Interventional Pain Medicine 1 (2022) 100120

Contents lists available at ScienceDirect

Interventional Pain Medicine

journal homepage: www.journals.elsevier.com/interventional-pain-medicine

Interpretation of meta-analyses

Pascal Richard David Clephas [a], Michael Heesen [b,*]

[a] Dept. of Anaesthesia, Erasmus University Medical Center, Dr. Molewaterplein 40, 3015GD, Rotterdam, the Netherlands
[b] Dept. of Anaesthesia and Pain Medicine, Kantonsspital Baden, Im Ergel 1, 5404, Baden, Switzerland

ARTICLE INFO

Keywords:
Systematic review
Meta-analysis
Network meta-analysis
Living meta-analysis
Trial sequential analysis
Bayesian statistics

ABSTRACT

Our article provides guidance on how to interpret a meta-analysis and introduces the reader to the basics of the underlying statistical analysis. The multiple steps of a meta-analysis including systematic literature search, risk of bias assessment, data extraction and data aggregation are addressed. The Grading of Recommendations Assessment, Development and Evaluation (GRADE) approach allows to score the quality of the evidence of the results revealed by a meta-analysis. Trial sequential analysis has been suggested in recent years as a method to assess the power of a meta-analysis and the risks of false positive or false negative conclusions. We also provide information on other more complex meta-analytical approaches including network meta-analysis for the comparison of several treatments as well as recent developments such as individual patient data meta-analysis and living meta-analysis.

1. Introduction: what are meta-analyses good for?

Conclusions derived from properly conducted systematic reviews (SR) and meta-analyses (MA) are considered the highest attainable level of evidence within Evidence Based Medicine (EBM), both according to traditional as well as newer pyramids of evidence [1]. This type of study is characterized by systematic methodology, making it reproducible and transparent. In addition, a MA pools multiple, preferably similar, studies which allows for greater statistical power to calculate an overall effect. With an increase in the practice of EBM, SRs and MAs have gained paramount importance in the development of clinical guidelines [2], as reflected by over 20.000 publication in 2021 in Pubmed carrying the term MA in the title (Fig. 1). While properly conducted SRs and MAs help to improve EBM through updating clinical guidelines, improperly conducted ones have the potential to do the opposite, which emphasizes the importance of using the right methods and resources for reporting and analysis of this type of study [3]. SRs and MAs can be challenging, especially given the advances in methodology that have been evolved in this field over the years [4]. Some of the greatest advances were the development of reporting checklists [5,6], which can aid authors in what to include in their article and helps to keep the reporting of SRs and MAs uniform, especially since most journals require the use of these reporting checklists. Excellent guidance is also available for conducting SRs and MAs in the well-known Cochrane Handbook for Systematic Reviews of Interventions [7].

There are 2 situations in which MAs are helpful. Often there are studies that show an effect of an intervention (or a diagnostic test or a prognostic factor or model) whereas others do not show an effect, a third group of studies may find the superiority of the comparator group which may be a control (placebo) group or an alternative intervention.

Aggregating all available data will help to answer the question whether there is an effect across all studies or not. Fig. 2 shows the example of a forest plot with studies displaying divergent results [8]. The forest plot gives the effect estimate (square) and the 95% confidence interval (CI) (line) of each study as well as the combined effect estimate and combined 95% CI (last line of the plot). There are studies on the right side of the plot (eg the study by Uysallar) contrasting with studies on the left side side (e.g. the study by Ko) of the vertical line (mean difference of 0) that indicates equipoise between the study arms. If the 95% CI represented by the line crosses the vertical 0 line (e.g. the study by Thoren) then the difference is not statistically significant different.

A less frequent situation is that all studies report an effect of the intervention, with studies finding a strong effect and other studies finding either a weak effect or a statistically not significant effect (95% CI line crossing 0). MA will allow to estimate the «true » effect size. Fig. 3 gives the forest plot of such an analysis of studies with similar results [9].

2. Structure of meta-analyses

2.1. Research question

Each SR and MA starts with a research question, which requires to

* Corresponding author.
E-mail address: Michael.Heesen@ksb.ch (M. Heesen).

https://doi.org/10.1016/j.inpm.2022.100120
Received 14 May 2022; Received in revised form 4 June 2022; Accepted 7 June 2022

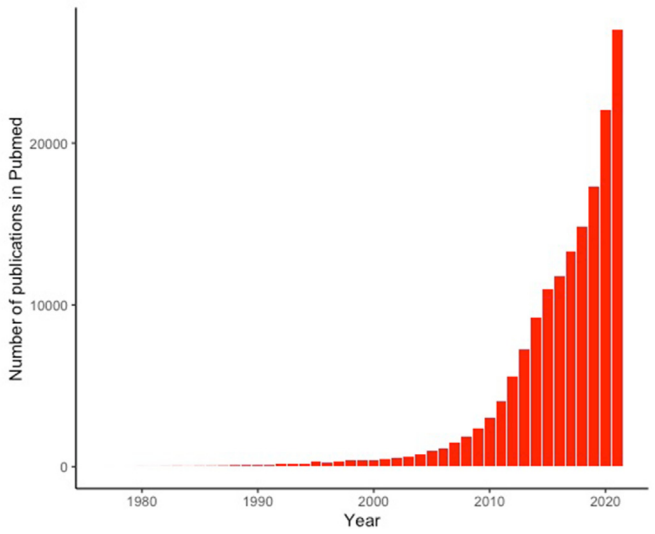

Fig. 1. Title: Number of publications in Pubmed with "meta-analysis" in the title by year.

define primary and secondary outcomes. Theoretical frameworks are available for conceptualizing and structuring a research question, the most well-known one being the PICO(S), an acronym which stands for P: Participants/Patients, I: Intervention, C: Comparator, S: Setting [10].

Other theoretical frameworks are available for more specific types of systematic reviews and meta-analyses [11]. Even in the presence of previous SRs and MAs a new SR and MA may be justified. This can be the case when new important studies have been published or when the previous SR and MA focused on a different subgroup. It is important for the SR and MA to be conducted that it fills in knowledge gaps and adds value to the current state of knowledge. The research question as well as details of the methodology should be defined in a study protocol. Most journals require an a priori registration of a SR and MAs as well as of the research question and the methodology e.g. in the international prospective register of systematic reviews (PROSPERO) or a publication of the research protocol. Registering a SR and MA forces authors to think in advance of what the research question and methodology will be, in addition registering a SR and MA makes it less likely for authors to deviate from their original plan, which avoids reporting bias. Moreover, registering a SR and MA informs other researchers that a SR and MA on a particular topic is already being conducted so that duplication of efforts as well as the waste of research resources can be prevented. It is also mandatory that SRs and MAs follow the guidelines of the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) statement [5].

