



# A guide to interpreting systematic reviews and meta-analyses in neurosurgery and surgery

Ignatius Esene<sup>1,2</sup> · Ourlad Alzeus G. Tantengco<sup>3</sup> · Faith C. Robertson<sup>4</sup> · Megan E. H. Still<sup>5</sup> · Alvan-Emeka K. Ukachukwu<sup>6</sup> · Ronnie E. Baticulon<sup>7</sup> · Ernest J. Barthélemy<sup>8</sup> · Daniela Perez-Chadid<sup>9</sup> · Laura Lippa<sup>10</sup> · Ana Cristina Veiga Silva<sup>11</sup> · Luxwell Jokonya<sup>12</sup> · Fahd D. Hassani<sup>13</sup> · Federico Nicolosi<sup>14</sup> · Berjo D. Takoutsing<sup>2</sup> · Jeff Ntalaja<sup>15</sup> · Samer S. Hoz<sup>16</sup> · Kazadi K. N. Kalangu<sup>12</sup> · Gilbert Dechambenoit<sup>17</sup> · Franco Servadei<sup>18</sup> · Najia El Abbadi<sup>13</sup> · Kee B. Park<sup>19</sup> · Angelos Kolias<sup>20,21</sup>

Received: 20 December 2023 / Accepted: 15 May 2024

© The Author(s), under exclusive licence to Springer-Verlag GmbH Austria, part of Springer Nature 2024

## Abstract

**Introduction** Systematic reviews (SRs) and meta-analyses (MAs) are methods of data analysis used to synthesize information presented in multiple publications on the same topic. A thorough understanding of the steps involved in conducting this type of research and approaches to data analysis is critical for appropriate understanding, interpretation, and application of the findings of these reviews.

**Methods** We reviewed reference texts in clinical neuroepidemiology, neurostatistics and research methods and other previously related articles on meta-analyses (MAs) in surgery. Based on existing theories and models and our cumulative years of expertise in conducting MAs, we have synthesized and presented a detailed pragmatic approach to interpreting MAs in Neurosurgery.

**Results** Herein we have briefly defined SRs and MAs and related terminologies, succinctly outlined the essential steps to conduct and critically appraise SRs and MAs. A practical approach to interpreting MAs for neurosurgeons is described in details. Based on summary outcome measures, we have used hypothetical examples to illustrate the Interpretation of the three commonest types of MAs in neurosurgery: MAs of Binary Outcome Measures (Pairwise MAs), MAs of proportions and MAs of Continuous Variables. Furthermore, we have elucidated on the concepts of heterogeneity, modeling, certainty, and bias essential for the robust and transparent interpretation of MAs. The basics for the Interpretation of Forest plots, the preferred graphical display of data in MAs are summarized. Additionally, a condensation of the assessment of the overall quality of methodology and reporting of MA and the applicability of evidence to patient care is presented.

**Conclusion** There is a paucity of pragmatic guides to appraise MAs for surgeons who are non-statisticians. This article serves as a detailed guide for the interpretation of systematic reviews and meta-analyses with examples of applications for clinical neurosurgeons.

**Keywords** Evidence-based neurosurgery · Forest plot · Meta-analysis · Neurosurgery · Surgery · Systematic review

## Introduction

The surge in research publications in recent years makes it difficult for healthcare providers to stay current with scientific knowledge [17, 19]. Systematic Reviews (SRs) and Meta-analyses (MAs) mitigate this challenge by synthesizing data and delivering comprehensive information [19]. However, readers can only critically appraise, interpret, and

use SRs and MAs if they understand the techniques used in conducting them.

Although guides to SRs methodology and appraisal exist [1, 12, 13], they contain complex statistical terminologies and formulae difficult for physicians to decipher the subject. There is thus a paucity of pragmatic guides for the appraisal of MAs by surgeons who are non-statisticians. Herein, we have defined basic concepts in SRs and MAs and succinctly outlined the essential steps in conducting and interpreting MAs. Furthermore, we have illustrated the most commonly used tools for assessing the methodological and reporting

Extended author information available on the last page of the article

quality of SRs and MAs. Case scenarios from the field of neurosurgery exemplify these concepts.

## What are systematic reviews and meta-analyses?

SRs and MAs represent the most formal, rigorous, and extensive review of the evidence about a specific research question and, therefore, reside at the top of the evidence hierarchy. They summarize and evaluate the quality of evidence to answer specific question(s), inform best practices, elucidate persistent knowledge gaps, and attempt to clarify controversies.

A SR is a literature-based hypothesis-driven research project that aims to analyze and summarize multiple closely related primary research studies investigating a clearly formulated clinical question using systematic, explicit, and reproducible methods [17]. Importantly, the quality of the combined data is contingent on the quality of prior publications. This approach systematically identifies, selects (searches), evaluates (critically appraises), and synthesizes all available high-quality scientific evidence relevant to answering the clinical research question. By doing so, it reduces biases and random errors that are inherent in individual cohort studies.

MAs are a group of statistical techniques that enable data from two or more studies to be combined and analyzed as a single, new dataset to draw an overall conclusion. A MA leverages the increased statistical power of a larger sample size

by integrating the quantitative findings from similar studies and provides a numerical estimate of the overall effects of interest.

A MA is often performed as a quantitative part of a SR, in conjunction with the descriptive and qualitative assessments. It is possible to do a MA of observational or experimental studies; however, the MA should report the findings for these two study designs separately. This method is especially appropriate when the studies that have been reported have small numbers of subjects or come to different conclusions.

## Essential steps to conducting a SR and MA

SRs and MAs should be conducted methodically with a thorough search for all studies, appraisal of their quality, and selection of the best studies answering the question. Details on the conduct of SRs and MAs are beyond the scope of this publication, which focuses on the reading and interpretation of the combined results. However, the Cochrane Library contains the most detailed guidance on performing and appraising SRs [10], and Table 1 includes a summary of the crucial steps that should be taken when conducting an SR and MA.

The Preferred Reporting Items for SRs and MAs (PRISMA) statement should be used by researchers as a guide to conducting systematic reviews. This was developed to facilitate transparent and complete reporting of systematic reviews [15]. The initial step in conducting a SR is

**Table 1** Ten steps for conducting a Systematic Review and Meta-analysis

1. Formulate the review question
  - a. The PICO format is often used to develop a research question
  - b. Develop the review's inclusion and exclusion criteria
2. Develop a data search strategy using relevant keywords and terms
  - a. Note that various databases have different search approaches
3. Identify relevant studies by running the search strategy on relevant databases of published literature
  - a. Multiple databases are typically reviewed in an attempt to include all relevant publications
4. Extract required data
  - a. Title and Abstract screening
  - b. Full-text review
  - c. Data extraction using an extraction template

\*NB. Steps a and b are typically performed by at least two reviewers to ensure consistency of methodology and application of inclusion/exclusion criteria, with a third author acting as arbitrator to resolve any disagreements.
5. Perform a risk of bias and quality assessment of included studies
6. Synthesize and analyze the data
7. Interpret the evidence
8. Consult with subject matter experts
9. Write up and publish the review
10. Update the review

PICO = Patient/Population, Intervention, Control/Comparator/Comparison, Outcome.

For prognostic studies we use PECO with E for Exposure.

formulating a research question (hypothesis). It should be specific and focused. The questions should not be too broad or too vague. One technique that can guide researchers is identifying the patient or population, type of intervention or exposure, comparison or control, and outcome (PICO outline). This will help researchers in formulating a succinct and answerable research question.

The next step is identifying keywords or terms that will be used for literature searches in research databases. The terms related to the patient or population and intervention or exposure can be used as keywords. Boolean operators (e.g., OR, AND, and NOT) and search modifiers (parentheses, asterisk, and quotation marks) should be used to enhance the literature search. The researchers should decide on the databases that will be used for the search, the range of dates that will be included in the search, and the target language(s) of search results. University librarians can be a good resource for expert guidance, and may be consulted before conducting a literature search.

The results yielded from different databases should be collated and deduplicated, as separate research databases will retrieve an overlapping portion of articles. The deduplicated articles should be screened based on the inclusion and exclusion criteria established, with at least two authors screening at each step. The authors should cite reasons for excluding articles during the title and abstract screening. If there are disagreements in the inclusion or exclusion of articles, a discussion should be made, ideally with a third author. After the title and abstract screening stage, the full text of the included articles should be retrieved and further screened independently by at least two authors. During the full-text analysis, the researchers should extract all necessary information using the predetermined data collection tool. Some important data that should be extracted are study

authors, year of publication, the population included in the study, sample size, and occurrence of outcomes in the experimental versus the control group. This information will be used for qualitative synthesis and/or MA, if applicable. If a MA can be done, specialized Softwares can be used, such as Revman, R, and STATA.

The quality and risk of bias of the included articles should also be assessed using validated assessment tools. For example, the Cochrane Risk of Bias 2 (RoB 2) tool can be used to assess the quality of randomized controlled trials [24], and the ROBINS-I tool may be used to evaluate the strength of observational studies [23].

