Systematic Reviews and Meta-Analysis: A Guide for Beginners

JOSEPH L MATHEW

From Department of Pediatrics, Advanced Pediatrics Centre, PGIMER, Chandigarh.

Correspondence to: *Prof Joseph L Mathew, Department of Pediatrics, Advanced Pediatrics Centre, PGIMER Chandigarh, India. joseph.l.mathew@gmail.com*

PII: S097475591600350

Note: *This early-online version of the article is an unedited manuscript that has been accepted for publication. It has been posted to the website for making it available to readers, ahead of its publication in print. This version will undergo copy-editing, typesetting, and proofreading, before final publication; and the text may undergo minor changes in the final version*

ABSTRACT

Systematic reviews involve the application of scientific methods to reduce bias in review of literature. The key components of a systematic review are a well-defined research question, comprehensive literature search to identify all studies that potentially address the question, systematic assembly of the studies that answer the question, critical appraisal of the methodological quality of the included studies, data extraction and analysis (with and without statistics), and considerations towards applicability of the evidence generated in a systematic review. These key features can be remembered as six 'A'; Ask, Access, Assimilate, Appraise, Analyze and Apply. Meta-analysis is a statistical tool that provides pooled estimates of effect from the data extracted from individual studies in the systematic review. The graphical output of meta-analysis is a forest plot which provides information on individual studies and the pooled effect. Systematic reviews of literature can be undertaken for all types of questions, and all types of study designs. This article highlights the key features of systematic reviews, and is designed to help readers understand and interpret them. It can also help to serve as a beginner's guide for both users and producers of systematic reviews and to appreciate some of its methodological issues.

Evidence-based (or evidence-informed) healthcare requires the integration of high-quality research evidence, clinical expertise and patient (consumer) values [1]. However, the immense volume of primary research and its diversity in terms of methodology, necessitate that it be reviewed and synthesized to make rational interpretations and decisions. This necessity has led to an entire field of secondary research to synthesize data from primary research. Systematic reviews are the key pillar of such secondary research. The broad principle of systematic review is to apply "scientific strategies that limit bias to the systematic assembly, critical appraisal, and synthesis of all relevant research studies on a specific topic" [2]. Thus, in contrast to traditional narrative reviews, there is a rigorous attempt to limit bias in the process of selecting, reviewing and synthesizing primary research studies in SR. These efforts at minimizing bias have led systematic review to be regarded superior to primary research study designs, thereby finding a place at the top of the hierarchy of research evidence. In terms of research methodology, bias can be described as systematic error that leads away from the truth [3]. This is largely avoidable, in contrast to random error which occurs by chance [3], hence is unpredictable. The ultimate goal of systematic review is to facilitate healthcare decisions that are objective, reproducible and transparent.

Meta-analysis is a statistical tool that is used to mathematically pool data derived from a SR, and generate a summary conclusion [4]. Meta-analysis of data is inappropriate if not derived from a systematic review. It would be akin to applying statistical tests on data which are not derived from primary research studies.

This article highlights the key features and methodological issues of systematic reviews and is designed to help readers understand and interpret them. This article is not intended to be a comprehensive handbook to interpret or conduct systematic reviews but can serve as a beginner's guide for both users and producers of systematic review.

Systematic reviews are initiated after preparing, registering, and publishing a review protocol. The process is similar to preparing protocols for primary research studies. Registration of systematic review protocols is broadly similar to registration of clinical trial protocols, however different platforms are used. One such platform is PROSPERO which serves as a database for registering protocols of systematic review [5]. This promotes transparency in the review process.

High quality reviews such as Cochrane reviews, publish systematic review protocols after stringent peer review. Some journals also publish systematic review protocols, whereas others expect them to be available online for access by anyone. Currently, it is difficult to publish a good quality systematic review without prior registration and publication (or disclosure) of the protocol. This is to ensure that appropriate methodology is used, detailed methods are disclosed beforehand (a priori), and no modifications are made after data become available (post hoc). This makes the review process and the product, systematic, objective, reproducible, and transparent (summarized by the acronym SORT).

MAKING SENSE OF A SYSTEMATIC REVIEW

Healthcare professionals reading, appraising or conducting a systematic review should focus on six key aspects (**Table I**).

Ask (Research Question)

The science of evidence-based medicine hinges on the art of framing and addressing research questions [6]. This is the most important step in any research study, including systematic review. The 'PICO format' [7] of research questions is better expanded to 'PICOTS' as follows.

