COMMISSIONED ARTICLE PART I

Systematic Reviews and Meta-Analyses in Neurosurgery Part I: Interpreting and Critically Appraising as a Guide for Clinical Practice

Authors and affiliations

Keng Siang Lee[1,2*], Julian PT Higgins[3], Daniel M Prevedello[4]

[1]Department of Neurosurgery, King's College Hospital, London, UK

[2]Department of Basic and Clinical Neurosciences, Maurice Wohl Clinical Neuroscience Institute, Institute of Psychiatry, Psychology and Neuroscience (IoPPN), King's College London, London, UK

[3]Population Health Sciences, Bristol Medical School, University of Bristol, Bristol, UK

[4]Department of Neurosurgery, Wexner Medical Center, The Ohio State University, Columbus, Ohio, USA

***Corresponding author's name and affiliations:**

Keng Siang Lee, MBChB (Dist.), MRes (Dist.)

Department of Neurosurgery, King's College Hospital, London, UK

Department of Basic and Clinical Neurosciences, Maurice Wohl Clinical Neuroscience Institute, Institute of Psychiatry, Psychology and Neuroscience (IoPPN), King's College London, London, UK

mrkengsianglee@gmail.com

Key Words: evidence-based medicine, meta-analysis, neurosurgery, research methodology, systematic review

ORCiD

Keng Siang Lee; ORCiD: 0000-0003-2308-0579

COMMISSIONED ARTICLE PART I

Abstract

Neurosurgeons are inundated with the Herculean task to keep abreast with the rapid pace at which clinical research is proliferating. Systematic reviews and meta-analyses (SRMAs) have consequently surged in popularity because, when executed properly, they constitute the highest level of evidence, and may save busy neurosurgeons many hours of combing the literature. Well-executed SRMAs may prove instructive for clinical practice, but poorly conducted reviews sow confusion and may potentially cause harm. Unfortunately, many SRMAs within neurosurgery are relatively lackluster in methodological rigor. When neurosurgeons apply the results of a systematic review or meta-analysis to patient care, they should start by evaluating the extent to which the employed methods have likely protected against misleading results. The present article aims educate the reader about how to interpret an SRMA, to assess the potential relevance of its results in the special context of the neurosurgical patient population.

Keywords: evidence-based medicine; systematic review and meta-analysis; meta-regression; neurosurgery; surgery; research methodology

INTRODUCTION

Neurosurgeons are inundated with the Herculean task to keep abreast with the rapid pace at which clinical research is proliferating [1, 2]. This shift towards evidence-based neurosurgery comes on the heels of an ongoing data-driven transformation in healthcare delivery. Keeping up to date with current knowledge of the comparative efficacy of different treatments is an essential part of making good clinical decisions during everyday clinical practice [3]. Systematic reviews and meta-analyses (SRMAs) have consequently been a beneficiary of such trends. Neurosurgical Review, for instance, published 127 systematic reviews in 2023, compared to 38 of such articles in 2018.

Unfortunately, many SRMAs within neurosurgery are relatively lackluster in methodological rigor and compliance.[4]. Results from SRMAs may vary widely in quality and deviate from the truth because of large variation in the quality of the studies that have been included, as well as clinical and methodological differences between the pooled studies [4]. Estimates produced from SRMAs are only as reliable as the make-up of the individual studies summarized. Notwithstanding, rigorously executed SRMAs are considered to provide the highest level of evidence and can answer well-defined research questions and advocate for a superior technique or illuminate the epidemiology of a rare disease condition. SRMAs of a body of evidence include a larger sample size than its individual studies. They are therefore better powered, leading to greater precision of estimates, without undue emphasis on the individual studies. This may bolster confidence when applying the results to the patient.

To employ SRMAs effectively, neurosurgeons need to be aware of the relative merits and limitations of the methodology, so that they can weigh up the results, and thus critically appraise the conclusions of the review. The present article aims to educate the reader about how to interpret an SRMAs, to assess the potential relevance of its results for an individual patient in the special context of the neurosurgical patient population. This guide does not provide comprehensive advice to neurosurgeons on how to design or conduct meta-analyses which will be covered in the companion article Part II.

WAS THE RESEARCH QUESTION FOCUSED AND CLEARLY FORMULATED TO ADDRESS A SENSIBLE CLINICAL PROBLEM?

Systematic reviews should have a clear focus and address questions that can be applicable to real-life situations in clinical practice. This question can be defined by the PICO framework – population (P), interventions (I), comparators (C), outcomes (O). PICO is an established tool for operationalizing a clear and specific clinical research question. An appropriate use of this framework, is when one intends to perform a comparative meta-analysis investigating whether or not bridging thrombolysis (I) compared with direct mechanical thrombectomy (C) improves mortality (O) in patients with stroke due to basilar artery occlusion (P) [5]. Focused and well-defined clinical questions such as the above example are more likely to be applied to a specific clinical scenario than broad reviews about "management of stroke due to basilar artery occlusion". The PICO may even be adapted to questions about single groups of individuals, for example to determine the rates of success (O) of endoscopic third ventriculostomy (I) for shunt malfunction in the pediatric population (P) [6]. In the latter example, there is no comparator (C) [6].

WERE THE ELIGIBILITY CRITERIA FOR SELECTING STUDIES APPROPRIATE?