2.2. Systematic literature search and selection criteria

It is of paramount relevance that all available evidence is gathered in a SR and MA. Therefore, the search strategy is crucial as it influences the

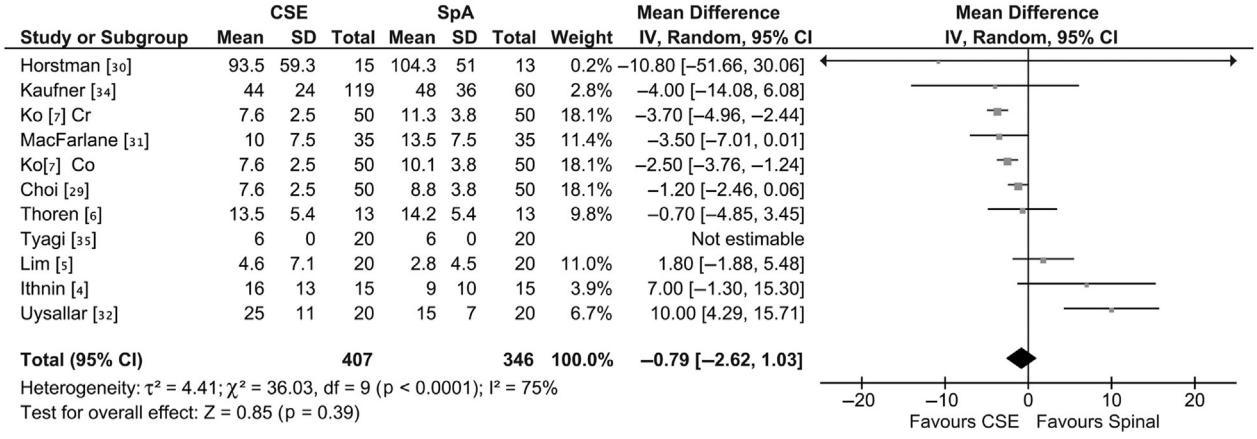

Fig. 2. Meta-analysis of studies with divergent results.
Forest plot of meta-analysis of comparative vasopressor use (mg ephedrine equivalent) for combined spinal-epidural aneasthesia versus spinal anaesthesia for caesarean section. Co: colloids; Cr: crystalloids; CSE: combined spinal-epidural anaesthesia; IV: inverse variance; SpA: Spinal anaesthesia.

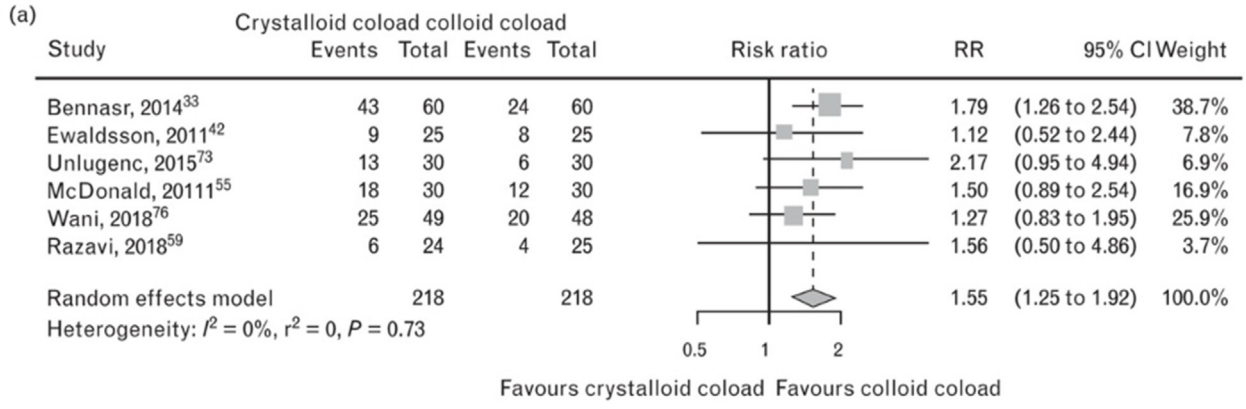

Fig. 3. Meta-analysis of studies with similar results.
Forest plot of meta-analysis of comparative effectiveness of crystalloid coload versus colloid coload in the treatment of shock and hypovolemia.

sensitivity (the number of relevant articles retrieved) and specificity (the number of non-relevant articles retrieved) of the systematic literature search. The search strategy consists of standardized keywords, such as MESH terms, and free text, combined with Boolean operators (e.g. AND, OR, NOT). The standardized keywords are database specific and therefore written search strategies often need to be adapted to different databases. It has been shown that a search should be performed in at least the following databases: MEDLINE, Embase, Web of Science, and Google Scholar (the first 200 references retrieved) [12]. The work that is involved when developing a search strategy with both good sensitivity and specificity and making it compatible for multiple databases is not easy by any means. The search strategy can be very elaborate and complex, as shown by the previously published study protocol of a SR and MA [13]. It is recommended to consult an information specialist or even to include one as a co-author [14]. Detailed guidance on developing systematic literature searches is available [15,16]. Inclusion and exclusion criteria for the selection of articles should be defined in the study protocol. The selection criteria include the type of study population, type of outcome, type of study or even the time of publication or language used, although the last one may not be needed as Google Translate has been reported to provide valid results [17]. Clear selection criteria can improve the inter-rater reliability and therefore the efficiency of the next step, which is the selection process.

2.3. Selection process

The selection process should be done by at least two independent reviewers to ensure accurate results, each reference should be screened by at least two reviewers. The selection process consists of two phases: a title and abstract phase and a full text phase. In the title and abstract phase, references are screened by title and abstract simultaneously. In this phase, it is not necessary to provide reasons for exclusions, not fulfilling the selection criteria is enough [5]. In the full text phase, the selection criteria are applied to the full text of the references and the reasons for exclusion should be provided in the PRISMA study selection flowchart [5]. Multiple software packages are available to perform the selection process, including EndNote, Rayyan, Covidence, and DistillerSR. The selection process can be time-consuming, suggestions have been made to streamline the process [18]. Disagreements about the questions which references are included and excluded can occur as a result of each reference being screened by at least two reviewers, resulting in at least two judgments. Disagreements should be resolved by discussion or, if not possible, by the final judgement of a third reviewer.

2.4. Data extraction

Similar to the selection process, the data extraction should be done by at least two reviewers independently. A standardized way of collecting the data is, however, needed to ensure uniform data and accurate results. This can be done with a standardized data collection form used by each reviewer.