## Critical appraisal of SRs and MAs

It is crucial to assess the quality of the methodology and reporting of a SR and MA and the appropriateness of the MA before diving into the fine points of the results and drawing conclusions on patient treatment. The key points in the quality assessment are summarized in Table 2.

## A practical approach to interpreting SRs and MAs for surgeons

Interpretation of a MA can be achieved by answering five basic questions (modified from 4 steps outlined by Rafael Perera [16]):

- What are the study characteristics?
- What is the summary measure?
- What does the Forest Plot show?

**Table 2** Critical appraisal of SRs and MAs

Core Questions	Criteria
When can a systematic review and MA give further insight into primary study results?	<ul style="list-style-type: none"> <li>• Existing studies gave disparate results</li> <li>• Bigger study population (sample size) can increase power, generalizability and precision of findings (effect estimate)</li> <li>• Subgroup analyses may be possible and could generate new hypotheses</li> </ul>
Is the MA clinically sensible?	<ul style="list-style-type: none"> <li>• Did the studies summarized in the systematic review address the same research question?</li> <li>• Are the studies included in the MA of comparable quality (selection bias, attrition rates, confounding variables)?</li> <li>• Are the studies comparable (e.g. population, duration/dosage of treatment, outcomes reported)?</li> </ul>
Will the results help in caring for my patients?	<ul style="list-style-type: none"> <li>• Are the studied populations comparable to my patients?</li> <li>• Are the results clinically important?</li> <li>• Are all clinically important outcomes considered?</li> <li>• Were benefits, harms and costs considered?</li> </ul>

Courtesy of Karen Reid et al. [17]

- What does the pooled effect (average effect) mean?
- Was it valid to combine studies?

These questions should be reviewed while reading SRs and MAs prior to the application of findings to clinical practice. Nonetheless, one can only truly assess and interpret a MA if one knows the scientific literature body of a specific topic in question.

### Definition of terminologies

To accurately answer these key questions, some common terminology must be understood (Table 3).

### Study characteristics

A description of study characteristics is the initial step in all SRs and MAs. Classically a summary table is produced, containing, but not limited to, the following information: Study identification (usually first author's name, e.g., John et al.), study year (the year the study was published), study site, period, sample size (intervention and control groups), mean age of participants, sex and other relevant sociodemographic features.

### What is the summary outcome measure?

After the description of study characteristics, the next step will be to define the summary measure (outcome measure). In neurosurgery, three types of outcome measures are usually reported in MAs: binary, continuous, and unary outcomes. A comparison of these measures is shown in Table 4.

### Interpreting different types of MAs with forest plots

In biomedicine, forest plots are the predominant display of data, a style of data visualization for MA results. In neurosurgery, the results of MAs are usually displayed graphically as a forest plot or “blobbogram,” which provides a graphical display of the observed effect, confidence interval, and weight of each included study, as well as the overall pooled effect of the MA. These can be preferred to tables for the ease of visual comparison of precision and spread of the studies and interpretability of the combined analysis. Figure 1 shows the main components of a forest plot for a MA of binary outcomes.

It is worth noting that MAs can also be visualized via other graphs, such as drapery plots [18]. A drawback of forest plots is that they can only display confidence intervals assuming a fixed significance threshold, conventionally  $p < 0.05$ . The significance of the effect size is based on these confidence intervals, whereas recent discussions about  $p$ -values are controversial [14]. Some believe that a  $p$ -value of 0.05 is not rigorous enough. Drapery plots are based on  $p$ -value functions circumventing the sole reliance on the  $p < 0.05$  significance threshold when interpreting the results of an analysis, thus offering an alternative to the conventional  $p$ -value restrictions.

### MAs of binary outcome measures (pairwise MA)

Table 5 summarizes five hypothetical surgical trials comparing post-operative complication (CSF Leak) in patients operated for pituitary macroadenomas using either Endoscopic (ETS) or Microscopic (MTS) Trans-sphenoidal approach. Since ETS surgery is a relatively new technique,

**Table 3** Definition of common Explanatory Terms

Term	Definition
Effect size/estimate	A measure of the direction and magnitude of the change of the probability of disease following a specific exposure or treatment.
Relative risk	Quotient of the disease probability in the exposed versus unexposed groups.
Odds ratio	Ratio of the probability of an event occurring versus the probability of not occurring in the cases (exposed) compared to the controls (unexposed) groups.
Confidence interval	Upper and lower limits of values within which the true population value lies. Usually, the 95% CI is used, corresponding to the typical 5% significance level.
Forest plot	Graphical visualization of the effect estimate and CI (significance) of individual studies, as well as the overall effect and CI (significance).
Heterogeneity	Measure that describes the difference of study effects.
Bias	Distortion of study results from a systematic error.
Publication bias	Bias in publication, such that negative studies with no effect are published less than positive studies with significant effect.
Funnel plot	Illustration of publication bias.
Fixed-effects model	An assumption that each study estimates the same (true population) treatment effect.
Random effects model	An assumption that each study estimates a treatment effect that falls within a distribution or range of effects.

**Table 4** Comparison of MA of Unary, Binary, and Continuous Variables

	Unary	Dichotomous (Binary)	Continuous
<b>Definition and Example</b>	Single group of an event without a comparison or control group. Examples include the proportion of surgeries with gross total resection, prevalence, and incidence	Something that happens or not, such as presence (YES) or absence (NO) of complication	Something that, in theory, can be any number in a range of values, such as: intracranial pressure, weight, height, length of hospital stay, or volume of blood loss during surgery.
<b>Outcome Measure</b>	Percentage or proportion	Ratios such as relative risk or Odds Ratio	Difference between means such as weighted or standardized mean difference (WMD or SMD)
<b>Null Line</b>	None	1	0
<b>Outcome value</b>	0-100	Always > 0	Can be negative or positive
<b>Treatment Scale</b>	Treatment scale range from 0-100.	Favors intervention on left hand side of the treatment scale (ratio < 1)	Favors intervention on left hand side of the treatment scale (WMD < 0)
<b>If outcome effect measure is negative (e.g. complication occurring)</b>			
<b>OR</b>			
<b>If decreased outcome effect measure is desirable (e.g. decrease in intracranial pressure)</b>			
<b>Treatment Scale</b>	Treatment scale range from 0-100.	Favors intervention on right hand side of the treatment scale (ratio > 1)	Favors intervention on right hand side of the treatment scale (WMD > 1)
<b>If outcome effect measure is desirable (e.g. stopped stroke)</b>			

Other outcome measures: ordinal, counts and incidence rate ratio (IRR)  
 Correlation Coefficient@: Strength of relationship between two variables  
 Time to event: Single group measure (Life Tables, K-M Curve) or comparative effect measure (hazard ratio (HR))  
 (Courtesy of Karin Ried, 2006) [17]

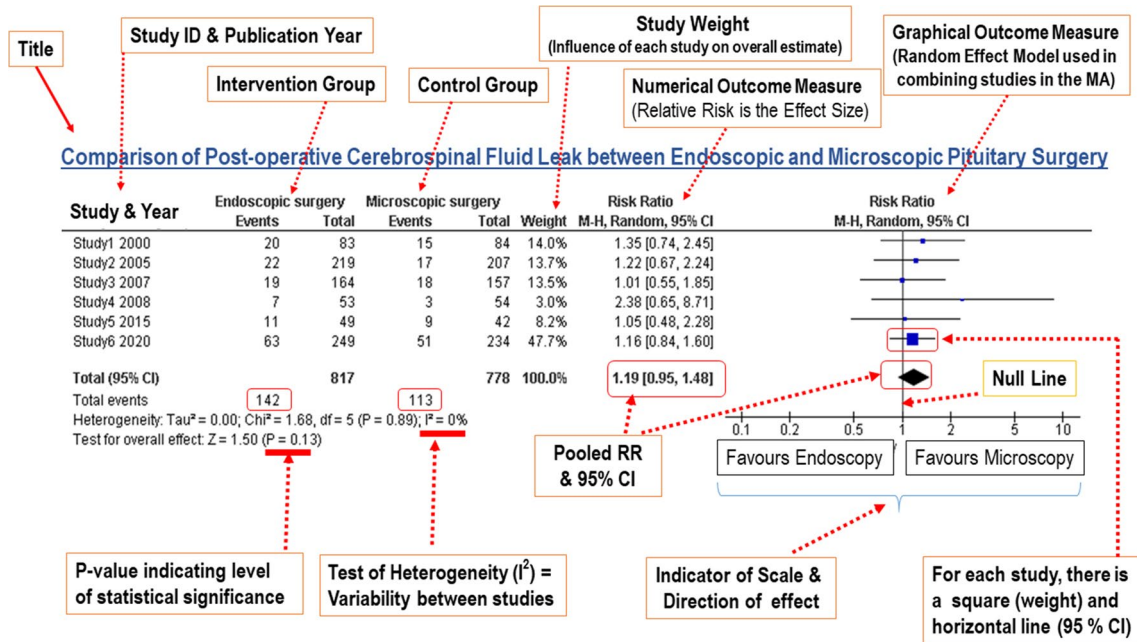


Fig. 1 Forest plot of pairwise MA (Comparison of CSF Leak between endoscopic and microscopic pituitary surgery)

Table 5 Comparison of CSF between endoscopic and microscopic pituitary surgery

Study	Year	Endoscopy			Microscopy (n/N)				
		Total Number of Patients (N)	CSF Leak (n)	No CSF Leak (n/N)	Total Number of Patients (N)	CSF Leak (n)	No CSF Leak (n/N)		
Study 1	2000	83	20	63	20/83	84	15	69	15/84
Study 2	2005	219	22	197	22/219	207	17	190	17/207
Study 3	2007	164	19	145	19/164	157	18	139	18/157
Study 5	2008	53	7	46	7/53	54	3	51	3/54
Study 6	2015	49	11	38	11/49	42	9	33	9/42
Study 7	2020	249	63	186	63/249	234	51	183	51/234

it is considered as the Intervention group (Exposed Group), while the MTS (the standard/old technique) is the control group (Non-Exposed Group).