There are pertinent issues to consider when crafting the study eligibility criteria, which can have technical or statistical implications for downstream meta-analyses. There is no consensus whether or not to include conference abstracts and other gray literature or unpublished results, may not have been peer-reviewed [1]. This conflict arises because of the counterpoise that: 1) excluding them could entrench publication bias, whilst 2) including them can be time consuming and entails risk of erroneous data. We recommend that such publications be considered for inclusion on a case-by-case basis, depending on resource constraints, and an a priori judgment about the likelihood that the gray literature will reveal important studies. Hallan et al. demonstrated that only half of studies presented as abstracts at the American Association of Neurological Surgeons (AANS) annual meetings were followed by a full-text publication [10]. Despite controversies about inclusion of unpublished studies, mainly related to potentially lower methodological quality, their omission can lead to a possible overrepresentation of studies with positive results. Publication bias exists even within top neurosurgical journals, with positive abstracts being 40% more likely to be published than negative abstracts [10]. On rare occasions, conference abstracts or presentations may contain all requisite information, but it has been shown that a fifth of papers subsequently published within neurosurgery differed in message from their original abstract [10]. Assuming these abstracts meet predefined eligibility criteria, reviewers who wish to include gray literature must be prepared to scrutinize the abstract. The threshold for including conference abstracts is relatively higher.

The goal of an SRMA should be to mitigate, rather than propagate biases from the individual included studies. Special attention should be paid to the eligibility criteria based on study design. Non-randomized studies of interventions form the bulk of the evidence base in many surgical disciplines [12, 13]. The quality of non-randomized studies in the neurosurgical literature varies widely — from small single-center studies that do not control for confounding, to studies that pursue rigorous quasi-experimental methodologies which emulate randomization [11]. Therefore, it is vital to discern which non-randomized study designs are eligible for inclusion, in order to ensure the reliability and validity of the summary treatment effect [14-18]. Apart from study design alone, eligibility criteria for including non-randomized studies could also be formed around risk-of-bias (RoB) assessments, and this potentially allows for a more comprehensive and systematic way of distinguishing ‘weak’ versus ‘strong’ non-randomized study designs. The topic of RoB will be discussed further later.

Whether or not to impose exclusion criteria against studies on the basis of sample size is another contentious topic. Meta-epidemiological investigations have revealed an association between “small” studies with higher RoB. Nonetheless, the definition of a “small study” tends to be arbitrary and studies in neurosurgery may necessarily be “small” because of the unique clinical constraints faced by the field.

WAS THE SEARCH STRATEGY EXHAUSTIVE, SYSTEMATIC AND REPRODUCIBLE?

A rigorous systematic search strategy is central to a valid systematic review. The search strategy should follow from the research question, eligibility criteria, and PICO items. The most fundamental aspect of a search strategy is its sensitivity. That is, the aim is to capture as many studies as possible that meet the eligibility criteria, and work towards accuracy and completeness of the evidence base. A high-quality search strategy is a core element of the systematic review search plan and its aim is to reduce bias in identifying and selecting source reports.

For a wholistic coverage, it is usually recommended that at least three electronic databases are searched for potentially eligible studies and we recommend routine use of OVID MEDLINE, OVID Embase, and the Cochrane Central Register of Controlled Clinical Trials (CENTRAL). Searching a single database may be insufficient and would risk presenting misleading results if they fail to secure a complete or representative sample of the available eligible studies [8]. Capturing only part of the literature may introduce bias with its direction either inflating or reducing treatment effects. Multiple synonyms and search terms to describe each concept are also needed in order to be robust. An example of a full search strategy when investigating the prevalence of small unruptured intracranial aneurysms (SUIA) in the general population is provided in Table 1. In this example, the concepts of “small,” “unruptured,” and intracranial aneurysms” were used in addition to synonyms and related terms.

Additionally, it is more reassuring to the reader for reference lists of included studies to be scrutinized to identify relevant studies fitting the inclusion criteria that may have been inadvertently overlooked in the search strategy – a technique known as snowballing to ensure literature saturation [9].

WERE THE STUDY SELECTION AND ASSESSMENT PROCESSES SYSTEMATIC AND REPRODUCIBLE?

A systematic and reproducible selection of articles is the hallmark of an SRMA. Along the way, some decisions will be subjective and may deviate from its established protocol. The selection process should be documented, ideally in a flow diagram, detailing the number of studies excluded at each step and the reasons for exclusion [21]. Employing two or more reviewers may reduce subjectivity and therefore errors. A third reviewer may also help resolve disagreements through discussions or further adjudication. At times, a measure of agreement between the reviewers on study selection and quality appraisal, such as the Cohen’s kappa, may be reported [22]. If there is good agreement, one may be more confident in the above process.

HAVE THE RESULTS BEEN PRESENTED APPROPRIATELY FOR INTERPRETATION AND CLINICAL APPLICATION?

SRMAs provide estimates a summary effect size which is a quantitative measure of the magnitude of difference between groups. The nature of the effect size follows that of the outcome. Table 3 of companion article Part II, provides more information on effect measures routinely seen in the neurosurgical literature.

Forest plots are typically how the results of meta-analyses are visually conveyed. The point estimate of each study is presented as a square with its size proportional to the corresponding weight given to the study. The corresponding horizontal line is the confidence interval (CI) around the point estimate. Finally, the combined summary effect, is portrayed as a diamond, with its width representing the 95% CI. The CI around the summary effect presents the range in which the true effect is most likely to lie. In contrast, 95% prediction intervals are an index of dispersion, informing us how widely the treatment effect size varies. A prediction interval provides estimates of what the effect size might be for similar studies conducted in the future. Example forest plots with its interpretation for the use of cerebrospinal fluid (CSF) lumbar drainage (LD) amongst patients with aneurysmal subarachnoid hemorrhage (aSAH) on delayed cerebral ischemia (DCI) are shown in Figure 1 [23].

