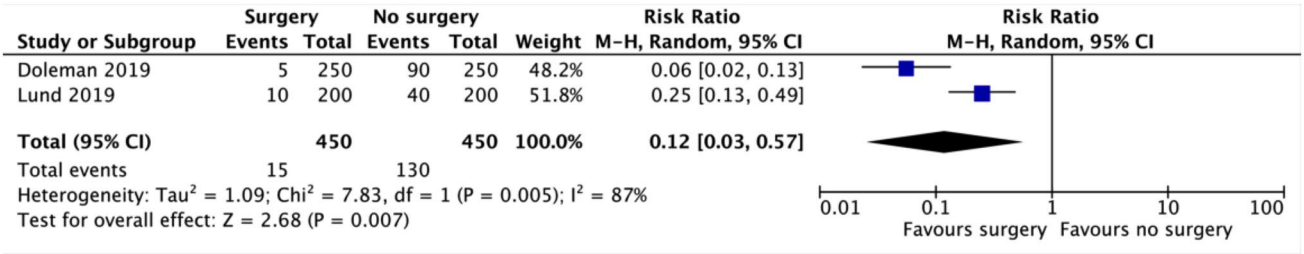

Fig. 2 Forest plot showing results from two trials demonstrating high statistical heterogeneity. This can be observed from lack of overlap of confidence intervals and, therefore, a wide confidence interval overall on random effects analysis (black diamond)

methodology (Doleman 2019 uses unblinded assessors). This leads to different results meaning we are more uncertain where the 'true' effect lies.

Heterogeneity is often used by review authors to inform the choice of statistical model used to aggregate results. However, the model should be selected on the basis of the assumptions of the model rather than any particular I^2 value. The fixed effect model assumes one underlying effect to estimate. For example, if a researcher conducted two RCTs using the same population (similar age, gender and intervention) from their hospital using standardised methodology and a meta-analysis was conducted on these two studies, then the fixed effect model would be reasonable (Fig. 1). However, nearly all meta-analyses include studies conducted in different populations and we would, therefore, argue that this assumption is rarely satisfied.

This leaves us with the more appropriate random effects model. This model assumes the true effect size may differ from study to study, as would be expected with different populations (Fig. 2). However, with this model, as statistical heterogeneity is incorporated into the calculation (increasing heterogeneity increases imprecision) as it assumes different underlying effects to estimate, this can lead to imprecision when using a random effects model (see width of diamond showing wider confidence intervals in Fig. 2 vs. Fig. 1).

Despite this, random effects models do not eliminate heterogeneity (as seen in many forest plots and high I^2 values in many meta-analyses). If heterogeneity is observed it can be used to generate new hypotheses on where interventions may be more effective: an under-recognised strength of meta-analyses [4]. From our fictitious examples of surgery on mortality (Fig. 2), Doleman 2019 includes emergency patients and, therefore, the intervention may be more effective in this group.

Publication bias

Publication bias is the preferential publication of 'positive' trials. Positive trials are both more likely to be published and published faster than trials with negative findings. Consider an example: a surgical researcher conducts a study on the effects of laparoscopic surgery on length of stay. Following completion, the results show no difference between the laparoscopic and the open surgery group. The researcher, therefore, does not submit the study for publication and files it away. Alternatively, the researcher submits to three journals who serially reject it as the results are not deemed 'interesting', taking several months for the manuscript to make it through the review process at each journal. This could lead to the study either not being published or being published later than if the study had positive results. Possible publication bias (or more correctly, imprecise study effects) can be observed in an asymmetrical funnel plot, which shows the relationship between the effect estimate and standard error (generally, the larger the study size, the smaller the standard error), with negative studies seen to be missing on the bottom right of the plot (Fig. 3).