2.5. Risk of bias

The reliability of the results of a SR and MA also depends on the bias of the studies included. Conclusions drawn from studies with low quality (high risk of bias) often have a low certainty whereas conclusions drawn from studies with high quality (low risk of bias) often have a high certainty. Evaluating the risk of bias of studies is therefore essential to determine the implications a conclusion of a SR and MA has. The most common types of bias are selection, performance, detection, reporting, publication, and attrition bias [19]. Selection bias occurs when the study population differs systematically from the population of interest. Selection bias is also possible when study groups differ systematically in ways other than the intervention of interest, resulting in bias by confounding [19]. Performance bias occurs when the care provided differs systematically between the study groups, which often inflates the effect estimate of the intervention [19]. Detection bias occurs when there are systematic differences between the study groups in the determination or detection of outcomes [19]. Reporting bias occurs when researchers withhold information related to the methods or results from a study [19]. Publication bias occurs when the likelihood of publishing depends on the results of the study, which causes a difference in publication between studies with negative and positive results [19]. Attrition bias occurs when there are systematic differences in participant loss or drop-out between the study groups [19]. Several checklists exist for assessing the quality of studies, often developed for specific types of studies. For randomized controlled trials (RCTs), version 2 of the Cochrane risk-of-bias tool for randomized trials (RoB 2) is nowadays the recommended checklist [20], while for observational studies the Newcastle-Ottawa Scale is preferred [21]. Other checklists available are the Risk Of Bias In Non-randomized Studies of Interventions (ROBINS-I) [22], Quality Assessment of Diagnostic Accuracy Studies 2 (QUADAS-2) [23], Quality In Prognosis Studies (QUIPS) [24], and Prediction model study Risk Of Bias Assessment Tool (PROBAST) [25]. These checklists each have study type-specific domains to score on risk of bias, which can then be used to give a judgment of the overall quality of the study. As with the selection process and the data extraction, it is recommended to have each reference assessed on risk of bias by at least two reviewers to ensure accuracy.

2.6. Meta-analysis

Prior to performing the MA, it needs to be determined whether the collected data from the studies are suitable for pooling. This decision mostly depends on the available data and the heterogeneity between the studies, the latter being discussed in the next section. Performing a MA is only useful when meaningful results can be obtained, which is not the case when the pooled studies have considerable heterogeneity in design, study population, (statistical) methods, or outcomes. However, when the studies are suitable for pooling, a MA increases the sample size and provides more statistical power for calculating the effect estimates, which is especially useful when the pooled studies have reported conflicting results. Different models are available for MA, with the main groups being the fixed effect and random effects models. These two models have different assumptions [26], which should influence which one to use. The fixed effect model assumes that all studies have one and the same effect and differences among the studies are attributed to sampling error. The random effects model assumes that there are different effects among the studies due to heterogeneity and accounts for this by including random effects for each study, at the cost of wider confidence intervals of the pooled effect estimate. Multiple software packages are available for performing the MA, with the most used one being Review Manager (RevMan). Other options include the Metafor package in R [27], and the metan command in STATA [28].

2.7. Heterogeneity

Heterogeneity plays a major role in performing MAs, as it determines whether a MA should be performed at all, and if one is performed, which model should be used. As mentioned earlier, heterogeneity between studies on similar topics can be present due to differences in study design, study population, (statistical) methods, or outcomes. It is possible to statistically quantify the heterogeneity with either the Cochrane's Chi squared test (Cochran's Q) or Higgins's I^2 statistic. The first one tests the null hypothesis that all studies have the same effect and the second one represents the variation percentage that is attributed to heterogeneity and not sampling error, with percentages $<25\%$ being considered low, 25–50% moderate, and $>75\%$ as high heterogeneity [29]. Most meta-analytical software packages provide both statistics. The sources of the heterogeneity can be explored when a considerable heterogeneity is present, mainly through subgroup analyses and meta-regressions.

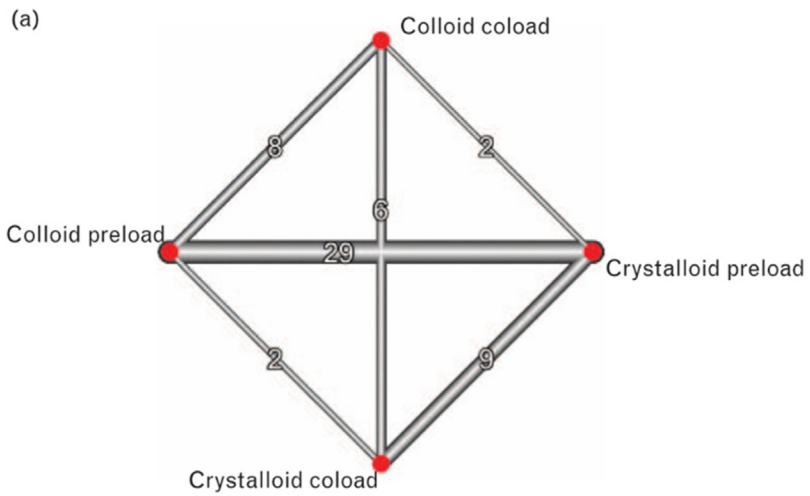

Fig. 4. Network meta-analysis, representation of number of studies.
Network geometry for the comparative effectiveness of crystalloid coload versus crystalloid preload versus colloid coload versus colloid preload in the treatment of shock and hypovolemia.

Subgroup analyses can be used to stratify for certain study characteristics, for example the study population, which can result in considerably different results within the subgroups. Meta-regression is similar to traditional forms of regression, with as the dependent outcomes the individual study effect estimates and as independent covariates certain study characteristics, such as time of publication of the study [30]. This allows to assess whether a certain study characteristic has a significant effect on the effect estimate. Similar to traditional regression analyses, meta-regression requires at least ten studies for each covariate included in the model.

2.8. Certainty of evidence/GRADE

The strength or certainty of evidence of a SR and MA is determined by the quality of the results. To score the quality of the results, the Grading of Recommendations Assessment, Development and Evaluation (GRADE) approach has been developed [31]. The quality of the results is scored based on the magnitude of effect, the results of the risk of bias assessment, consistency of results, directness of the evidence, imprecision, and publication bias [31]. The certainty of evidence can be scored as high, moderate, low, or very low.