It may be reasonable to report more than one effect measure for the same outcome, to help put a result into perspective by weighing up its benefits and harms. For example, it may be justifiable to provide risk difference (RD), the numbers-needed to treat for benefit (NNTB) or harm (NNTH), in addition to the relative risk (RR), for dichotomous outcomes, since the magnitude of effect could appear substantial based on the RR (relative effect), whereas the RD and NNT (absolute effects) may convey otherwise, when the prevalence of the event is low. Patients at high baseline risk can expect more benefit than those at lower baseline risk from the same intervention (the same relative effect). For instance between an individual with high baseline risk for DCI events (30%) and the other with a lower risk (5%), a 41% relative risk reduction obtained from an SRMA of CSF LD in aSAH, can be applied onto each of these baseline risks [23]. The resulting absolute risk reduction attributable to CSF LD would then calculated to be 12.3% for the high-risk individual and 2.1% for the low-risk individual. We may also look at the NNT which is the inverse of the absolute RD. In the example found in Table 2, the NNT is 6, which means that for every six patients with aSAH that are treated with CSF LD as opposed to control, one case of DCI can be prevented. On the other hand, the NNH is 111, which means only one case of infection would occur for every 111 patients with aSAH that are treated with CSF LD as opposed to control. As such, CSF LD would be have merits as a good therapeutic intervention when trading off its benefits and harm.

Similarly, the same style of presentation could be applied to continuous endpoints. For example in elderly patient with adult spinal deformities undergoing minimally invasive surgery, pain on the 100 mm visual analog scale (VAS) can be better understood by informing readers that the minimal amount considered by patients to be important on that scale is a positive change of 1.2 points [25].

WERE STATISTICAL METHODS USED TO COMBINE THE STUDIES SOUND? WHAT WAS THE CHOICE OF WEIGHTING ESTIMATORS AND MODELS TO HANDLE HETEROGENEITY?

Fixed-effects and random-effects models are the most commonly employed statistical models for SRMAs. The fixed-effect model is predicated on the assumption that there is one single underlying, true effect size, and that individual studies produce estimates that deviate from this ground-truth due to sampling errors, in other words, by chance [26]. Random-effects models, on the other hand, is contingent upon multiple existing ground-truths – that each study estimates an effect size that is specific to the characteristics of its population, interventions, and study design, which results in genuine variation in effect sizes across different studies [26].

Random-effects models may be used when heterogeneity is anticipated a priori. Heterogeneity is defined as inconsistency in the treatment effect across primary studies. While homogenous results do not exclude clinical or methodological aspects between studies, on the contrary, heterogeneous results probably reflect some underlying differences. More granular discussion on exploring and addressing heterogeneity would be covered below.

ARE THE RESULTS CONSISTENT ACROSS STUDIES? IDENTIFYING AND ADDRESSING HETEROGENEITY

Heterogeneity was briefly explained above and is estimated as part of a random-effects SRMA. One could identify heterogeneity in one of two ways: 1) visually inspecting the point estimates and CIs of a forest plot, and 2) performing a statistical comparison [27].

Via forest plots, authors may easily note differences in the point estimates between individual studies and then the extent to which CIs overlap. Stark differences in point estimates or CIs that do not overlap suggest that the difference in results between studies is unlikely attributable to random error and therefore our confidence in the summary effect estimate diminishes. In contrast, individual studies are relatively homogeneous when the point estimates are similar with overlapping CIs, such as in the subgroup analysis of only RCTs in Figure 1.

The amount of statistical heterogeneity is often tested for using a χ^2 test. The I^2 statistic is derived from the test statistic from this χ^2 test and measures the proportion of variability in estimated treatment effects that is due to heterogeneity rather than sampling variation within the studies [28, 29]. The I^2 statistic may be viewed as a measure of how inconsistent the results are with each other and should not be misinterpreted as a measure of the amount of heterogeneity [29, 30]. Authors often use arbitrary rules to decide whether there is heterogeneity based on certain thresholds: 1) I^2 statistic >30%, and/or 2) P value <0.10. When substantial heterogeneity exists, pooling data and presenting a single summary estimate can be erroneous.

Authors should explore explanations when substantial heterogeneity exists. Subgroup analyses are simple methods for exploring and addressing heterogeneity. Subgroups can be defined based on a priori clinical knowledge to delineate the presence of effect modification. For example, an SRMA on the use of antiplatelet therapy for aSAH, might stratify study effect sizes according to primary treatment modality study (clipping versus endovascular coiling), based on clinical observations that patients that have undergone clipping have increased rates of in-hospital mortality [31, 32], as shown in Figure 3. A more sophisticated approach to exploring causes of heterogeneity is through the use of random-effects meta-regression [33, 34]. Both subgroup analyses and meta-regression may have limited feasibility in neurosurgery [6, 35] – 1) 10 or more included studies are recommended for each covariate, and few topics in neurosurgery feature sufficient numbers of high-quality studies; 2) meta-regression has low statistical power and requires a large numbers of studies; and 3) the use of aggregate data as a covariate leads to ecological bias, which can cause overestimation of its effects [36]. Finally, especially in neurosurgery, heterogeneity could reflect the existence of learning curves, such as the inception of new technologies such as robotic surgery, or novel surgical approaches [37].

WHAT IS THE CONFIDENCE IN THE ESTIMATES OF EFFECT?

Confidence ratings begin by considering study design. RCTs are initially assigned high confidence and observational studies are given low confidence, but a number of factors may modify these initial ratings. Confidence may decrease when there is high risk of bias, inconsistency, imprecision, indirectness, or concern about publication bias. An increase in confidence rating is uncommon and occurs primarily in observational studies when the effect size is large.