- **Title of graph:** First, identify the details of the MA displayed at the top (above) the graph usually in the “PICO format” (Problem or Research question), Intervention and Control groups and the Outcome measure). For example, A Comparison of CSF Leak in Endoscopic versus Microscopic Pituitary Adenoma Surgery.
- **Study identities and Publication Year:** Studies included in the MA and incorporated into the forest plot will generally be identified in chronological order on the left-hand side by author and date. No significance is given to the vertical position assumed by a particular study. The first Column (Study) lists the individual study types (study identification) included in the MA. The first author’s name is usually displayed. The Second Column is the Publication Year of the articles.
- It is common to see columns 1 and 2 (Author’s name and Year) combined in a single column.
- **Events in Intervention and Control:** These are presented in columns 3 and 4, respectively, as number of events with outcome (n) on the total number of events (N) and displayed as n/N. Sometimes they are displayed to the right of the forest plot.
- **Study Results.**
- The individual study results are displayed in rows.

- *Outcome effect measure* is displayed graphically and numerically. For binary data, outcome values are always greater than “0”.
- *The Weight* of each study is the influence of each study on the overall estimate. To calculate the summary or pooled effect, different weights are assigned to the different studies included in the MA. The weight of each study varies inversely to the standard error (and therefore indirectly to the sample size) reported in the studies. Studies with large standard errors and small sample sizes are given less weight in calculating the pooled effect size and vice versa.
- *Null line*: Classically, a forest plot has a vertical reference line in the center. It indicates the point on the x-axis equal to no effect called the “line of no effect”. If a result (horizontal Line) touches or crosses it, then the result is not statistically significant (RR or OR = 1), implying no association between exposure (intervention) and outcome or no difference between the two interventions.
- *Value axis and Scale of measurement* is the Indicator of Scale and Director of measurement. It is worth noting that the direction or orientation of the outcome measure is not standardized. Thus, the details of the value axis should always be checked. Some forest plots display the intervention on the left (as in our example) or right of the null line.
- *Boxes and horizontal lines (whiskers)*: Each horizontal line with a square (box) in the middle represents the results of the different studies included in the MA.
- The squares or boxes (sometimes circles) represent the point estimates (studies’ “results”). The size is proportional to the weight the study has in the MA. The bigger the box, the larger the sample size of the study and the higher the statistical power and, thus, the more confident we can be about the result.
- The length of that horizontal line represents the length of the confidence interval (CI). This gives us an estimate of how much uncertainty there is around that result – the wider it is, the less confident we can be about the result, and vice versa.
- The size of the point estimate is echoing the length of the confidence interval. They are two perspectives on the same information. Large squares with short lines provide more confidence than small squares with long lines.
- If the CI crosses the line of no effect (RR = 1), then the study results are not statistically significant.
- *Arrowed horizontal line*: Sometimes the confidence interval is too wide beyond the scale used in the forest plot and is thus truncated and displayed as an arrow as in study 5 in Fig. 1.
- Conventionally, a forest plot should also contain the “effect size data” that was used to perform the MA for transparency and reproducibility of results.
- **Overall Results**: The overall effect size is represented graphically by the ‘diamond’ and numerically by its ‘pooled effect size and CI’.
- *The diamond* is the summary estimate. It represents the summary of the results from all the studies combined. It is the combined effect size and confidence interval. If the diamond touches the line of no effect, it means the overall pooled effect size is not statistically significant. If the diamond is to the left, then there are more episodes of the outcome in the treatment group, and if to the right, there are more episodes in the treatment group (Fig. 1).
- *The overall numeric pooled effect size*: It is the weighted average with larger studies having more events counting more. It is not the mathematical summation of the results of the seven studies (Fig. 1) and then dividing them by 7!
- *The Confidence Interval*: The length of the diamond symbolizes the confidence interval of the pooled result on the x-axis. The lower and the upper limits of the CI of the pooled effect size are the left and right tips of the diamond. The tips get closer together with the addition of each study to the plot (narrow CI) and will move left or right depending on the direction the study’s result tips the scales toward.
- *Heterogeneity (Test of Heterogeneity)*: The agreement or disagreement between the studies is assessed using different measures of heterogeneity. Typically a forest plot is enhanced to display heterogeneity measures such as  $I^2$  or  $\tau^2$ . Here  $I^2 = 0\%$ , meaning there is no heterogeneity and less variability thus, studies are combinable. The  $I^2$  should be less than 50%.
- *Model*: It is the Method of combining studies. In Fig. 1, the weights are from the Random Effect Model, although the fixed effect model could be used since  $I^2 = 0\%$ .
- It is worth noting that, in a forest plot, the “Effect size” and “confidence intervals” are typically displayed on a linear scale. However, a logarithmic scale on the x-axis is commonly used when the summary measure is a ratio such as odds ratios or risk ratios. Here, the reference line is at 1, indicating no effect. This makes sense for ratios since these effect size metrics cannot be interpreted in a “linear” fashion (i.e., the “opposite” of RR = 0.50 is 2, not 1.5 if it were to be linear). Table 6 summarizes the essential points in interpreting the forest plot of a pairwise MA.

### MA of proportion (proportional MA)

Neuroepidemiological and Neurosurgical research often involves MAs of studies that report certain proportions (e.g., disease prevalence, incidence, and proportion of

**Table 6** Essential Elements to Interpret a Forest Plot

1. Names on the left	First author of primary studies
2. Solid Squares (Boxes)	Effect size of studies (RR, OR or WMD)
3. Solid Square Size	Weight of each study
4. Horizontal lines	95% Confidence Interval
5. Vertical Line	Line of No Effect
6. Diamond	Overall Treatment Effect
7. Diamond Centre	Pooled or Combined Treatment Effect
8. Tips of Diamond	95% Confidence Interval of pooled treatment effect
9. Heterogeneity and significance	Gives an idea about the variability of the studies combined. It should be $I^2 < 50\%$
10. Model	Fixed or Random Effect Model Used

**Table 7** MA of the proportion of patients with Gross Tumor Resection

Study	year	Number of Cases with Gross Tumor resection	Total	Proportion of Cases with Gross Tumor resection (%)
Study 1	2000	12	15	80
Study 2	2001	53	68	78
Study 3	2002	91	146	62
Study 4	2005	45	63	71
Study 5	2005	10	15	67
Study 6	2005	278	300	93
Study 7	2009	51	57	89

(Courtesy of Tabaei et al. 2009) [5]

patients with gross tumor resection). Proportional MAs are statistical methods used in MAs to combine results from multiple independent studies reporting proportions or rates and addressing the same research question. They are a subtype of “Single Group MAs”. They differ from “Pairwise MAs” comparing therapeutic effects aimed at estimating different relative effects such as odds ratio, risk ratio, or risk difference [14]. In Neurosurgery, an intervention-oriented specialty, MAs of proportions can be used to address the effectiveness of treatment or intervention, albeit unable to provide information about causality. MAs of proportions are restricted within 0–100% and focus on estimating the overall (population-averaged) proportion.

Table 7 shows hypothetical data from seven descriptive cohort studies [5] assessing the proportion of patients with gross tumor resection after endoscopic trans-sphenoidal pituitary adenoma surgery.