3. Network meta-analysis (NMA)

In many clinical scenarios more than one treatment option is available, and the clinician is interested in the best treatment for his patient. Whereas conventional MA compares a treatment against placebo or treatment A versus treatment B, network MA (NMA) allows for the analysis of a multitude of treatments. In general, the algorithms of network MA programs combine head-to-head comparisons and network

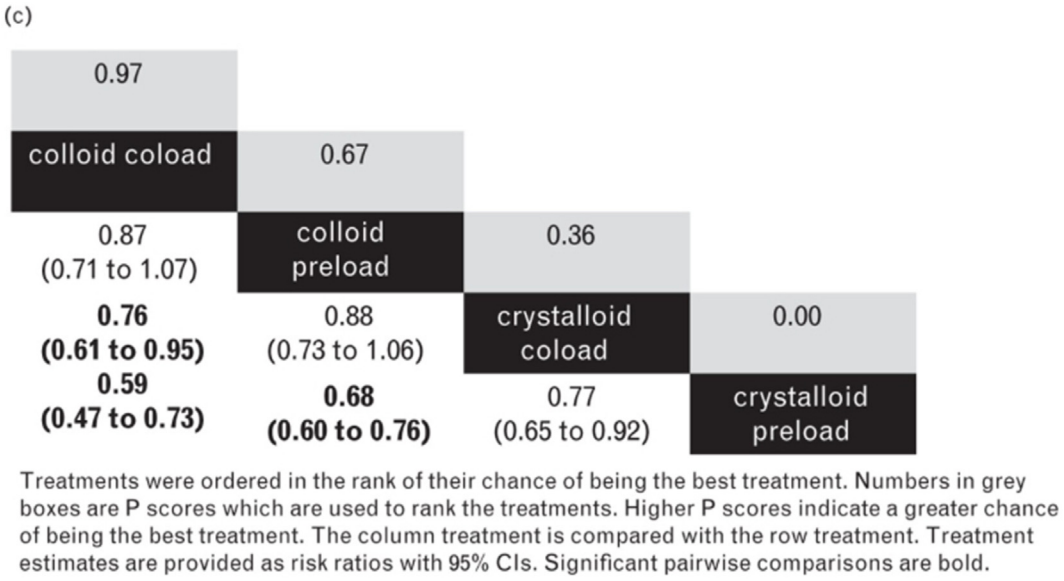

Fig. 5. Network meta-analysis, ranking of results.
League table of crystalloid coload versus crystalloid preload versus colloid coload versus colloid preload in the treatment of shock and hypovolemia sorted by rank.

comparisons. Thereby, NMA also produce comparisons of treatments for which no head-to-head studies are available. This can be best illustrated by the following example: we assume there are studies comparing treatment A versus B and studies comparing B versus C but no studies of A versus C. A NMA approach is able to compute an effect estimate of the difference between treatment A and C. In addition to this advantage a NMA may be more rigorous than two conventional MAs (one comparing A versus B and another one comparing B versus C) when all comparisons are subjected to the same study methodology. Figs. 4 and 5 come from a recent NMA and provide the typical information given in a NMA [9]. Faltinsen et al. summarized important information about NMA [32].

4. Statistics in meta-analyses

4.1. P-value

The p-value is a measure of evidence against a null-hypothesis. After a researcher has set up a null-hypothesis, data are collected, and a p-value can be computed. Since the intrinsic characteristic of a null-hypothesis is that it can only be rejected, a p-value can only quantify evidence against and not for this hypothesis. Therefore, it is important to keep in mind that a statistically significant p-value (a p-value below the pre-defined alpha level, aka the level of type 1 error rate) means that the null-hypothesis is unlikely, but not that a non-significant p-value can prove a null-hypothesis to be likely. In other words: No evidence of effect is not evidence of no effect. However, evidence of no effect can be calculated using Bayes factors. The reader is pointed towards a recent review of Bayes factors for a more detailed explanation of this [33]. Due to the very limiting property of the p-value, the p-value is often misinterpreted to be something more than it really is. The only correct definition of the p-value is: the likelihood of seeing data, this or more extreme, even though the null-hypothesis were true. In the recently refined statistical guidelines for authors of the New England Journal of Medicine (NEJM), some of the limitations of p-values are addressed [34]. The NEJM recommends replacing p-values with estimates of effect or association along with a 95% confidence interval (unless p-values were adjusted for multiplicity). The reader is referred to publications from the American Statistical Association for further delineation of the limitations of p-values. We want to close the discussion of p-values with a quote from a recent publication about statistical reporting guidelines: 'Finally, the notion that a treatment is effective for a particular outcome if $P < 0.05$ and ineffective if that threshold is not reached is a reductionist view of medicine that does not always reflect reality.' [35].

4.2. Effect estimates and confidence intervals

While the commonly used p-value, can only give information about how unlikely a null-hypothesis is, an effect-estimate can quantify the effect that an intervention or a prognostic variable has, and is therefore clinically more relevant.

In general, every point effect-estimate should be accompanied by a corresponding confidence interval. This is due to the fact that a point-estimate does not convey a lot of information since there is lots of uncertainty around its true value (statisticians sometimes call the point-estimate a 0% confidence interval).

4.2.1. 95% Confidence interval (CI)

A CI is calculated using the standard error which is a measure of dispersion. A big standard error, i.e., an estimate with lots of uncertainty around its true value, for example due to low sample size, will lead to a larger CI. Therefore, the width of the interval can tell us about the precision of our estimation. A CI is also commonly used for hypothesis testing. If the CI crosses the value stated as the null-hypothesis (most often the point of no effect, i.e., OR or RR = 1, MD = 0), the null-hypothesis cannot be rejected. The only correct interpretation of CI is: When repeating an experiment 100 times, the true value will lie 1-alpha times (when using a 95% CI: 95 times) within the CI. Therefore, the CI does not tell us about the probability of the point estimate being true or the probability that the point-estimate lies within the CI. Because the latter is the definition of the Bayesian Credible Interval, it is important to keep the definition of the CI and of the Bayesian Credible Interval apart.

Table 1
2 × 2 Contingency table depicting the calculation of Odds ratio (OR) and Relative risk (RR).

	Response	Non-response	
Treatment	a	b	a+b
Control	c	d	c + d
	a+c	b + d	TOTAL

Calculation of RR: (a/a+b)/(c/c + d).
Calculation of OR: (a/b)/(c/d).

There are several commonly used effect estimates for categorical data: Odds ratio, Relative risk, Risk reduction, Absolute risk reduction and more. Most commonly used in MAs are Odds ratio and Relative risk.

4.2.2. Odds ratio (OR)

The OR is the ratio between the odds of a certain event A (e.g., treatment response) happening in a group X (e.g., treatment arm of a trial) and the odds of this certain event A happening in a group that is not X (e.g., placebo arm of a trial). The difference to the relative risk is that it is computed in a different way (Table 1), but it is therefore also interpreted in a different way. A OR of 2.0 means that the odds (not the risk or probability) are increased by 100% or by a factor of 2. Note here that there is a relationship between odds and probability where odds are Probability/1-Probability (a probability of 50% is equal to odds of 1). A OR of 1 is the point of no effect because then the odds of event A are equally likely in both groups that are compared. Furthermore, a OR of 1 (or 50% probability) is equal to the probability of flipping a (fair) coin. The OR is probably more commonly used than the Relative risk and there might be a few reasons for this: ORs can be obtained from logistic regression which uses the logit (log odds) for modelling, while Relative risks cannot readily be obtained from logistic regression. Furthermore, the odds of a complement event can be computed by inverting the odds ratio. If for example the OR of treatment response was 2, then the OR of no treatment response would be ½. This is not true for the Relative risk. However, a disadvantage of the OR is that it is more difficult to understand than the Relative risk since the human mind is more used to thinking in probabilities than thinking in odds. This is the reason why the OR is often misinterpreted as the Relative risk. In a given situation, the OR tends to be larger than the Relative risk and misinterpretation would therefore inflate the (mis-)interpreted effect [36]. This is the reason why the statistical reporting guidelines of the NEJM advise that ORs should be avoided [34].