Only when a pooled statistic is presented with its heterogeneity measures and its certainty are readers able to start to make an informed judgement about the spread of variance of the SRMAs results. An increasingly adopted framework is the GRADE (Grading of Recommendations Assessment, Development, and Evaluation) recommendations, within which authors assess the quality (certainty) of evidence from the body of evidence based on specific study characteristics. The GRADE method assesses the overall quality of evidence based on the five domains: risk of bias, heterogeneity (inconsistency), indirectness, precision, and publication bias. The subsequent subsections would address the confidence in the summary estimates by walking through each of the five domains. The quality of evidence can be downgraded or upgraded based on consideration of these multiple factors. An illustrative example is provided in Table 2. The overall quality of evidence for each outcome can then be appraised based on four tiers: Very low (⊖⊖⊖⊖), Low (⊕⊕⊖⊖), Moderate (⊕⊕⊕⊖), or High (⊕⊕⊕⊕).

There is a need for the reading neurosurgeon to also interpret the clinical appropriateness of the outcome being reported, for not all outcomes in SRMAs necessarily can applied to practice. Clinical reasoning and intuition is required by both the author and the reader to

justify as to whether or not an outcome reported has merit clinically, which can be influenced after knowing how certain the outcome is. As much as possible, a GRADE assessment should be reported for each outcome [38], nonetheless, readers can consider these factors regardless of whether authors of the SRMA formally used this approach.
[27, 38-47].

HOW SERIOUS IS THE RISK OF BIAS?

Ideally, authors should evaluate and report the RoB for each of the main outcomes measured in each individual study. Numerous tools are available for assessing methodological quality or RoB, such as version 2 of the Cochrane risk-of-bias assessment for RCTs (RoB 2) for RCTs [49], and ROBINS-I [50], Newcastle-Ottawa scale [51] or the Joanna Briggs Institute (JBI) checklist [52], for non-randomized studies. The latter tools take into account additional sources of bias that are not otherwise addressed by analytical design features, such as biases that may arise from protocol deviation, outcome assessment, missing data and adequacy of follow-up. For example, a large propensity-score matched study could falsely manifest the appearance of a ‘strong’ non-randomized study on the basis of their matched study design, but a thorough RoB assessment may reveal weaknesses in terms of residual confounding or selective reporting. Within an SRMA, quality and RoB assessment tools are frequently used as a basis of sensitivity or subgroup analyses to examine the impact of excluding the weaker studies, such as in Figure 1 where different summary effect are shown based on the study design, rather than excluding low-quality studies [23, 31]. A judgment about the overall risk of bias for all of the included studies may then result in decreasing the confidence in estimates [41, 53].

HOW PRECISE ARE THE RESULTS?

Studies may mislead through 1) systematic error or bias, as discussed above, and 2) random error. The latter is negatively correlated with sample sizes, and prevalence of events. When sample size and number of events are small, random error is therefore large and we refer to the results as “imprecise”.

SRMAs generate CIs around the summary estimate, usually depicted as the diamond shape on forest plots, as shown in Figure 1. Neurosurgeons can assess precision by considering the boundaries of the CI, and determine if their advice to their patients would be the same if the extreme ends of the boundaries were to represent the truth [44]. Considering the same example of CSF LD in aSAH (Figure 1 and Table 2), the CI around the absolute effect of CSF LD is a reduction from 84 (the minimum) to 219 (the maximum) DCI in 1000 patients with aSAH given CSF LD. Albeit a subjective judgement, the neurosurgeon must ask if their patients would make different choices about the use of CSF LD if their risk of DCI decreased by only 84 in 1000 or by as much as 219 in 1000. If the neurosurgeon is confident that patients would derive similar signifance on reductions of 84 and 219 in 1000 infarctions, concern about imprecision would be trivial

IS THERE CONCERN ABOUT REPORTING OR PUBLICATION BIAS?

When the decision to publish certain material is based upon the statistical significance or magnitude of the results, a systematic error termed selective outcome reporting bias ensues [54, 55]. Publication bias refers to the higher probability of studies with positive results being published [56, 57]. The preference towards studies with positive results leaves many small studies with negative results unpublished, exacerbating the 'file-drawer' problem [58]. Consequently, inclusion of more studies with positive results can lead to an overestimation of the treatment effect, with risk of a Type 1 false-positive error.

DO THE RESULTS APPLY DIRECTLY TO MY PATIENT? HOW MAY I TRANSLATE THE RESULTS INTO CLINICAL PRACTICE?

The ideal body of evidence for clinical decision-making stems from research that has directly compared interventions in which we are concerned about, evaluated in our intended populations. If the PICO in studies differ from those of our interest, this evidence is indirect. Indirect evidence occurs when, for instance, neurosurgeons must choose between interventions that have not been tested in head-to-head comparisons [45]. As a case study, RCTs have compared direct mechanical thrombectomy versus best medical management in basilar artery occlusion (BAO) [60], but only three cohort studies have compared direct mechanical thrombectomy against bridging thrombectomy [5]. The juxtaposition of these two interventions under these circumstances requires extrapolation from indirect comparisons and multiple assumptions. Ultimately these factors will lower confidence in the summary estimates [45]. Nonetheless, on practical terms, the current summary estimate still remains the best estimate of the treatment effect to inform their decisions, as long as it is interpreted judiciously.