- *P* (*Population and/or Patient and/or Problem*): It refers to the people in/for whom the systematic review is expected to be applied.
- *I* (*Intervention*): In the context of systematic reviews examining effects of treatment, 'I' encompasses medicines, procedures, health education, public health measures, or bundles/combinations of these. 'I' also includes preventive measures such as vaccination, prophylaxis, health education tools, and packages of such interventions. In some contexts, the intervention is not administered by the study investigators, but by nature, and the investigators are merely observing the effects. Therefore, 'I' can be better expressed as 'Exposure' abbreviated as 'E'. This is also true for systematic review of diagnostic test studies (wherein participants are 'exposed to' diagnostic tests), prognostic markers (wherein participants are exposed to one or more factors), and prevalence of certain conditions (wherein participants are naturally exposed to the condition).

- *C* (*Comparison*): Those not receiving the intervention could receive an alternate intervention, or placebo, or nothing (depending on the research question). However, for some study designs and/or research questions, it may not be feasible to include a Comparison.
- *O* (*Outcome*): This refers to the broad parameters by which the effect of 'I' on 'P' in comparison to 'C' can be measured. In general, systematic review of interventions focus on efficacy, safety, and sometimes cost. SRs of diagnostic tests focus on measures of accuracy, reliability, and cost. Multiple specific outcome measures can be analyzed for each outcome being evaluated.
- *T* (*Time-frame*): Outcomes are meaningful only when the time-frame in which they are recorded are specified. For example, 'mortality' as an outcome can be recorded in various time-frames or duration. Different outcomes in a systematic review may have different time-frames which should be specified clearly.
- *S* (*Study design*): Multiple study designs may be used in primary studies to address the same research question. However, study designs have inherent risks of bias (by virtue of the design itself) which results in a hierarchy of primary research study designs. Randomized controlled trials (RCT) are associated with the least risk of bias for evaluating interventions. Bias increases in non-randomized trials, other clinical trials, cohort studies (with and without comparison groups), case-control studies, case series, and case reports (in that order). Since the focus of systematic review is to review literature minimizing bias as far as possible, some systematic review include only methodologically high-quality study designs (such as RCT), whereas others may include various study designs and examine the impact of lower-quality designs separately.

There are other formats (besides PICOTS) for framing and/or presenting research questions. The SPICE acronym covers issues such as setting, population, intervention, comparison and evaluation [8]. It is generally considered helpful to develop questions relating to qualitative research, and for evaluating project proposals and quality improvement. Another tool is SPIDER, which helps to structure qualitative research questions. It summarizes sample, phenomenon of interest, design, evaluation and research type [9]. Yet another format is ECLIPSE [10], that is reportedly helpful for questions addressing healthcare policies or services. The acronym covers expectation, client, location, impact, professionals, and service.

However, the PICO format remains the most popular version perhaps because it is the oldest, covers a variety of research questions, is 'portable' across study designs, and can be extended to secondary research, health technology assessment, guidelines, and policy issues.

The research question in a systematic review is usually clearly specified in the introduction section. Often, no research question may be found but enough information may be provided for readers to frame one in the PICOTS format. However, systematic review that do not specify a research question, or facilitate the construction of one by readers, are likely to result in biased interpretations and should be read with caution. Research questions that have very narrow or highly focused 'P' run the risk of producing systematic review with limited generalizability. On the other hand, very broad questions can generate more noise than signal. The key is to have a research question wherein the elements are balanced to include the population of interest in a non-restrictive manner, yet have a high signal to noise ratio. The PICOTS template is applicable for systematic reviews addressing all types of research questions (**Table II**).

ACCESS (LITERATURE SEARCH)

This step is designed to identify all literature that can potentially answer the research question. It includes several components to facilitate systematic, objective, reproducible, and transparent (SORT) search and inclusion of studies.

Types of studies: Systematic review authors may include only studies conforming to the most appropriate study design, or choose to include various types of study designs. The advantage of the first approach is that studies with higher risk of bias are eliminated upfront, however the disadvantage is that there may be insufficient studies of high methodological quality, and these may not truly represent the real-world scenario. The second approach may yield more studies (hence larger sample size) but reduce the confidence in the overall result due to inclusion of lower quality primary studies. The way out is for systematic review authors to either include only the highest quality study design, or include multiple designs but perform separate analyses of high quality versus lower quality designs, and explore the difference.

Types of participants: This refers to the participant characteristics in the primary studies, such as age group, socio-demographic characteristics, duration of disease, and severity. Here also, choosing a very narrow set of criteria limits the generalizability of the systematic review. whereas very broad criteria may end up combining apples and oranges to obtain a pooled result. A useful method is to ensure that the inclusion criteria are broad, but include objective methods of diagnosis and measurement of disease severity. For example, in diagnostic test studies, the participants should include people 'suspected to have the disease' or those 'with potential to have the disease', and not only those confirmed to have the disease.