4.2.3. Relative risk (RR)

The RR gives is a ratio between the probability of a certain outcome in group A and the probability of a certain outcome in group B. Just like with OR, a value of 1 indicates no effect. Neither the OR nor the RR can be calculated when there are 0 events in the control group (as division through 0 is not possible). However, this problem is handled by most statistics programs by adding 0.5 to each field in the contingency table.

Since RR is a more intuitive estimate it may be useful to calculate a RR from an OR. The Cochrane handbook of systematic reviews provides a formula for this calculation [37].

Moreover, from the risk ratio the number needed to treat can be computed, details are again given by the Cochrane Handbook for Systematic Reviews of Interventions [38].

Neither the OR nor the RR can be calculated when there are 0 events in the control group (division through 0 is not possible). In order to make computation possible 0.5 to each group is added.

Since RR is a more intuitive estimate it may be useful to calculate a RR

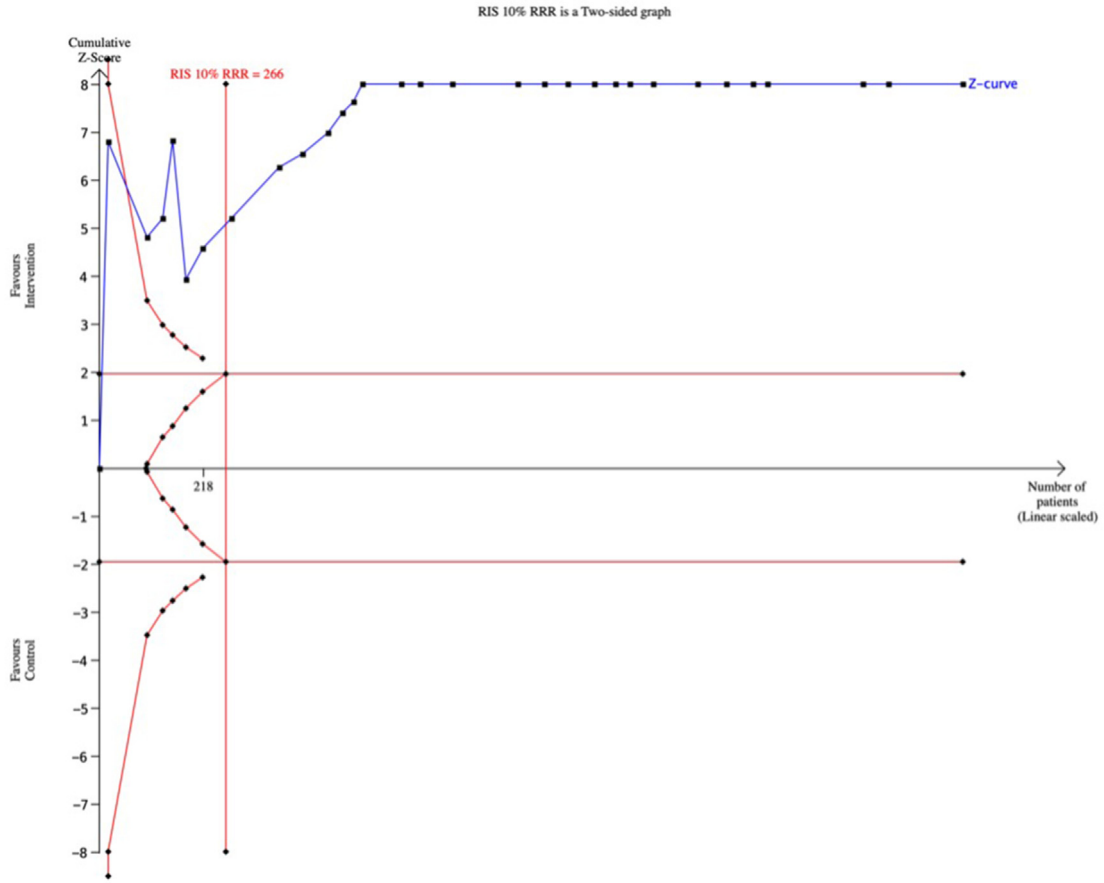

Fig. 6. Trial sequential analysis (TSA) plot.
Blue line (z-curve): cumulative Z score with each square adding the results of the individual studies; horizontal red lines: conventional threshold for significance (p value of 0.05); vertical red line: required information size (RIS); dotted small red lines: trial sequential monitoring boundaries.

from an OR. The Cochrane handbook of systematic reviews provides a formula for this calculation [39].

When working with continuous data, there are several effect estimates to choose from. Some of the most commonly used effect estimates are mean difference or standardized mean difference.

4.2.4. Mean difference (MD) and standardized mean difference (SMD)

The MD is simply calculated by subtracting the mean outcome value (e.g., hemoglobin value) in group A from the mean outcome value in group B. The magnitude of the effect will vary widely depending on the absolute mean values of the outcome. In order to increase interpretability, the SMD was introduced which divides the mean difference by the pooled standard deviation and thereby "standardizes" this effect estimate. This effect estimate (obtained by dividing the MD by the grouped standard deviation) is also called Cohen's d. The greater Cohen's d value, the greater the effect (and the greater the difference in outcome between group A and B) with values above 0.8 indicating a strong effect [40]. The log OR can be approximated from the SMD [41].

5. Trial sequential analysis (TSA)

TSA was introduced to control for type I and II errors of statistical analysis. Similar to an individual RCT a meta-analysis carries the risk of false-positive (type I error) or false negative results (type II error) [42]. Fig. 6 gives the example of a TSA, as published by Koning et al. [43]. The horizontal red lines give the conventional threshold for significance with a constant Z-value of 1.96 (p value of 0.05). The vertical red line gives the required information size (RIS) [44]. The blue line (Z-curve) is the cumulative Z-score and each square adds the results of the individual trials to the score. If the Z-score crosses the RIS line, then there is firm evidence of an effect of the intervention and the meta-analysis has sufficient power. In addition, trial sequential monitoring boundaries (based on the O'Brien-Fleming alpha-spending function) are constructed (dotted small red lines) [45]. Firm evidence can also be concluded when the Z-score crosses the monitoring boundaries but does not reach the RIS threshold. In Fig. 6, the Z-score crosses both, monitoring boundary and RIS line. In case of a negative result (i.e., no effect of the intervention was found) it is unclear whether there is a true absence of evidence or whether the sample size was too small to draw a firm conclusion. Addressing this issue TSA incorporates futility boundaries and a Z-score crossing the futility threshold allows to conclude a true negative result. A review about TSA is provided by Wetterslev et al. [46].