When translating the results of an SRMA into clinical practice, it is paramount to bear in mind that what neurosurgeons may consider important not necessarily be significant to patients. Patients may value certain risks and benefits differently from clinicians, for example, neurosurgeons may be interested in surrogate end points for functional outcomes such as the modified Rankin scale (mRS), whereas patients are most concerned about other outcomes such as return to work. Clinicians should not view results in terms of their statistical significance but take into account its clinical importance. A statistically significant difference has little meaning if it is not clinically relevant. Any difference may be statistically significant with an adequate sample size . Conversely, non-significant resultsdo not equate to no difference and the reader should scrutinize the boundaries of the CI – as the absence of evidence is not evidence of absence.

CONCLUSIONS

The neurosurgical literature has historically faced a problem with paucity of high-quality randomized evidence [13]. This is an inevitable mainstay, not least because of the relatively smaller volume of patients worldwide afflicted with neurosurgical conditions and the heterogeneity of such diseases, but also the practical and ethical challenges of performing RCTs involving high-risk procedures in humans. In spite of this, the approach outlined herein should empower readers and authors with the fundamental framework to identify an appropriate SRMA, scrutinize methodology to determine its credibility, interpret its summary results, and judge their certainty. In the hands of an informed user, an SRMA can be a powerful tool to inform decision-making, when based on the totality of the best evidence.

REFERENCES

1. Khan, N.R., et al., *An analysis of publication productivity for 1225 academic neurosurgeons and 99 departments in the United States.* J Neurosurg, 2014. **120**(3): p. 746-55.
2. Davidoff, F., et al., *Evidence based medicine.* BMJ, 1995. **310**(6987): p. 1085-6.
3. Lee, K.S., et al., *Tenets for the Proper Conduct and Use of Meta-Analyses: A Practical Guide for Neurosurgeons.* World Neurosurg, 2022. **161**: p. 291-302.e1.
4. Klimo, P., et al., *Methodology and reporting of meta-analyses in the neurosurgical literature.* J Neurosurg, 2014. **120**(4): p. 796-810.
5. Lee, K.S., et al., *Bridging thrombolysis improves survival rates at 90 days compared with direct mechanical thrombectomy alone in acute ischemic stroke due to basilar artery occlusion: a systematic review and meta-analysis of 1096 patients.* J Neurointerv Surg, 2023. **15**(10): p. 1039-1045.
6. Lee, K.S., et al., *Endoscopic third ventriculostomy for shunt malfunction in the pediatric population: a systematic review, meta-analysis, and meta-regression analysis.* J Neurosurg Pediatr, 2023. **31**(5): p. 423-432.
7. McGowan, J., et al., *PRESS Peer Review of Electronic Search Strategies: 2015 Guideline Statement.* J Clin Epidemiol, 2016. **75**: p. 40-6.
8. Hopewell, S., et al., *Handsearching versus electronic searching to identify reports of randomized trials.* Cochrane Database Syst Rev, 2007. **2007**(2): p. MR000001.
9. Greenhalgh, T. and R. Peacock, *Effectiveness and efficiency of search methods in systematic reviews of complex evidence: audit of primary sources.* BMJ, 2005. **331**(7524): p. 1064-5.
10. Hallan, D.R., et al., *Charting the course from abstract to published article.* J Neurosurg, 2022. **136**(6): p. 1773-1780.
11. Lee, K.S., et al., *Radiological surveillance of small unruptured intracranial aneurysms: a systematic review, meta-analysis, and meta-regression of 8428 aneurysms.* Neurosurg Rev, 2021. **44**(4): p. 2013-2023.
12. Yarascavitch, B.A., et al., *Levels of evidence in the neurosurgical literature: more tribulations than trials.* Neurosurgery, 2012. **71**(6): p. 1131-7; discussion 1137-8.
13. Mansouri, A., et al., *Randomized controlled trials and neurosurgery: the ideal fit or should alternative methodologies be considered?* J Neurosurg, 2016. **124**(2): p. 558-68.
14. Lonjon, G., et al., *Comparison of treatment effect estimates from prospective nonrandomized studies with propensity score analysis and randomized controlled trials of surgical procedures.* Ann Surg, 2014. **259**(1): p. 18-25.
15. Ioannidis, J.P., et al., *Comparison of evidence of treatment effects in randomized and nonrandomized studies.* JAMA, 2001. **286**(7): p. 821-30.
16. Benson, K. and A.J. Hartz, *A comparison of observational studies and randomized, controlled trials.* N Engl J Med, 2000. **342**(25): p. 1878-86.
17. Venkataramani, A.S., J. Bor, and A.B. Jena, *Regression discontinuity designs in healthcare research.* BMJ, 2016. **352**: p. i1216.
18. Moscoe, E., J. Bor, and T. Bärnighausen, *Regression discontinuity designs are underutilized in medicine, epidemiology, and public health: a review of current and best practice.* J Clin Epidemiol, 2015. **68**(2): p. 122-33.
19. Grainge, M. *Excluding small studies from a systematic review or meta-analysis. Presented at: CSG Annual Meeting 2015; March 12-18, 2015; Dresden, Germany.* 10 February 2024]; Available from: https://skin.cochrane.org/sites/skin.cochrane.org/files/public/uploads/CSG-COUSIN_March%202015_M%20Grainge.pdf .