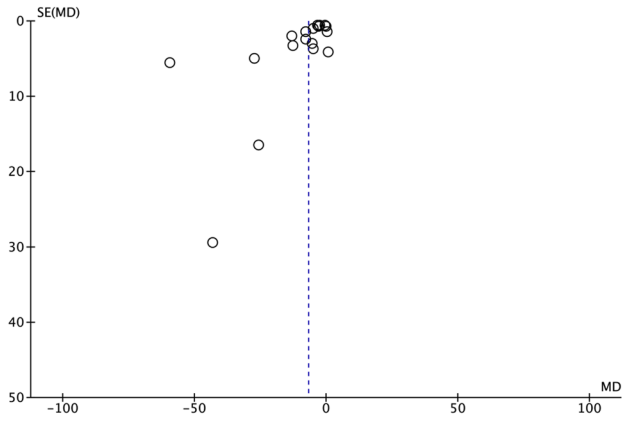

Fig. 3 Funnel plot showing possible publication bias (imprecise study effects) with larger studies towards the top and smaller studies at the bottom. With no publication bias, smaller negative studies should be present in the bottom right of the plot giving an inverted funnel appearance. However, these studies are missing, suggesting possible publication bias

The incidence of publication bias may be as high as 50–80% [5]. Although it is important for reviews to evaluate for potential publication bias, it is vital that reviews attempt to prevent it by searching for unpublished data from clinical trial databases and grey literature searches. However, thorough search strategies are only seen in around 20% of published reviews. Therefore, if the majority of published reviews are based on a set of studies which are susceptible to publication bias, then the results from these reviews will be biased towards 'positive' effects of interventions, which may partly explain the poor PPV found in previous studies [2, 3]. Similarly, the issue of publication bias may extend to reviews themselves, with review authors either not submitting negative reviews or journals less likely to publish them.

Error

One of the benefits of conducting a meta-analysis of smaller RCTs is to improve the power of the collected data, helping to identify beneficial interventions which were 'hidden' within underpowered studies. The classic example of identifying benefits hidden in underpowered small studies is the use of corticosteroids in preterm labour, a forest plot which forms the logo of the Cochrane Collaboration. However, the conducting of a meta-analysis does not guarantee such power is achieved and reviews may remain underpowered (shows no significant difference when one exists or type II error).

In primary studies, post hoc power calculations can be used. However, until recently, no such equivalent was commonly used in meta-analysis. Conversely, as the publication of a new trial in an area presents an opportunity to perform another meta-analysis, multiple testing could cause type I errors in analyses (shows significant difference when one does not exist). These risks led to development of trial sequential analysis which can reduce both type I and II errors in reviews.

In trial sequential analysis, control for type I error rates is performed as each study is added by requiring a greater degree of statistical significance early in evidence accrual. In addition, it helps calculate the required number of participants required to provide a result which satisfies a particular power (specified by the user), similar to a post hoc power calculation in a primary study. In a sample of 50 anaesthesia meta-analyses, only 12% had adequate power (> 80%) and only 32% preserved type I error rates (were statistically significant after adjustment for multiple comparisons) [6]. Such common type I and type II errors will contribute to the poor PPV and NPV described above [2, 3].

Risk of bias

Meta-analyses are only as good as the trials they include. The ways in which the trials are conducted can lead to bias in the overall results of a meta-analysis [4]. Deficiencies in trial conduct, such as incomplete blinding, can exaggerate the effects of an intervention [4]. Lack of blinding is a particular problem in surgical studies where great imagination and effort is required to ensure subjects and observers are blinded to the obvious external signs of the intervention. Although it can often be difficult or impossible to blind, this does not reduce the bias of lack of blinding which can lead to incorrect conclusions.

Although risk of bias assessment is now widely used in meta-analyses, excluding high risk of bias trials or adjusting estimates of poor-quality studies [4] is not always performed [7] and readers of meta-analyses need to know to look for this assessment and note its findings or the absence when deciding how much confidence to put in the findings. Solutions to reducing risk of bias in analysis include review authors only including trials of the highest quality that score low risk for all domains. Although this would drastically reduce the number of included studies (and in most cases leave none!), it would provide more confidence in the results and may improve the predictive ability of meta-analyses.

How do we make the majority of meta-analyses true?