6. Individual patient data meta-analysis

Conventional meta-analysis aggregates summary effect estimates reported in the individual studies and they may be different, i.e. one study has reported risk ratio and the other has reported odds ratio. Moreover, when there is substantial heterogeneity in effect estimates, an average value may no longer be informative [47]. Similarly, the methods of statistical analysis may differ across the studies. These issues are addressed by individual patient data meta-analysis (IPD-MA), a new approach that seeks the raw data from the various individual studies. This approach not only allows to standardize the statistical analysis but also permits to do subgroup analyses that were not conducted in (all) individual studies [48]. Also, a detailed analysis of patients' characteristics will be made possible [49]. These benefits have to be balanced against possible disadvantages [49], including higher costs, a longer duration until

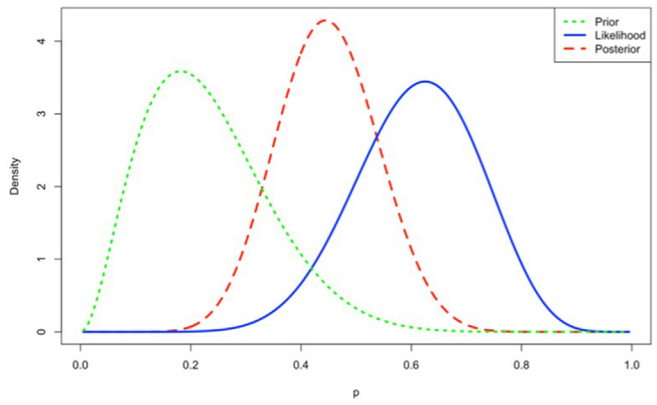

Fig. 7. Concept of prior distribution, likelihood, and posterior distribution of a binomial model in Bayesian statistics.
A beta-prior and binomial likelihood were assumed to calculate the conjugate beta posterior distribution. As an example, the posterior distribution might represent the distribution of a true response rate.

completion of the analyses as well as the need to receive data from all studies published on the research question. A pitfall in the conduct of an IPD-MA is that patient data may be considered as if they came from one single trial whereas clustering of patients within studies needs to be taken into consideration to avoid spurious conclusions [50]. In recent years recommendations for planning and conducting IPD-MAs have become available [47]. More detailed information about IPD-MA is given by Riley et al. [51].

7. Living systematic reviews and meta-analysis

Another emerging type of meta-analysis is that of living systematic reviews [52], which incorporates up-coming data in regular, pre-defined time-intervals. According to guidance by the Cochrane collaboration three pre-requisites have to be fulfilled when the conduct of a living MA is considered: the review question should be of high importance for clinical decision-making, the existing evidence is uncertain and insufficient to inform clinical practice, additional relevant information is likely to be produced and also likely to impact the conclusions. The interim guidance by the Cochrane collaboration has recently been revised and updated [53].

In the field of pain medicine, a living meta-analysis of plant-based treatments for chronic pain management has been commissioned by the Agency for Healthcare Research and Quality Evidence-based Practice Center Program [54], reflecting the increasing interest in this modern approach to produce up-to-date evidence.

8. Bayesian statistics

In statistics, there are two streams of approaching a statistical problem: Frequentist and Bayesian. The commonly used statistics (p-value, 95%CI etc.) are the frequentist way of making inference. Bayesian statistics are less commonly used but have some very interesting properties that make Bayesian statistics a desirable alternative to frequentist statistics. Clinicians might be most familiar with the Bayesian theorem that is used to calculate the positive or negative predictive value from sensitivity/specificity and prevalence. Bayesian statistics, compared to frequentist statistics, does not infer a single effect estimate (with 95% CI), but a distribution of an effect estimate. This is called the posterior distribution of a certain parameter (e.g., the mean difference in hemoglobin between two arms of a trial). It is called posterior because it is calculated from the likelihood (which corresponds to the data) and the prior distribution. The "prior" includes previous knowledge (e.g., results from a trial or meta-analysis) and the posterior distribution is therefore also dependent on the prior distribution (Fig. 7). This dependence on a pre-specified value is one of the reasons that Bayesian statistics are less commonly seen in medical literature. However, the influence of the specified prior distribution is usually checked with a sensitivity analysis by defining a non-informative prior (a prior which does not convey any previous information). When using a non- or weekly-informative prior, the posterior distribution is then mostly dependent on the likelihood (i.e., the data). A detailed description of Bayesian statistics is given by Held and Bove [55].