20. Dechartres, A., et al., *Influence of trial sample size on treatment effect estimates: meta-epidemiological study.* BMJ, 2013. **346**: p. f2304.
21. Page, M.J., et al., *The PRISMA 2020 statement: an updated guideline for reporting systematic reviews.* BMJ, 2021. **372**: p. n71.
22. Cohen, J., *A coefficient of agreement for nominal scales.* Educational and Psychological Measurement, 1960. **20**(1): p. 37–47.
23. Lee, K.S., et al., *Effectiveness of Cerebrospinal Fluid Lumbar Drainage Among Patients with Aneurysmal Subarachnoid Hemorrhage: An Updated Systematic Review and Meta-Analysis.* World Neurosurg, 2024.
24. Murad, M.H., et al., *Estimating risk difference from relative association measures in meta-analysis can infrequently pose interpretational challenges.* J Clin Epidemiol, 2009. **62**(8): p. 865-7.
25. Park, P., et al., *Can a Minimal Clinically Important Difference Be Achieved in Elderly Patients with Adult Spinal Deformity Who Undergo Minimally Invasive Spinal Surgery?* World Neurosurg, 2016. **86**: p. 168-72.
26. DerSimonian, R. and N. Laird, *Meta-analysis in clinical trials.* Control Clin Trials, 1986. **7**(3): p. 177-88.
27. Guyatt, G.H., et al., *GRADE guidelines: 7. Rating the quality of evidence--inconsistency.* J Clin Epidemiol, 2011. **64**(12): p. 1294-302.
28. Higgins, J.P., et al., *Measuring inconsistency in meta-analyses.* BMJ, 2003. **327**(7414): p. 557-60.
29. Higgins, J.P. and S.G. Thompson, *Quantifying heterogeneity in a meta-analysis.* Stat Med, 2002. **21**(11): p. 1539-58.
30. Turner, R.M., et al., *Predicting the extent of heterogeneity in meta-analysis, using empirical data from the Cochrane Database of Systematic Reviews.* Int J Epidemiol, 2012. **41**(3): p. 818-27.
31. Lee, K.S., et al., *Antiplatelet therapy in aneurysmal subarachnoid hemorrhage: an updated meta-analysis.* Neurosurg Rev, 2023. **46**(1): p. 221.
32. Molyneux, A.J., et al., *The durability of endovascular coiling versus neurosurgical clipping of ruptured cerebral aneurysms: 18 year follow-up of the UK cohort of the International Subarachnoid Aneurysm Trial (ISAT).* Lancet, 2015. **385**(9969): p. 691-7.
33. Thompson, S.G. and J.P. Higgins, *How should meta-regression analyses be undertaken and interpreted?* Stat Med, 2002. **21**(11): p. 1559-73.
34. Higgins, J.P. and S.G. Thompson, *Controlling the risk of spurious findings from meta-regression.* Stat Med, 2004. **23**(11): p. 1663-82.
35. Lee, K.S., et al., *Surgical revascularizations for pediatric moyamoya: a systematic review, meta-analysis, and meta-regression analysis.* Childs Nerv Syst, 2023. **39**(5): p. 1225-1243.
36. Berlin, J.A., et al., *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**(3): p. 371-87.
37. Lee, K.S., et al., *Accuracy of robot-assisted stereotactic MRI-guided laser ablation in children with epilepsy.* J Neurosurg Pediatr, 2023. **32**(2): p. 214-222.
38. Guyatt, G.H., et al., *GRADE guidelines: 12. Preparing summary of findings tables-binary outcomes.* J Clin Epidemiol, 2013. **66**(2): p. 158-72.
39. Guyatt, G., et al., *GRADE guidelines: 1. Introduction-GRADE evidence profiles and summary of findings tables.* J Clin Epidemiol, 2011. **64**(4): p. 383-94.
40. Guyatt, G.H., et al., *GRADE guidelines: 2. Framing the question and deciding on important outcomes.* J Clin Epidemiol, 2011. **64**(4): p. 395-400.

41. Balshem, H., et al., *GRADE guidelines: 3. Rating the quality of evidence.* J Clin Epidemiol, 2011. **64**(4): p. 401-6.
42. Guyatt, G.H., et al., *GRADE guidelines: 4. Rating the quality of evidence--study limitations (risk of bias).* J Clin Epidemiol, 2011. **64**(4): p. 407-15.
43. Guyatt, G.H., et al., *GRADE guidelines: 5. Rating the quality of evidence--publication bias.* J Clin Epidemiol, 2011. **64**(12): p. 1277-82.
44. Guyatt, G.H., et al., *GRADE guidelines 6. Rating the quality of evidence--imprecision.* J Clin Epidemiol, 2011. **64**(12): p. 1283-93.
45. Guyatt, G.H., et al., *GRADE guidelines: 8. Rating the quality of evidence--indirectness.* J Clin Epidemiol, 2011. **64**(12): p. 1303-10.
46. Guyatt, G.H., et al., *GRADE guidelines: 9. Rating up the quality of evidence.* J Clin Epidemiol, 2011. **64**(12): p. 1311-6.
47. Schünemann, H., et al. *Chapter 14: Completing 'Summary of findings' tables and grading the certainty of the evidence. In: Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, Welch VA (editors). Cochrane Handbook for Systematic Reviews of Interventions version 6.2 (updated February 2021).* 1 April 2021]; Available from: https://training.cochrane.org/handbook/current/chapter-14.
48. Moher, D., et al., *Does quality of reports of randomised trials affect estimates of intervention efficacy reported in meta-analyses?* Lancet, 1998. **352**(9128): p. 609-13.
49. Sterne, J.A.C., et al., *RoB 2: a revised tool for assessing risk of bias in randomised trials.* BMJ, 2019. **366**: p. l4898.
50. Sterne, J.A., et al., *ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions.* BMJ, 2016. **355**: p. i4919.
51. Wells, G., et al. *The Newcastle-Ottawa Scale (NOS) for assessing the quality of nonrandomised studies in meta-analyses.* 1 April 2021]; Available from: ohri.ca/programs/clinical_epidemiology/oxford.asp.
52. Tufanaru, C., et al. *Chapter 3: Systematic reviews of effectiveness. Aromataris E, Munn Z, editors. JBI Manual for Evidence Synthesis*. 2020; Available from: https://doi.org/10.46658/JBIMES-20-04.
53. Guyatt, G.H., et al., *GRADE: an emerging consensus on rating quality of evidence and strength of recommendations.* BMJ, 2008. **336**(7650): p. 924-6.
54. Smyth, R.M., et al., *Frequency and reasons for outcome reporting bias in clinical trials: interviews with trialists.* BMJ, 2011. **342**: p. c7153.
55. Chan, A.W., et al., *Empirical evidence for selective reporting of outcomes in randomized trials: comparison of protocols to published articles.* JAMA, 2004. **291**(20): p. 2457-65.
56. Egger, M., et al., *Bias in meta-analysis detected by a simple, graphical test.* BMJ, 1997. **315**(7109): p. 629-34.
57. Egger, M. and G.D. Smith, *Bias in location and selection of studies.* BMJ, 1998. **316**(7124): p. 61-6.
58. Stern, J.M. and R.J. Simes, *Publication bias: evidence of delayed publication in a cohort study of clinical research projects.* BMJ, 1997. **315**(7109): p. 640-5.
59. Lau, J., et al., *The case of the misleading funnel plot.* BMJ, 2006. **333**(7568): p. 597-600.
60. Tao, C., et al., *Endovascular Treatment Versus Best Medical Management in Acute Basilar Artery Occlusion Strokes: Results From the ATTENTION Multicenter Registry*. Circulation, 2022. **146**(1): p. 6-17.