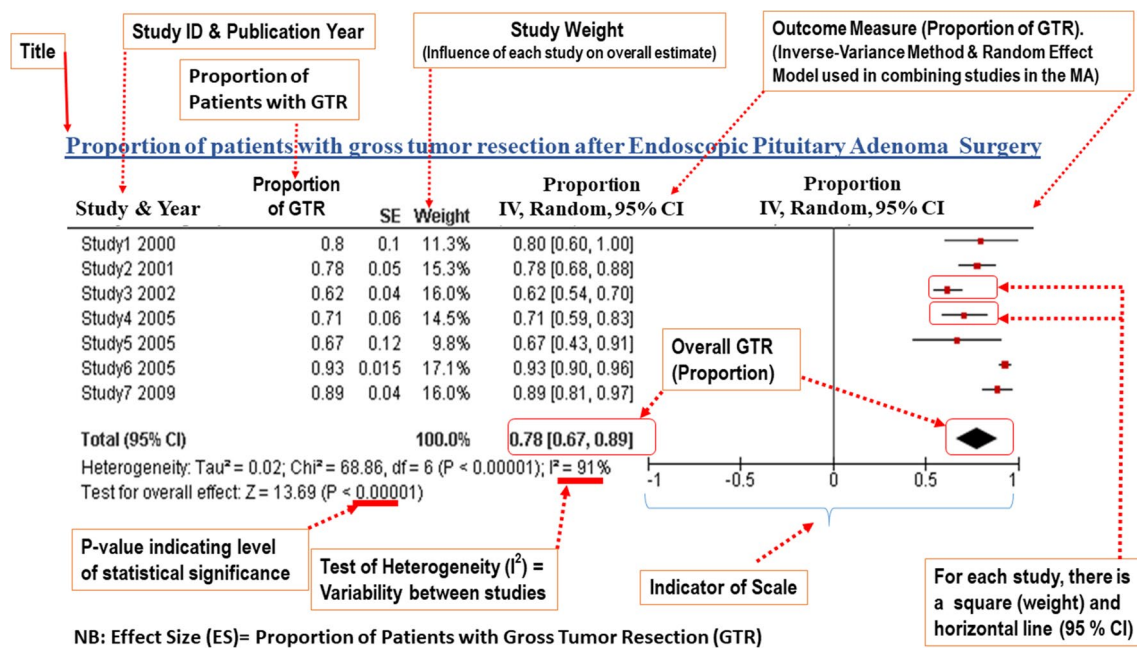
The forest plot below was obtained after pooling the data together (Fig. 2). The presentation of a proportional MA forest plot may appear differently depending on the software used in designing it. Nonetheless, the figure ought to provide the following information: Title of the graph, Study identities and Publication Year, Events in Intervention (Exposed) group (actual number of events and total group size), Study Results (Summary estimates from

each study with their confidence interval, the weight of each study) and Overall Results (final pooled proportional estimate, confidence interval, measure of heterogeneity, formula and model used).

The interpretation of the forest plot is similar to that of comparative pairwise (two groups) analysis but with distinct differences underscored below.

- *Title of graph and Study identities* (Publication Year) are interpreted as for comparative MAs.
- *Events in Intervention (Exposed) group*: Actual number of events in the intervention (exposed) group and total group size for each study. Here there is no control group as for forest plots of pairwise (comparative) MAs.
- *Studies' Results* (Summary estimates from each study with their confidence interval and weight of each study). The individual study results are displayed in rows *with the Outcome effect measure* displayed graphically and numerically.
- *The proportion and their respective 95% CI* are displayed for each study. *The weight* of each study is the influence of each study on the overall pooled proportion (based on sample size).
- *Null line*: Unlike pairwise MA, there is no “Null Line” (“line of no effect”) for proportional MA.
- *The Scale of measurement* is the Indicator of Scale and Director of measurement: MAs of proportions are restricted within 0–100%.
- *Boxes and horizontal lines*: Studies included in the MAs are represented in the forest plot as squares (boxes) with horizontal lines on both sides of the box.
- The squares or boxes are graphical representations of the effect sizes or point estimates (“proportion”). This point estimate is supplemented by a horizontal line, which represents the range of the confidence interval calculated for the observed effect size. The length of the horizontal line (confidence interval) varies from 0 to 100%. There are no negative values.
- *Overall Results*. Just like for pairwise MAs, the overall effect size is represented graphically by the ‘diamond’





**Fig. 2** Forest plot of a proportional MA (proportion of patients with gross tumor resection)

and numerically by its ‘pooled proportion and its Confidence Interval’.

- *The diamond* is the summary estimate of the pooled proportion. In our example (Fig. 2) it is 0.78 or 78% meaning the proportion of people with gross tumor resection was 78% (i.e. 540 of the 664 patients had their pituitary adenomas grossly resected).
- A vertical line via the “middle of the diamond” signifies the overall pooled proportion.
- *The Confidence Interval*: The 95% CI is the most commonly used in clinical research and in our example, it ranged from 82 to 97%. The lower and the upper limits of the confidence interval of the pooled proportion are the left and the right tips of the diamond.
- The numeric ‘pooled proportion and its Confidence Interval’ are reported to the right of the diamond.
- The *P*-value is the test of the overall effect and represents the statistical significance of the overall pooled results. Typically a *p*-value of < 0.05 implies statistical significance.
- *Heterogeneity (Test of Heterogeneity)*: There are no specific tests to assess heterogeneity in proportional MA. I<sup>2</sup> is the recommended surrogate statistic since it was developed for pairwise MAs [2].
- *Model*: A random effect model is recommended for proportional MAs [2].

## MA of continuous variables

*Notion of Mean Difference, Weighted Mean Difference, and Standardized Mean Difference.*

In statistics, a quantitative variable can be discrete or continuous if they are obtained by counting or measuring, respectively. A continuous variable can take an uncountable set of values.

A MA of continuous variables pools many studies by comparing the mean values of the variables of interest. The weighted mean difference (WMD) is presented if the means are reported in the same unit and the standardized mean difference, (SMD) is used when the means are reported in different units. Table 8 summarizes the differences between mean difference, weighted mean difference, and standardized mean differences (Table 8).

Here, we have used a hypothetical example comparing the length of hospital stay between endoscopic and microscopic pituitary surgery to illustrate the interpretation of forest plots of continuous variables MA (Table 9).

The forest plot obtained after pooling the mean length of hospital stay is displayed in Fig. 3.

The interpretation of the forest plot for continuous data is slightly different from that of comparative pairwise (two groups) or proportional MAs.

**Table 8** Definitions of mean difference, weighted mean difference, and standardized mean difference

	When it is used	Why it is used	Statistical Equation	Interpretation
Mean Difference	-Within one study -Measures in the same unit	Compares one group to another group	$MD = M_1 - M_2$	$MD > 0$ = Intervention group had a positive effect compared to control group $MD = 0$ = No difference between the intervention and control groups $MD < 0$ = Intervention group had a negative effect compared to control group
Weighted Mean Difference	-MA -Measures in same unit	-Pooling many studies together -Compares one group to another group	$WM = \sum(\text{mean value} \times \text{Weight value})$ $WMD = WM_1 - WM_2$ OR $WMD = \sum(MD \times \text{Weight value})$	$WDM > 0$ = Intervention had a positive effect compared to control group $WMD = 0$ = No difference between the intervention and control group $WMD < 0$ = Intervention had a negative effect compared to control group
Standardized Mean Difference	-MA -Measures in DIFFERENT units	-Pooling many studies together -Compares one group to another group	$SMD = MD/SD$	$SDM > 0$ = Intervention had a positive effect compared to control group $SMD = 0$ = No difference between the intervention and control group $SMD < 0$ = Intervention had a negative effect compared to control group

(Courtesy of Kristy Yee

Youtube Video: Link: <https://youtu.be/b91npgFfmGs?list=PLzoVKyjcmSw67c9aNdjADm-ZrD6zVvjYP>)

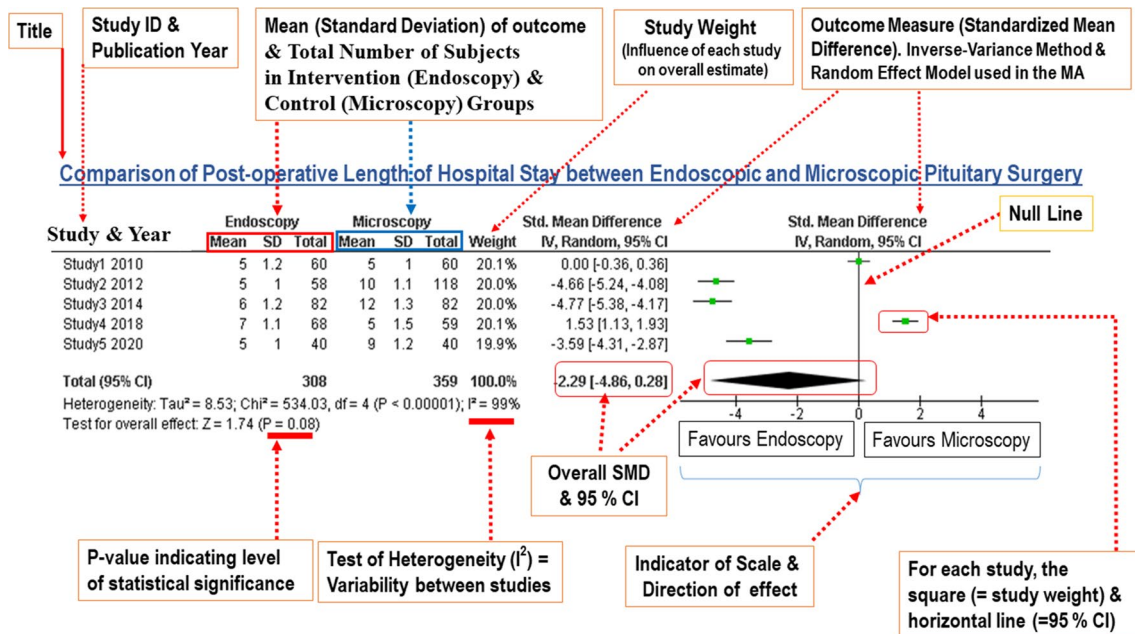
**Table 9** Comparison of length of hospital stay between endoscopic and microscopic pituitary surgery