For the many reasons listed above, high-quality evidence from systematic reviews (even within the Cochrane Collaboration) may be lower than first thought [8]. Potential solutions previously discussed in this article include, (as a part of

journal requirements for publication): having a search plan which includes a thorough search for unpublished studies with tenacious pursuit of study authors to provide missing data and narrowing inclusion to the highest quality primary studies with the acceptance that some reviews may include no studies of a high enough quality for inclusion. Trial sequential analysis and investigation of heterogeneity should be performed and may generate future research hypotheses or help identify why the results of studies in a review differ.

Publication bias could be further reduced by journals guaranteed publication schemes for primary studies. For example, journals could agree to publish studies before recruitment of participants if it answers an important clinical question, has rigorous methodology and is adequately powered. If completed, the study is published regardless of the direction of results. This may help channel resources into priority research questions whilst encouraging both completion and submission of studies for publication. These changes could be tested empirically to identify whether PPVs can be improved with the above measures. The challenge is that until the hegemony of, and preoccupation with, impact factor is overcome, all but the most secure journals are unlikely to sign up to this, for fear of an obligation to publish negative (and hence less cited) results.

Where do we go from here?

There is no doubt that meta-analyses of high-quality RCTs (especially individual patient data), not subject to the limitations above, provide high-quality evidence as they help show consistency of results across different populations and potentially increase power. Unfortunately, these meta-analyses are the exception rather than the rule. However, even meta-analyses of small RCTs can be helpful in generating hypotheses by investigating where interventions may be more effective [4] but they should be treated with healthy scepticism by readers and conclusions tempered by review authors. There is a temptation for readers of meta-analysis merely to look at the forest plot summative diamond without further qualification of the reliability of the data. One must resist this tempting shortcut if one is to draw accurate conclusions, and all should be educated in the correct interpretation of such a ubiquitous tool.

Meta-analyses may also prompt the conduct of large RCTs where an intervention has been identified as potentially beneficial. Only rarely should a meta-analysis result be regarded as providing definitive evidence for an intervention, rather, it should prompt the conduct of a definitive RCT in the area. Ultimately, rather than resources being expended on small and underpowered RCTs, research groups and clinical trial networks should aim to collaborate on large multicentre trials, free from industry control that are adequately powered to definitively answer research questions and avoid the limitations of meta-analyses described in this article.

Compliance with ethical standards

Conflict of interest Jonathan Lund is Editor in Chief of Techniques in Coloproctology. Brett Doleman is a Junior Editor of Techniques in Coloproctology.

Ethical approval This article does not contain any studies with human participants or animals performed by any of the authors.

Informed consent For this type of study, informed consent is not required.

References

1. Ioannidis JP (2005) Why most published research findings are false. PLoS Med 2:e124
2. LeLorier J, Gregoire G, Benhaddad A, Lapierre J, Derderian F (1997) Discrepancies between meta-analyses and subsequent large randomized, controlled trials. N Engl J Med 337:536–542
3. Sivakumar H, Peyton PJ (2016) Poor agreement in significant findings between meta-analyses and subsequent large randomized trials in perioperative medicine. Br J Anaesth 117:431–441
4. Doleman B, Sutton AJ, Sherwin M, Lund JN, Williams JP (2018) Baseline morphine consumption may explain between-study heterogeneity in meta-analyses of adjuvant analgesics and improve precision and accuracy of effect estimates. Anesth Analg 126:648–660
5. Hedin RJ, Umberham BA, Detweiler BN, Kollmorgen L, Vassar M (2016) Publication bias and nonreporting found in majority of systematic reviews and meta-analyses in anesthesiology journals. Anesth Analg 123:1018–1025
6. Imberger G, Gluud C, Boylan J, Wetterslev J (2015) Systematic reviews of anesthesiologic interventions reported as statistically significant: problems with power, precision, and type 1 error protection. Anesth Analg 121:1611–1622
7. Detweiler BN, Kollmorgen LE, Umberham BA, Hedin RJ, Vassar BM (2016) Risk of bias and methodological appraisal practices in systematic reviews published in anaesthetic journals: a meta-epidemiological study. Anaesthesia 71:955–968
8. Conway A, Conway Z, Soalheira K, Sutherland J (2017) High quality of evidence is uncommon in Cochrane systematic reviews in anaesthesia, critical care and emergency medicine. Eur J Anaesth 34:808