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] Murad MH, Asi N, Alsawas M, Alahdab F. New evidence pyramid. Evid Base Med 2016;21(4):125–7. https://doi.org/10.1136/ebmed-2016-110401.
[2] Manchikanti L, Benyamin RM, Helm S, Hirsch JA. Evidence-based medicine, systematic reviews, and guidelines in interventional pain management: part 3: systematic reviews and meta-analyses of randomized trials. Pain Physician 2009; 12(1):35–72.
[3] Berlin JA, Golub RM. Meta-analysis as evidence: building a better pyramid. JAMA 2014;312(6):603–5. https://doi.org/10.1001/jama.2014.8167.
[4] Gurevitch J, Koricheva J, Nakagawa S, Stewart G. Meta-analysis and the science of research synthesis. Nature 2018;555(7695):175–82. https://doi.org/10.1038/nature25753.
[5] Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, Shamseer L, Tetzlaff JM, Akl EA, Brennan SE, Chou R, Glanville J, Grimshaw JM, Hrobjartsson A, Lalu MM, Li T, Loder EW, Mayo-Wilson E, McDonald S, McGuinness LA, Stewart LA, Thomas J, Tricco AC, Welch VA, Whiting P, Moher D. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. J Clin Epidemiol 2021;134:178–89. https://doi.org/10.1016/j.jclinepi.2021.03.001.
[6] Brooke BS, Schwartz TA, Pawlik TM. MOOSE reporting guidelines for meta-analyses of observational studies. JAMA Surg 2021;156(8):787–8. https://doi.org/10.1001/jamasurg.2021.0522.
[7] Higgins J, Thomas J, Chandler J, Cumpston M, Li T, Page M, Welch V. Cochrane handbook for systematic reviews of interventions. 2022. version 6.3 (updated February 2022), https://training.cochrane.org/handbook/current. [Accessed 11 May 2022].
[8] Klimek M, Rossaint R, van de Velde M, Heesen M. Combined spinal-epidural vs. spinal anaesthesia for caesarean section: meta-analysis and trial-sequential analysis. Anaesthesia 2018;73(7):875–88. https://doi.org/10.1111/anae.14210.
[9] Rijs K, Mercier FJ, Lucas DN, Rossaint R, Klimek M, Heesen M. Fluid loading therapy to prevent spinal hypotension in women undergoing elective caesarean section: network meta-analysis, trial sequential analysis and meta-regression. Eur J Anaesthesiol 2020;37(12):1126–42. https://doi.org/10.1097/EJA.0000000000001371.
[10] Stern C, Jordan Z, McArthur A. Developing the review question and inclusion criteria. Am J Nurs 2014;114(4):53–6. https://doi.org/10.1097/01.NAJ.0000445689.67800.86.
[11] Munn Z, Stern C, Aromataris E, Lockwood C, Jordan Z. What kind of systematic review should I conduct? A proposed typology and guidance for systematic reviewers in the medical and health sciences. BMC Med Res Methodol 2018;18(1):5. https://doi.org/10.1186/s12874-017-0468-4.
[12] Bramer WM, Rethlefsen ML, Kleijnen J, Franco OH. Optimal database combinations for literature searches in systematic reviews: a prospective exploratory study. Syst Rev 2017;6(1):245. https://doi.org/10.1186/s13643-017-0644-y.
[13] Clephas PRD, Hoeks SE, Trivella M, Guay CS, Singh PM, Klimek M, Heesen M. Prognostic factors for chronic post-surgical pain after lung or pleural surgery: a protocol for a systematic review and meta-analysis. BMJ Open 2021;11(6):e051554. https://doi.org/10.1136/bmjopen-2021-051554.
[14] Rethlefsen ML, Farrell AM, Osterhaus Trzasko LC, Brigham TJ. Librarian co-authors correlated with higher quality reported search strategies in general internal medicine systematic reviews. J Clin Epidemiol 2015;68(6):617–26. https://doi.org/10.1016/j.jclinepi.2014.11.025.
[15] Bramer WM, Rethlefsen ML, Mast F, Kleijnen J. Evaluation of a new method for librarian-mediated literature searches for systematic reviews. Res Synth Methods 2018;9(4):510–20. https://doi.org/10.1002/jrsm.1279.
[16] Bramer WM, de Jonge GB, Rethlefsen ML, Mast F, Kleijnen J. A systematic approach to searching: an efficient and complete method to develop literature searches. J Med Libr Assoc 2018;106(4):531–41. https://doi.org/10.5195/jmla.2018.283.
[17] Jackson JL, Kuriyama A, Anton A, Choi A, Fournier JP, Geier AK, Jacquerioz F, Kogan D, Scholcoff C, Sun R. The accuracy of Google translate for abstracting data from non-English-language trials for systematic reviews. Ann Intern Med 2019;171(9):677–9. https://doi.org/10.7326/M19-0891.
[18] Bramer WM, Milic J, Mast F. Reviewing retrieved references for inclusion in systematic reviews using EndNote. J Med Libr Assoc 2017;105(1):84–7. https://doi.org/10.5195/jmla.2017.111.