Study	Year	Endoscopy			Microscopy		
		n	mean	SD	n	mean	SD
Study 1	2010	60	5	1.2	60	5	1
Study 2	2012	58	5	1	118	10	1.1
Study 3	2014	82	6	1.2	82	12	1.3
Study 4	2018	68	7	1.1	59	5	1.5
Study 5	2020	40	5	1	40	9	1.2

- *Title of graph and Study identities* (Publication Year) are interpreted as comparative and proportional MAs.
- *Variables in Intervention and Control Groups*: These are presented in columns 3 and 4 and respectively represent the Intervention and control groups. Three variables are displayed for each group viz.: the total number of participants (n), their arithmetic means ( $\bar{x}$ ), and the respective standard deviations (SD) of the outcome measure in that order.
- *Study Results*: The individual study results are displayed in rows with the *Outcome effect measure* displayed graphically and numerically.
- *Null line*: The vertical reference line in the center is the 'line of no effect' and typically has a value of zero ("0")

for continuous variables. Studies such as "Study 1" (in Fig. 3) cross the null line, meaning there is no statistically significant difference in the length of hospital stay between the Intervention (Endoscopy) and Control group (Microscopy) ( $SMD = 0$ ).

- The value axis is at the bottom of the graph.
- *The Scale of measurement of the treatment effect* is the Indicator of Scale and Director of the measurement. In our example, (Fig. 3) the intervention (endoscopy) is on the left, and the control (microscopy) is to the right of the null line.
- *Boxes and horizontal lines (whiskers)*: The squares or boxes representing the point estimates are situated in line with the outcome value of the included studies (SMD).



**Fig. 3** Forest plot of continuous variables MAs

The sizes of the boxes are directly related to the weight of the studies included in the MA. The bigger the box, the larger the sample size of the study and the higher the statistical power.

- **The confidence interval (CI)** is depicted by the length of the horizontal lines via the boxes. The longer the length, the wider the CI, and the less confident or precise we can be about the result. If the CI crosses the line of no effect (SMD = 0) then the study results are not statistically significant.
- The **Weight** of each study (expressed as a percentage) is the influence or weighting of each study included in the MA on the overall pooled estimate. The weight or influence of a study on the pooled estimate varies with the sample size of the study and precision (confidence interval). Classically, the larger the sample size, the narrower the confidence interval, and the greater the weight of the study (the more influence it has on the overall pooled estimate).
- **Overall Results.** The overall effect size is represented graphically by the ‘diamond’ and numerically by its ‘pooled WMD and its Confidence Interval’.
- **The diamond** (last row of the forest plot) is the summary estimate of pooled SMD. In our example (Fig. 3) the numerical pooled SMD is -2.3 days and 95% CI (-4.9 -0.28). This means patients who were operated on for pituitary adenomas using endoscopy had a shorter hospital stay of 2.3 days compared to the microscopic group. The difference was not statisti-

cally significant since the diamond crossed the null line ( $p$ -value = 0.08). A vertical line via the “middle of the diamond” signifies the overall SMD.

- **The Confidence Interval:** The lower and the upper limits of the confidence interval of the pooled SMD are the left and right tips of the diamond. The 95% CI is the most commonly used in clinical research. In our example, the numerical values for the 95% CI ranged from -4.9 to 0.28 days. This 95% CI includes a zero value for SMD so the results are not statistically significant.
- **The P-value:** The probability value represents the statistical significance of the overall pooled results (= test of overall effect). Typically, a  $p$ -value of < 0.05 implies statistical significance.
- **Heterogeneity (Test of Heterogeneity):** Heterogeneity is assessed as for pairwise MA. I<sup>2</sup> is the recommended statistic. In Fig. 3, it was 99% (very high).
- **Model:** The REM was thus used because of the high heterogeneity.

**Forest plot layout types** There is the “RevMan5” layout which produces a forest plot similar to the ones generated by Cochrane’s Review Manager 5. Another common type is the **JAMA layout**, which displays the forest plot in accordance with the guidelines of the Journal of the American Medical Association. This layout is accepted by most medical journals.

### Appropriateness of pooling studies

The correct interpretation and critical appraisal of SRs and MAs enhances the appropriate application of evidence to patient care. Four main domains have been identified to improve on the robustness and transparency: heterogeneity, modeling, certainty, and bias [4, 9, 13].

### Heterogeneity [4, 7, 9, 13, 17]

Heterogeneity tells us if the studies included in the MA were similar to each other and were combinable. It depicts the degree of overlap or variability that exists amongst studies pooled in the MA. This is usually displayed in the forest plot at the bottom left. Studies that do not overlap well are referred to as heterogeneous. Results from such data are not definitive and uncombinable, but more conclusive results are obtained when the studies' results are more similar or homogenous. Heterogeneity or inconsistency can be clinical, methodological, or statistical in nature.

- **Clinical heterogeneity.** Results from differences in participants, interventions, or outcomes. It results from normal or expected variation between studies and stems from different patient assessment criteria, drugs, doses, and assessment tools among others. Clinical sense or judgment should be used in whether to accept this variation or not.
- **Methodological heterogeneity:** Differences in study design that can affect the comparability of the data and variations may increase risk of bias.

- **Statistical heterogeneity** is the true heterogeneity that affects MA results stemming from the variation in intervention effects or results. The variability among studies in a MA is the sum of true heterogeneity and within-study error. The degree of inconsistency of studies included in a MA can be evaluated using different techniques including a visual assessment of the reported 95% CI and via statistical testing for heterogeneity (Cochran's,  $p$ -value,  $T^2$  and Tau and  $I^2$  statistics) [4].

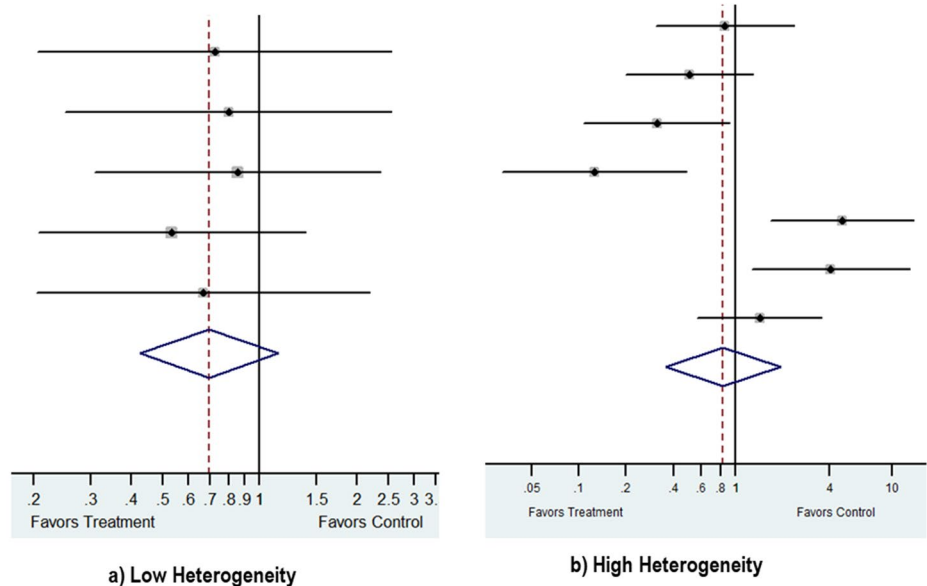
#### Visual Method.