Declarations

Ethical Approval

Not applicable

Funding

None

Table 1. Full search phrases used for the of three electronic databases – Medline, Embase, and Cochrane CENTRAL were undertaken from database inception to March 2020 for published studies reporting the growth and rupture risks of small unruptured intracranial aneurysms [11].

Ovid MEDLINE	
Aneurysm concept	
1	exp Intracranial Aneurysm/
2	Intracranial aneurysm*.tw
3	Brain aneurysm*.tw
4	Cerebral aneurysm*.tw.
5	1 or 2 or 3 or 4
Unruptured concept	
6	unruptured.tw.
Unruptured aneurysm concept	
7	5 or 6
8	UIA.tw
9	SUIA.tw
10	8 or 9
11	7 or 10
Small concept	
12	Small.tw
13	Tiny.tw
14	12 or 13
15	1 mm.tw
16	2 mm.tw
17	3 mm.tw
18	4 mm.tw
19	5 mm.tw
20	6 mm.tw
21	7 mm.tw
22	15 or 16 or 17 or 18 or 19 or 20 or 21
23	14 or 22
Small unruptured aneurysm concept	
24	11 and 23
Embase	

Aneurysm concept	
1	exp intracranial aneurysm/
2	exp brain artery aneurysm/
3	intracranial aneurysm*.tw
4	brain aneurysm*.tw
5	cerebral aneurysm*.tw.
6	1 or 2 or 3 or 4 or 5
Unruptured concept	
7	unruptured.tw.
Unruptured aneurysm concept	
8	6 or 7
9	exp unruptured intracranial aneurysm/
10	8 or 9
11	UIA.tw
12	SUIA.tw
13	11 or 12
14	10 or 13
Small concept	
15	small.tw
16	tiny.tw
17	15 or 16
18	1 mm.tw
19	2 mm.tw
20	3 mm.tw
21	4 mm.tw
22	5 mm.tw
23	6 mm.tw
24	7 mm.tw
25	18 or 19 or 20 or 21 or 22 or 23 or 24
26	17 or 25
Small unruptured aneurysm concept	
27	14 and 26
Cochrane Controlled Register of Trials CENTRAL	
Aneurysm concept	
1	MeSH descriptor: [Intracranial Aneurysm] explode all trees

2	(intracranial aneurysm):ti,ab,kw
3	(brain aneurysm):ti,ab,kw
4	(cerebral aneurysm):ti,ab,kw
5	#1 OR #2 OR #3 OR #4
Unruptured concept	
6	(unruptured):ti,ab,kw
Unruptured aneurysm concept	
7	#5 OR #6
8	(IUA):ti,ab,kw
9	(SUIA):ti,ab,kw
10	#8 OR #9
11	#7 OR #10
Small concept	
12	(small):ti,ab,kw
13	(tiny):ti,ab,kw
14	#12 OR #13
15	(1 mm):ti,ab,kw
16	(2 mm):ti,ab,kw
17	(3 mm):ti,ab,kw
18	(4 mm):ti,ab,kw
19	(5 mm):ti,ab,kw
20	(6 mm):ti,ab,kw
21	(7 mm):ti,ab,kw
22	#15 OR #16 OR #17 OR #18 OR #19 OR #20 OR #21
23	#14 OR #22
Small unruptured aneurysm concept	
24	#11 AND #23

Table 2. Evidence summary of the use of cerebrospinal fluid lumbar drainage in aneurysm subarachnoid hemorrhage, extracted from Lee et al [22].