[19] Aronson J, Badenoch D, Banerjee A, Bankhead C, Brassey J, Chalmers I, Davis R, Friedemann-Smith C, Heneghan C, Lach J, Mahtani K, McCall M, McFadden E, Nunan D, O'Sullivan J, Onakpoya I, Pluddemann A, Richards G, Spencer E, Turk A. Catalogue of bias. 2019. https://catalogofbias.org. [Accessed 7 May 2022].
[20] Higgins J, Sterne J, Savovic J, Page M, Hróbjartsson A, Boutron I, Reeves B, Eldridge S. A revised tool for assessing risk of bias in randomized trials. 2016. https://methods.cochrane.org/bias/resources/rob-2-revised-cochrane-risk-bias-tool-randomized-trials. [Accessed 7 May 2022].
[21] Wells G, Shea B, O'Connell D, Peterson J, Welch V, Losos M, Tugwell P. The Newcastle-Ottawa Scale (NOS) for assessing the quality of nonrandomised studies in meta-analyses. 2011. http://www.ohri.ca/programs/clinical_epidemiology/oxford.asp. [Accessed 1 May 2022].
[22] Sterne JA, Hernan MA, Reeves BC, Savovic J, Berkman ND, Viswanathan M, Henry D, Altman DG, Ansari MT, Boutron I, Carpenter JR, Chan AW, Churchill R, Deeks JJ, Hrobjartsson A, Kirkham J, Juni P, Loke YK, Pigott TD, Ramsay CR, Regidor D, Rothstein HR, Sandhu L, Santaguida PL, Schunemann HJ, Shea B, Shrier I, Tugwell P, Turner L, Valentine JC, Waddington H, Waters E, Wells GA, Whiting PF, Higgins JP. ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. BMJ 2016;355:i4919. https://doi.org/10.1136/bmj.i4919.
[23] Whiting PF, Rutjes AW, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, Leeflang MM, Sterne JA, Bossuyt PM, Group Q-. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med 2011;155(8):529–36. https://doi.org/10.7326/0003-4819-155-8-201110180-00009.
[24] Hayden JA, van der Windt DA, Cartwright JL, Cote P, Bombardier C. Assessing bias in studies of prognostic factors. Ann Intern Med 2013;158(4):280–6. https://doi.org/10.7326/0003-4819-158-4-201302190-00009.
[25] Wolff RF, Moons KGM, Riley RD, Whiting PF, Westwood M, Collins GS, Reitsma JB, Kleijnen J, Mallett S, Groupdagger P. PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. Ann Intern Med 2019;170(1):51–8. https://doi.org/10.7326/M18-1376.
[26] Barili F, Parolari A, Kappetein PA, Freemantle N. Statistical Primer: heterogeneity, random- or fixed-effects model analyses? Interact Cardiovasc Thorac Surg 2018;27(3):317–21. https://doi.org/10.1093/icvts/ivy163.
[27] Viechtbauer W. Conducting meta-analyses in R with the metafor package. J Stat Software 2010;36(3):1–48. https://doi.org/10.18637/jss.v036.i03.
[28] Harris RJ, Deeks JJ, Altman DG, Bradburn MJ, Harbord RM, Sterne JAC. Metan: fixed- and random-effects meta-analysis. STATA J 2008;8(1):3–28. https://doi.org/10.1177/1536867X0800800102.
[29] Higgins JP, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. BMJ 2003;327(7414):557–60. https://doi.org/10.1136/bmj.327.7414.557.
[30] Thompson SG, Higgins JP. How should meta-regression analyses be undertaken and interpreted? Stat Med 2002;21(11):1559–73. https://doi.org/10.1002/sim.1187.
[31] Schünemann H, Brożek J, Guyatt G, Oxman A. GRADE handbook for grading quality of evidence and strength of recommendations. 2013. https://gdt.gradepro.org/app/handbook/handbook.html. [Accessed 7 May 2022].
[32] Faltinsen EG, Storebo OJ, Jakobsen JC, Boesen K, Lange T, Gluud C. Network meta-analysis: the highest level of medical evidence? BMJ Evid Based Med 2018;23(2):56–9. https://doi.org/10.1136/bmjebm-2017-110887.
[33] Held L, Ott M. On p-values and Bayes factors. Annual Review of Statistics and Its Application 2018;5(1):393–419. https://doi.org/10.1146/annurev-statistics-031017-100307.
[34] Harrington D, D'Agostino Rb Sr, Gatsonis C, Hogan JW, Hunter DJ, Normand ST, Drazen JM, Hamel MB. New guidelines for statistical reporting in the journal. N Engl J Med 2019;381(3):285–6. https://doi.org/10.1056/NEJMe1906559.
[35] Wasserstein RL, Schirm AL, Lazar NA. Moving to a world beyond "p < 0.05". Am Statistician 2019;73(sup1):1–19. https://doi.org/10.1080/00031305.2019.1583913.
[36] Higgins J, Thomas J, Chandler J, Cumpston M, Li T, Page M, Welch V. Cochrane handbook for systematic reviews of interventions. 2022. version 6.3 (updated February 2022). [Chapter 10].4.3, https://training.cochrane.org/handbook/current. [Accessed 11 May 2022].
[37] Higgins J, Thomas J, Chandler J, Cumpston M, Li T, Page M, Welch V. Cochrane handbook for systematic reviews of interventions. 2022. version 6.3 (updated February 2022). [Chapter 6].4.1.1, https://training.cochrane.org/handbook/current. [Accessed 11 May 2022].
[38] Higgins J, Thomas J, Chandler J, Cumpston M, Li T, Page M, Welch V. Cochrane handbook for systematic reviews of interventions. 2022. version 6.3 (updated February 2022). [Chapter 15].4.4.2, https://training.cochrane.org/handbook/current. [Accessed 11 May 2022].
[39] Higgins J, Thomas J, Chandler J, Cumpston M, Li T, Page M, Welch V. Cochrane handbook for systematic reviews of interventions. 2022. version 6.3 (updated February 2022). [Chapter 15].4.4.4, https://training.cochrane.org/handbook/current. [Accessed 11 May 2022].
[40] Cohen J. Statistical power analysis for the behavioral sciences. 2th edition. Hillsdale: Lawrence Erlbaum Associates; 1988.
[41] Higgins J, Thomas J, Chandler J, Cumpston M, Li T, Page M, Welch V. Cochrane handbook for systematic reviews of interventions. 2022. version 6.3 (updated February 2022). [Chapter 15].5.3.3, https://training.cochrane.org/handbook/current. [Accessed 11 May 2022].
[42] Brok J, Thorlund K, Wetterslev J, Gluud C. Apparently conclusive meta-analyses may be inconclusive–Trial sequential analysis adjustment of random error risk due to repetitive testing of accumulating data in apparently conclusive neonatal meta-analyses. Int J Epidemiol 2009;38(1):287–98. https://doi.org/10.1093/ije/dyn188.
[43] Koning MV, Klimek M, Rijs K, Stolker RJ, Heesen MA. Intrathecal hydrophilic opioids for abdominal surgery: a meta-analysis, meta-regression, and trial sequential analysis. Br J Anaesth 2020;125(3):358–72. https://doi.org/10.1016/j.bja.2020.05.061.
[44] Wetterslev J, Thorlund K, Brok J, Gluud C. Estimating required information size by quantifying diversity in random-effects model meta-analyses. BMC Med Res Methodol 2009;9:86. https://doi.org/10.1186/1471-2288-9-86.
[45] Thorlund K, Engstrøm J, Wetterslev J, Brok J, Imberger G, Gluud C. In: User manual for trial sequential analysis (TSA). 2th edition. Copenhagen: Copenhagen Trial Unit; 2017.
[46] Wetterslev J, Thorlund K, Brok J, Gluud C. Trial sequential analysis may establish when firm evidence is reached in cumulative meta-analysis. J Clin Epidemiol 2008;61(1):64–75. https://doi.org/10.1016/j.jclinepi.2007.03.013.
[47] Debray TP, Moons KG, van Valkenhoef G, Efthimiou O, Hummel N, Groenwold RH, Reitsma JB. Get Real Methods Review G. Get real in individual participant data (IPD) meta-analysis: a review of the methodology. Res Synth Methods 2015;6(4):293–309. https://doi.org/10.1002/jrsm.1160.
[48] Cooper H, Patall EA. The relative benefits of meta-analysis conducted with individual participant data versus aggregated data. Psychol Methods 2009;14(2):165–76. https://doi.org/10.1037/a0015565.
[49] Stewart LA, Tierney JF. To IPD or not to IPD? Advantages and disadvantages of systematic reviews using individual patient data. Eval Health Prof 2002;25(1):76–97. https://doi.org/10.1177/0163278702025001006.
[50] Abo-Zaid G, Guo B, Deeks JJ, Debray TP, Steyerberg EW, Moons KG, Riley RD. Individual participant data meta-analyses should not ignore clustering. J Clin Epidemiol 2013;66(8):865–73. https://doi.org/10.1016/j.jclinepi.2012.12.017. e4.
[51] Riley RD, Lambert PC, Abo-Zaid G. Meta-analysis of individual participant data: rationale, conduct, and reporting. BMJ 2010;340:c221. https://doi.org/10.1136/bmj.c221.
[52] Elliott JH, Turner T, Clavisi O, Thomas J, Higgins JP, Mavergames C, Gruen RL. Living systematic reviews: an emerging opportunity to narrow the evidence-practice gap. PLoS Med 2014;11(2):e1001603. https://doi.org/10.1371/journal.pmed.1001603.
[53] Guidance for the production and publication of Cochrane living systematic reviews: Cochrane Reviews in living mode. 2019. version December 2019, https://community.cochrane.org/sites/default/files/uploads/inline-files/Transform/201912_LSR_Revised_Guidance.pdf. [Accessed 26 April 2022].
[54] McDonagh MS, Chou R, Wagner J, Ahmed AY, Morasco BJ, Iyer S, Kansagara D. Living systematic reviews: practical considerations for the agency for Healthcare research and quality evidence-based practice center program. Rockville (MD): Agency for Healthcare Research and Quality (US); 2022.
[55] Held L, Bové D. In: Likelihood and bayesian inference: with applications in biology and medicine (statistics for biology and health). 2th edition. Zürich: Springer; 2020.