- **Eye-ball Test:** When horizontal lines or whiskers in a forest plot overlap, the studies are said to be homogenous or consistent. Here the focus is on the overlapping confidence intervals rather than on which side the effect estimates fall (Fig. 4).
- **Imaginary line method:** Visualizes the forest plot by drawing an imaginary vertical line through the upper and lower tips of the diamond (the overall pooled estimate). The studies are said to be homogenous if the line crosses all the confidence interval (horizontal) lines. In Fig. 4, the imaginary line crosses the horizontal CI lines in plot (a) (homogeneity) but not in plot (b) (heterogeneity).
- **Numerical value of CI method:** The studies included in the MA are said to be homogenous if ALL the lower limit of CI numbers of each study are lesser than the upper CI numbers of all studies. For example, consider an MA of three studies with the following 95% confidence intervals:

Study 1: CI (0.1–4.5), Study 2: CI (0.5–2.1), and Study 3 CI (1.4–5.5). Here, the lower limits of CI for the three

Fig. 4 Visual assessment of heterogeneity

Visual Methods for the assessment of Heterogeneity: Eye Balling Test and Imaginary Line Method



studies (0.1, 0.5, 1.4) are smaller than the upper limits of CI (4.5, 2.1, 5.5). Thus, these three studies would be considered homogeneous. However, the 95% CI is limited in its ability to specifically quantify the inconsistency of the studies [4].

**Statistical methods** Quantification of heterogeneity is commonly done with  $X^2$ ,  $P$ -value,  $T^2$ , and  $I^2$ .

- **Cochran's Q Chi<sup>2</sup> (X<sup>2</sup>) Test (Q Statistics):** If  $X^2 > \text{degree of freedom}$ , it implies heterogeneity, but if  $X^2 \leq \text{degree of freedom}$ , it implies the results could be homogeneous. (Degree of freedom (df) = Number of Studies – 1). However, Q is statistically underpowered (fails to detect heterogeneity) when the number of studies is low and the sample size within the studies is small. Similarly, the power is too high for MA with many studies.
- **$P$ -value of the  $X^2$  test:** If the  $p$ -value  $< 0.10$ , it implies heterogeneity, but if the  $p$ -value is  $\geq 0.1$ , it implies the results could be homogeneous. Cochran's Q Chi<sup>2</sup> (X<sup>2</sup>) and  $P$ -value of the  $X^2$  tests have limited discriminatory power to distinguish homogeneity from true heterogeneity. They provide information about the statistical significance of the heterogeneity of the studies but not the extent. Q tests have low power when the number of studies is small, as in our example (and most MAs). In addition, a non-significant Q test does not provide evidence that the effect sizes are constant but may be due to a lack of power because of the small number of studies or small studies having overlapping confidence intervals (high within-study variance). Similarly, the Chi<sup>2</sup> (X<sup>2</sup>) Test assumes the null hypothesis that each study is measuring an identical effect (i.e., all the studies are homogeneous). The  $p$ -value test gives us a  $p$ -value to test this hypothesis. If the  $p$ -value of the test is low, we can reject the hypothesis, and heterogeneity is present. Notice the cut-off for a  $p$ -value of  $< 0.1$  is used instead of the classical  $< 0.05$  as the cut-off. The test is often not sensitive enough, and the wrong exclusion of heterogeneity is common. For these reasons, other tests should be used to check for heterogeneity, such as the  $I^2$  test.
- **$T^2$  and  $Tau$ :** Both  $T^2$  and  $Tau$  are statistical tests often displayed in MA result tables. They are measures of the dispersion (variance) of true effect sizes between studies in terms of the scale of the effect size.  $T^2$  is not used in itself as a measure of heterogeneity but is used to compute  $Tau$  and to assign weights to the studies in the MA under the random-effects model (REM).  $Tau$  estimates the standard deviation of the distribution of true effect sizes under the assumption that these true effect sizes are normally distributed.  $Tau$  is a useful first indication of the extent of this dispersion.  $Tau$  is used for computing the *prediction interval* (a more direct and more easily interpretable indicator).

- **$I^2$  test:** The  $I^2$  test is the most reliable to assess heterogeneity. Unlike the Q test, the  $I^2$  test quantifies the heterogeneity and is independent of the number of studies included. The  $I^2$  test reflects the extent of overlap of CIs: the larger the  $I^2$ , the lesser the overlap. Higgins et al. proposed a scale for the assessment of heterogeneity [11]. It is expressed as a percentage with a range from 0 to 100%. In common practice, if  $I^2 \leq 50\%$ , the studies are considered homogeneous, and if  $I^2 \geq 50\%$ , they are considered heterogeneous. Heterogeneity is a fundamental problem that should be avoided during the conduct of MAs and managed after pooling by post-hoc analyses [9]. Inconsistency should be explored graphically in the forest plot and statistically by assessing the 95% CI, Cochran's Q, and  $I^2$  statistical tests. The assessment of inter- and intra-study variability or comparability determines the choice of model or MA technique for pooling. The fixed effect model can be used if the studies are homogeneous ( $I^2 \leq 25\%$ ). The random effect model should be used if heterogeneity is high ( $I^2 \geq 75\%$ ).

For proportional MA,  $I^2$  is only a surrogate statistic. True heterogeneity is expected in prevalence/incidence estimates due to differences in the time and place where the study was carried out. Thus, high  $I^2$  in proportional MA does not necessarily imply inconsistencies in data [2]. Chi-squared test and  $Tau$  can also be applied to assess heterogeneity [2].

*How to deal with heterogeneity?* In the presence of heterogeneity, several methods can be used, including changing the outcome measure (e.g., relative risk instead of odds ratio), presenting the results qualitatively as an SR (no pooling of results), meta-regression, post-hoc subgroup analysis, and sensitivity analysis. Sometimes, the

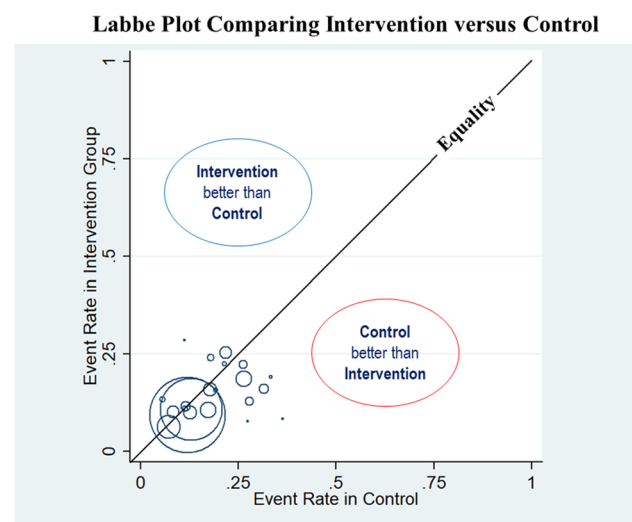


Fig. 5 Interpretation of a labbe plot

heterogeneity is ignored if the 95% CI of all studies lies on the same side of the overall pooled estimate.

The Labbe Plot is commonly used to explore and explain heterogeneity (Fig. 5). It is a plot of the event rate in the intervention group (y-axis) versus control group (x-axis). The intervention is better than the control if most of the studies (circles) lie to the left of the line of equality and vice versa.

## Modeling

Two common models are used in MAs: fixed effect model (FEM) and random effect model (REM). There are four widely used MA methods for dichotomous outcomes: three fixed-effect methods (Mantel-Haenszel, Peto, and inverse variance) and one random-effects method (DerSimonian and Laird inverse variance). Each method has specific indications determined by the statistician.

FEM is performed assuming that all the included studies share a common effect size and that factors that can influence the effect size are the same in all studies. REM assumes that the true effect sizes are normally distributed. Generally speaking, whereas the FEM focuses solely on the selected studies included in the MAs, the REM takes into consideration that there might be other studies unpublished, not picked up in the search strategy or to be undertaken in future which were not included in the MA at hand [6]. The FEM is appropriate if the studies in the MA share a common effect size (homogeneity) for the identified population. The goal is not to generalize to other populations. On the contrary, the REM should be used if it is very unlikely that the pooled studies share a common effect size and the goal is to generalize the results to a wider population. However, REM is commonly used because the relative weights of studies are more balanced in REM than FEM. The choice of the model used is important for pairwise MA as FEM and REM give different results, whereas, for continuous variables, the results of the MA using either model are often identical [6]. Particularly for proportional MA, REM is recommended. This is because when considering epidemiological factors typically measured using proportional data (prevalence/incidence), they are well known to vary between population characteristics, whereas FEM assumption that there is one true estimate is unlikely to hold true [2].

## Certainty [8]

How certain the results of a MA are is an essential question. If appropriate pooling is done, the certainty is increased. Certainty of evidence or in the effect estimates is the extent of confidence that the treatment effect revealed by the research is accurate. The GRADE (Grading of

Recommendations Assessment, Development, and Evaluation) criteria, although not fully endorsed and integrated into neurosurgical literature, is an accepted tool for assessing certainty. Details of this grading system are beyond the scope of our review.

## Bias [3, 4, 25]

The extent to which a MA can draw precise conclusions about the effect size depends on the validity of the data combined from the primary studies. Bias can lead to misinformation about the true effect size of the intervention or exposure and can stem from the primary studies or the MA itself.