Outcomes	**No. of patients (no. of included studies)**	**Relative risk (95%CI)**	**Risk difference per 1000**	**Statistical heterogeneity**	**Quality of evidence (GRADE)**
Delayed cerebral ischemia	1497 (10 studies)	0.59 (0.45 – 0.79)	164 fewer (84 fewer to 219 fewer)	I^2 = 31.6% (*P* = 0.156)	⊕⊖⊖⊖[a,b,c]
Rebleeding	973 (3 studies)	0.84 (0.13 – 5.39)	9 fewer (50 fewer to 250 more)	I^2 = 43.2% (*P* = 0.172)	⊕⊖⊖⊖[a,b,d]
Infection	1099 (5 studies)	1.05 (0.86 – 1.30)	1 more (2 fewer to 4 more)	I^2 = 0% (*P* = 0.503)	⊕⊖⊖⊖[a,b,d]

[a] Initial downgrade due to study design

[b] Downgraded by one level for inconsistency

[c] Downgraded by one level for possible funnel plot asymmetry

[d] Downgraded by one level for statistical imprecision

The relative risk (RR) for delayed cerebral ischemia (DCI) in aneurysmal subarachnoid hemorrhage (aSAH) is 0.59. The baseline risk of DCI – which is the risk without cerebrospinal fluid (CSF) lumbar drainage (LD) can be obtained from the trial that is the largest and likely enrolled most representative population (57/143, approximately 399 per 1000). The risk of DCI with CSF LD intervention would be (399 per 1000 x 0.59, approximately 235 per 1000). The absolute risk difference would be (235 per 1000 – 399 per 1000 = −164 per 1000, approximately 164 fewer DCI per 1000, with CSF LD). The same process can be used to calculate the confidence intervals around the risk difference, substituting the boundaries of the confidence interval (CI) of the RR for the point estimate. The number needed to treat to prevent one patient from DCI can also be calculated as the inverse of the absolute risk difference (1/0.164 = 6 patients).

FIGURE LEGENDS

Figure 1. This figure shows a forest plot, which is typically used to present the results of a meta-analysis by Lee et al [23]. In this example, the outcome parameter is delayed cerebral ischemia (DCI) – a dichotomous outcome (present or absent) – amongst patients with aneurysmal subarachnoid hemorrhage (aSAH). The relative risk of developing DCI when with or without cerebrospinal fluid (CSF) lumbar drainage (LD) is displayed on the x-axis. Instead of relative risks, odds ratios could also be presented. The line-of-no-effect (vertical line) separates outcomes that favor or disfavor CSF LD. The squared grey boxes represent the point estimates and the horizontal lines represent the associated 95% confidence intervals for each study. Meta-analysis provides a weighted average of the results of the individual studies in which the weight of the study depends on its precision. Studies that are more precise and narrower CIs will have greater weight and thus more influence on the combined estimate. For binary outcomes such as DCI, the precision depends on the number of events and sample size. The EARLYDRAIN trial by Wolf S et al., had the largest number of DCIs (98) and the largest sample size (287); therefore, it had the narrowest CI and the largest weight. Smaller trials with smaller numbers of events in that plot have a much wider CI, and their effect size is quite different from the combined effect (ie, had less weight in meta-analysis). This example demonstrates the power of a meta-analysis. Whereas the confidence intervals of all individual studies cross the line-of-no-effect (relative risk = 1) representing no significant differences, the confidence interval of the summary estimate (bottom-most red diamond shape) lies entirely to the left of the line-of-no-effect representing a significantly lower DCI risk with CSF LD (RR=0.59, 95%CI: 0.45; 0.79). Overlapping confidence intervals of the individual studies and an I^2 value of 32% with a non-significant p-value (0.16) indicates moderate statistical heterogeneity of the studies, and justifies presenting a summary estimate. In contrast, 95% prediction intervals (red horizontal line) are an index of dispersion, informing us how widely the treatment effect size varies. A prediction interval provides estimates of what the effect size might be for similar studies conducted in the future. This information tells us that we can expect the effect size of a new study from the same overall target population to fall between 0.34 and 1.05 with 95% probability. Thus, the treatment effect from new studies would not all be expected to report results that recommend CSF LD for a similar cohort of patients with aSAH. An additional subgroup analysis based on study type was performed which showed that no statistically significant difference between CSF LD and control in RCTs compared with observational studies However, given that different subgroups may include variable information and ability to detect treatment effects, performing a statistical comparison for subgroup differences may be misleading. Therefore, hypothetically, if the treatment effect was statistically significant in

one subgroup and nonsignificant in another subgroup, this would not signify that the subgroup variable explains heterogeneity.

Figure 2. Meta-analysis of cerebrospinal fluid lumbar drainage versus control treatment for prevention of delayed cerebral ischemia in aneurysmal subarachnoid hemorrhage, extracted from Lee et al [23]. Upper panel A: Funnel plot assessing for publication bias by plotting the relative risk point estimates of individual study results (grey circles) against the sample size of the study, demonstrates asymmetry. For example, there is an overrepresentation of smaller studies with significant findings on the bottom-left portion of the diagram, with few or no studies in the bottom-right part, indicating publication bias. Lower panel B: Trim-and-fill analysis to adjust for funnel plot asymmetry, as visual inspection of the funnel plot of observed studies in the upper panel indicated possible funnel plot asymmetry. The ‘trim and fill’ method aims to identify and correct for funnel plot asymmetry arising from publication bias. The basis of the method is to (1) ‘trim’ (remove) the smaller studies causing funnel plot asymmetry, (2) use the trimmed funnel plot to estimate the true ‘centre’ of the funnel, and (3) replace the omitted studies and their missing ‘counterparts’ around the centre (filling).

Figure 3. Forest plots with random-effects model, stratified by timing of antiplatelet use in aneurysmal subarachnoid hemorrhage (aSAH), for delayed cerebral ischemia (DCI), extracted from Lee et al [31]. The subgroup analysis showed that no statistically significant difference between pre-ictal antiplatelet therapy and control. However, the treatment effect was statistically significant in when comparing post-ictal antiplatelet therapy and control.