For the primary studies, errors or bias can result from the systematic differences in the baseline properties of the study groups ('selective bias'), in therapeutic care (performance bias), as a result of dropouts in the study groups (attrition bias), outcome assessment (detection bias) or selective reporting of outcome (outcome reporting bias).

Reporting bias is an umbrella term that refers to systematic differences between reported and unreported findings. Reporting biases can include publication bias, citation bias, language bias, and time lag bias. In MAs, it is usually checked for during sensitivity analysis to explore the influence of included studies, graphically via funnel plot or quantitatively using statistical tests [20]. Although subjective, publication bias can be detected visually by looking for the asymmetrical distribution of the dots (studies) around the pooled effect line. If both heterogeneity and bias are absent, then 95% of studies are expected to lie within the diagonal dotted '95% Confidence Interval' lines.

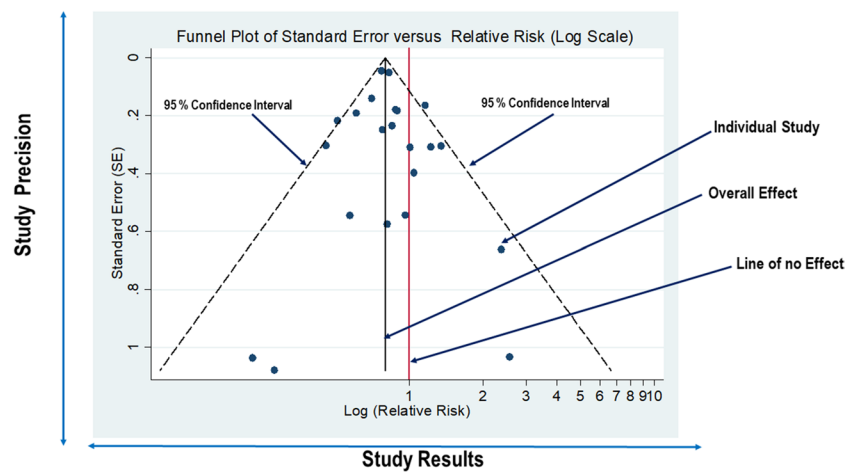
Biases in published SRs and MAs are of significant concern, as their outcomes have far-reaching implications for clinical decisions, policymaking, and patient care. It is imperative to have a comprehensive understanding of these biases because healthcare choices and recommendations can be compromised when the evidence presented in SR and MA is incomplete or skewed.

## Funnel plots

A funnel plot is a scatter plot that compares the precision (how close the estimated effect size of the intervention is to the true effect size) and the results of individual studies. In MAs, it is a visual aid to detect and assess publication bias if there is asymmetrical distribution around the overall effect or pooled result line.

In funnel plots, the precision of the estimated intervention effect increases with the sample size of the study. Large study effect estimates are typically concentrated at the top of the graph. Larger studies have smaller spreads as they are more precise and closer to the true effect (Fig. 6).

**Fig. 6** Funnel plot of standard error versus relative risk (Log Scale)



Key aspects of the funnel plot include the following:

**Y-axis** The y-axis represents a measure of study precision. Smaller studies with lower precision are at the bottom, while studies with larger sample sizes, thus, greater precision, are displayed at the top. The standard error is the most common measure of study precision. Others include the reciprocal of the standard error or sample size and the variance of the estimated effect. The lesser the standard error, the larger the sample size. Hence, studies with bigger sample sizes are in the uppermost portion of the funnel plot.

**X-axis** The x-axis represents the study estimated effect size of the outcome. The x-axis scale can be a ratio (plotted on a logarithmic scale) such as RR and OR or a continuous measure such as weighted mean difference (WMD).

**Middle vertical line** The middle vertical line of the plot represents the overall or pooled estimate results.

**Scattered dots** Individual studies included in the MA are represented as dots.

**Dash lines** The dashed or confidence interval lines represent the 95% CI. The further the dash lines are from the pooled effect line, the wider the confidence interval.

In MA, publication bias is checked quantitatively (via statistical tests) or graphically (using a funnel plot). Statistical tests for bias lack power if the number of trials included is less than 25. Funnel plots and tests for symmetry are used to check for bias in MAs. They should only be used when the number of studies included in the MA is at least 10 studies, as the power of the tests is low when there are fewer studies.

Numerous statistical tests exist to assess publication bias. Egger's and Begg's tests are common. A publication

bias is said to be present if the test is statistically significant ( $p < 0.05$ ). Unlike the graphical visualization, these tests are more objective. However, their applicability for proportional MA has not been validated [2]. Publication bias should, therefore, be assessed qualitatively [2].

**Graphical Asymmetry:** Although subjective, publication bias can be detected visually by looking for the asymmetrical distribution of the dots (studies) around the pooled effect line. If both heterogeneity and bias are absent, then 95% of studies are expected to lie within the diagonal dotted '95% Confidence Interval' lines.

Common sources of funnel plot asymmetry include, but are not limited to, non-reporting bias, true heterogeneity, data irregularities (poor methodological quality leading to exaggerated effects), artifact (heterogeneity due to poor choice of effect measurement), or random error because of the small number of studies in the MA. ROBIS (Risk of Bias in Systematic Reviews) tool is a common tool to assess the risk of bias in SRs and MA rather than in the primary studies [23, 24].

## Assessment of the overall quality of methodology and reporting of MAs

ROBIS is a tool designed to assess the risk of bias in systematic reviews [3]. ROBIS has three phases: (1) assess relevance (optional) (2), identify concerns with the review process (study eligibility criteria, identification and selection of studies, data collection and study appraisal, and synthesis and findings), and (3) judge the risk of bias. The results of the ROBIS assessment could include summarizing the number of systematic reviews that had a low, high, or unclear concern for each phase 2 domain and the number of reviews at high or low risk of bias. The advantage of using ROBIS is that it covers systematic reviews on interventions, diagnosis, prognosis, and etiology.

Another tool for the appraisal of SRs and MAs is the AMSTAR 2 tool. This instrument is only applicable to a systematic review of healthcare interventions. The old version AMSTAR tool [21] was developed to assess the quality of systematic review of randomized trials. However, due to the increasing number of systematic reviews that include randomized and non-randomized studies of healthcare intervention, the AMSTAR 2 [22] was developed. This instrument consists of 16 items that assess various aspects, including the research question, inclusion and exclusion criteria, adherence to a well-developed protocol, selection of study designs, comprehensive literature search strategy, study selection, data extraction, excluding studies, and justification of exclusion. It also evaluates the description of included studies, assesses the risk of bias in individual studies, reports on the sources of funding for the studies included in the review, uses an appropriate statistical combination of results, assesses the potential impact of risk of bias in individual studies on the results of the MA, discussion of any heterogeneity, publication bias, and potential sources of conflict of interest. These questions will guide the reviewers in assessing their overall confidence in the review results. The ratings can be critically low (more than one critical flaw with or without non-critical weaknesses), low, moderate, or high (with zero or one non-critical weakness).

## Applicability of evidence from MAs

While MAs are extremely valuable to the field, neither conducting nor interpreting a MA is easy. Fundamentally, the quality of data outputs is highly contingent on the inputs; MAs that aggregate randomized control trial data are much less susceptible to bias compared to non-randomized primary studies. Thus, one should pay close attention to the types of studies included in the assessment. After a judicious evaluation of the effect measure and pooled effect size, one needs to ascertain the validity of the results. Although intuition and clinical reasoning are necessary for judging the results, the application of findings should follow the principles of evidence-based neurosurgery: integrating the best available research evidence, surgical experience, expertise and judgment; and patient values and preferences to provide optimal patient care. This requires checking for similarity between the results of the MA and your individual patients, an assessment of the risk versus benefit, and finally, the cost of care. The clinical and statistical significance of the results should be critically analyzed before applying evidence.

**Supplementary Information** The online version contains supplementary material available at <https://doi.org/10.1007/s00701-024-06133-8>.

**Funding** None.

**Data availability** Not applicable.

**Code Availability** Not applicable.

## Declarations

**Ethics approval** Not applicable.

**Consent to participate** Not applicable.

**Consent for publication** Not applicable.

**Conflict of interest** The authors declare no competing interests.

## References

- Barker FG, Carter BS (2005) Synthesizing medical evidence: systematic reviews and metaanalyses. *Neurosurg Focus* 19(4):E5
- Barker TH, Migliavaca CB, Stein C, Colpani V, Falavigna M, Aromataris E et al (2021) Conducting proportional meta-analysis in different types of systematic reviews: a guide for synthesisers of evidence. *BMC Med Res Methodol* 21(1):189
- Borenstein M, Hedges LV, Higgins JPT, Rothstein HR (2009) *Introduction to Meta-Analysis*. First Ed. ed: John Wiley & Sons, Ltd
- Esene IN, Commentary (2020) Pearls for interpreting neurosurgical systematic reviews and Meta-analyses: lessons from a collaborative effort. *Neurosurgery* 87(5):E594–E5
- Esene IN, Ngu J, El Zoghby M, Solaroglu I, Sikod AM, Kotb A et al (2014) Case series and descriptive cohort studies in neurosurgery: the confusion and solution. *Childs Nerv Syst* 30(8):1321–1332
- Fleiss JL (1993) The statistical basis of meta-analysis. *Stat Methods Med Res* 2(2):121–145
- Fletcher J (2007) What is heterogeneity and is it important? *BMJ* 334(7584):94–96
- Guyatt G, Oxman AD, Akl EA, Kunz R, Vist G, Brozek J et al (2011) GRADE guidelines: 1. Introduction-GRADE evidence profiles and summary of findings tables. *J Clin Epidemiol* 64(4):383–394
- Haines SJ, Commentary (2020) Pearls for interpreting neurosurgical systematic reviews and Meta-analyses: lessons from a collaborative effort. *Neurosurgery* 87(3):E275–E6
- Higgins J, Thomas J, Chandler J, Cumpston M, Li T, Page M et al (2023) *Cochrane Handbook for Systematic Reviews of Interventions*: Wiley
- Higgins JP, Thompson SG, Deeks JJ, Altman DG (2003) Measuring inconsistency in meta-analyses. *BMJ* 327(7414):557–560
- Lee KS, Zhang JY, Nga VDW, Ng CH, Tai BC, Higgins JPT et al (2022) Tenets for the proper Conduct and Use of Meta-analyses: a practical guide for neurosurgeons. *World Neurosurg* 161:291–302e1
- Lu VM, Graffeo CS, Perry A, Link MJ, Meyer FB, Dawood HY et al (2020) Pearls for interpreting neurosurgical systematic reviews and Meta-analyses: lessons from a collaborative effort. *Neurosurgery*
- Mayo DG, Hand D (2022) Statistical significance and its critics: practicing damaging science, or damaging scientific practice? *Synthese* 200(3):220
- Page MJ, Moher D (2017) Evaluations of the uptake and impact of the Preferred reporting items for systematic reviews and Meta-analyses (PRISMA) Statement and extensions: a scoping review. *Syst Rev* 6(1):263




16. Perera R, Heneghan C (2008) Interpreting meta-analysis in systematic reviews. *Evid Based Med* 13(3):67–69
17. Ried K (2006) Interpreting and understanding meta-analysis graphs—a practical guide. *Aust Fam Physician* 35(8):635–638
18. Rucker G, Schwarzer G (2021) Beyond the forest plot: the drapery plot. *Res Synth Methods* 12(1):13–19
19. Scheidt S, Vavken P, Jacobs C, Koob S, Cucchi D, Kaup E et al (2019) Systematic reviews and Meta-analyses. *Z Orthop Unfall* 157(4):392–399
20. Sedgwick P (2015) What is publication bias in a meta-analysis? *BMJ* 351:h4419
21. Shea BJ, Grimshaw JM, Wells GA, Boers M, Andersson N, Hamel C et al (2007) Development of AMSTAR: a measurement tool to assess the methodological quality of systematic reviews. *BMC Med Res Methodol* 7:10
22. Shea BJ, Reeves BC, Wells G, Thuku M, Hamel C, Moran J et al (2017) AMSTAR 2: a critical appraisal tool for systematic reviews that include randomised or non-randomised studies of healthcare interventions, or both. *BMJ* 358:j4008
23. Sterne JA, Hernán MA, Reeves BC, Savović J, Berkman ND, Viswanathan M et al (2016) ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. *BMJ* 355:i4919
24. Sterne JAC, Savović J, Page MJ, Elbers RG, Blencowe NS, Boutron I et al (2019) RoB 2: a revised tool for assessing risk of bias in randomised trials. *BMJ* 366:14898
25. Whiting P, Savović J, Higgins JP, Caldwell DM, Reeves BC, Shea B et al (2016) ROBIS: a new tool to assess risk of bias in systematic reviews was developed. *J Clin Epidemiol* 69:225–234

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

## Authors and Affiliations

Ignatius Esene<sup>1,2</sup>  · Ourlad Alzeus G. Tantengco<sup>3</sup> · Faith C. Robertson<sup>4</sup> · Megan E. H. Still<sup>5</sup> · Alvan-Emeka K. Ukachukwu<sup>6</sup> · Ronnie E. Baticulon<sup>7</sup> · Ernest J. Barthélemy<sup>8</sup> · Daniela Perez-Chadid<sup>9</sup> · Laura Lippa<sup>10</sup> · Ana Cristina Veiga Silva<sup>11</sup> · Luxwell Jokonya<sup>12</sup> · Fahd D. Hassani<sup>13</sup> · Federico Nicolosi<sup>14</sup> · Berjo D. Takoutsing<sup>2</sup> · Jeff Ntalaja<sup>15</sup> · Samer S. Hoz<sup>16</sup> · Kazadi K. N. Kalangu<sup>12</sup> · Gilbert Dechambenoit<sup>17</sup> · Franco Servadei<sup>18</sup> · Najia El Abbadi<sup>13</sup> · Kee B. Park<sup>19</sup> · Angelos Koliass<sup>20,21</sup>

✉ Ignatius Esene  
ignatiusesene@yahoo.co.uk

Ourlad Alzeus G. Tantengco  
ogtantengco@up.edu.ph

Faith C. Robertson  
faith\_robertson@hms.harvard.edu

Megan E. H. Still  
megan.still@neurosurgery.ufl.edu

Alvan-Emeka K. Ukachukwu  
alvan.ukachukwu@duke.edu

Ronnie E. Baticulon  
ronnie.baticulon@gmail.com

Ernest J. Barthélemy  
globalneurosurgeon@gmail.com

Daniela Perez-Chadid  
danielaperezchadid@gmail.com

Laura Lippa  
laura.lippa@gmail.com

Ana Cristina Veiga Silva  
christyanne77@hotmail.com

Luxwell Jokonya  
doclux@gmail.com

Fahd D. Hassani  
fahdderkaoui@gmail.com

Federico Nicolosi  
fn@upsurgeon.com

Berjo D. Takoutsing  
takoutsingberjo@gmail.com

Jeff Ntalaja  
drntalaja@yahoo.fr

Samer S. Hoz  
hozsamer2055@gmail.com

Kazadi K. N. Kalangu  
kazkal2003@gmail.com

Gilbert Dechambenoit  
gdechambenoit@gmail.com

Franco Servadei  
franco.servadei@hunimed.eu

Najia El Abbadi  
najiaelabbadi@gmail.com

Kee B. Park  
kee\_park@hms.harvard.edu

Angelos Koliass  
ak721@cam.ac.uk

- 1 Neurosurgery Division, Faculty of Health Sciences, University of Bamenda, Bamenda, Cameroon
- 2 Research Division, Winners Foundation, Yaounde, Cameroon
- 3 Department of Physiology, College of Medicine, University of the Philippines Manila, Manila, Philippines
- 4 Department of Neurosurgery, Massachusetts General Hospital, Boston, MA, USA
- 5 Department of Neurosurgery, University of Florida, Gainesville, FL, USA
- 6 Department of Neurosurgery, Duke University Health System, Durham, NC, USA

- 7 Division of Neurosurgery, Department of Neurosciences, Philippine General Hospital, University of the Philippines, Manila, Philippines
- 8 Global Neurosurgery Laboratory, Division of Neurosurgery, SUNY Downstate Health Sciences University, Brooklyn, NY, USA
- 9 Faculty of Medicine, Universidad CES, Medellin, Colombia
- 10 Department of Neurosurgery, ASST Ospedale Niguarda, Milan, Italy
- 11 Neurosurgery Department of Post Graduation Program in Neuropsychiatry and Behavioral Sciences (PosNeuro), Federal University of Pernambuco, Recife, Brazil
- 12 Department of Surgery, College of Health Sciences, University of Zimbabwe, Harare, Zimbabwe
- 13 Department of Neurosurgery, Cheikh Zaid International Hospital, Abulcasis International University of Health Sciences, Rabat, Morocco
- 14 Department of Medicine and Surgery, Neurosurgery Unit, University of Milano-Bicocca, Milan, Italy
- 15 Clinique Ngaliema, République Démocratique du Congo, Kinshasa, Congo
- 16 Department of Neurosurgery, University of Cincinnati College of Medicine, Cincinnati, OH, USA
- 17 Centre Medical Chirurgical Obstetrical Cote d'Opale, Saint Martin Boulogne, France
- 18 Department of Biomedical Sciences, Humanitas University, Milan, Italy
- 19 Program in Global Surgery and Social Change, Harvard Medical School, Boston, MA, USA
- 20 Division of Neurosurgery, Addenbrooke's Hospital, University of Cambridge, Cambridge, United Kingdom
- 21 NIHR, Global Health Research Group on Acquired Brain and Spine Injury, Department of Clinical Neurosciences, University of Cambridge, Cambridge, UK