Types of intervention/exposure: The PICOTS question in the Introduction section identifies the broad contours of the intervention/exposure, whereas the methods section provides greater detail of the intervention such as, dosage, frequency of administration, mode of administration, duration of administration, and similar issues. When the intervention is a procedure, the skill/training of the operator and the healthcare setting may be additional factors. For studies measuring behavior change (in response to health education, legislation etc.), the 'intervention' may consist of a 'bundle' involving many different components, with or without reinforcement.

The intervention is actually an 'exposure' in diagnostic test studies, prognosis studies, and prevalence/incidence studies.

Types of comparison: All the details specified for the intervention should be specified for the comparison also. In intervention studies, the comparator may be another intervention (such as the current standard of care), placebo (if that is deemed safe and appropriate on ethical grounds), or no intervention (if safe/appropriate). In diagnostic test studies, there is no separate group of individuals for comparison, but the same group of participants receives the index test (exposure) and the reference test (comparison). Some primary research studies may not have comparison group (examples are clinical trials without a comparison group, cohort studies without comparison, and prevalence/incidence studies). The information derived from such studies is inferior to those with comparison groups.

Types of outcome measures: Just as in primary research studies, SYSTEMATIC REVIEWgenerally have one primary outcome and multiple secondary outcomes. Each outcome may have several methods of measurement/recording. Thus the broad term 'efficacy' may include outcomes like clinical cure, resolution, survival/mortality, need for escalation of therapy, duration of hospitalization, or quality of life measurements. Other surrogate outcomes of efficacy could be laboratory parameters, biomarker levels, radiological findings, or results of combinations of investigations. Each of these outcomes could be measured in multiple ways, and may be recorded at multiple time points, and/or using multiple instruments/tools, all of which are generally reported in the systematic review. Similarly, safety outcomes could include development of adverse events, count of serious adverse events, number of patients developing such events, number of events per patient, need for enhanced monitoring, etc. It is impossible to include every possible outcome measure in a systematic review. However, no important outcomes should be missed; patient-centric outcomes should be included; outcomes measured objectively are preferred; hard outcomes are considered superior to soft outcomes, and purely indirect/surrogate outcomes are less preferred. The methods section should include the time-frame of recording each of the included outcomes. Where the outcomes are recorded multiple times, separate analyses would be necessary for each.

Search methods for identification of studies

Where? This section defines the literature databases accessed to identify all the relevant evidence. High quality systematic reviews search multiple electronic databases such as Medline, Embase, Cochrane Register of Trials, and other repositories. At the very least, two databases should be searched. Depending on the review question, additional literature databases may also be searched. In addition, most reviewers search other sources of literature including reference lists of included studies (this is referred to as hand-searching), clinical trials registries (for registered trials), conference abstract books/proceedings, and databases of non-indexed journals. In the Indian context, many journals are indexed in IndMED [11], although not in Medline. Similarly, Wangfang Data is a source of Chinese literature [12], and LILACS database includes Latin American and Caribbean literature [13]. There are also specific databases for different types of clinical problems and/or healthcare specialists. All these additional searches are focused on published sources of evidence. Some authors go further and search sources of unpublished literature (sometimes referred to as grey literature). These may be available through repositories of such studies (for example OpenGrey database includes over 7 lakh references of grey literature in Europe) [14].

How? Databases of published and unpublished literature have specific approaches to ensure comprehensive searches for all eligible primary studies. Systematic reviews thus undertake multiple searches of each database, with various combinations of keywords, exploiting the inbuilt filters in some of the databases. Although it may be convenient to search only English language publications, high-quality reviews do not restrict by language or any other criteria. This is so that no bias creeps in through selective inclusion (or exclusion) of primary studies. Such rigour increases the cost, duration, and workload of systematic review authors, but minimizes a major source of bias.

When? Systematic review authors are expected to declare the date of literature search, period over which each database was searched, and also provide updated searches just before the systematic review is published. All these efforts ensure that the evidence is current and the searches are reproducible.

Who? Literature searching is a key step of SRs, and is generally conducted independently by more than one author. The outputs, eligibility, and selection are compared and is resolved by another independent author where there is mismatch. Although not essential, reference managers such as Endnote, Zotero, or Mendeley can be used to compile the search output, remove duplicate publications and obtain the final list of the preliminary search.

ASSIMILATE (INCLUSION AND EXCLUSION OF STUDIES)

Generally, a three-step approach is used to confirm the eligibility of primary studies for inclusion in the SR. This includes a preliminary screening of each study title, followed by screening the abstract of short-listed titles. The third step is to read the full-text of the short-listed abstracts. Thereafter, the full-text of each publication is matched against the set of eligibility criteria described above, to decide on inclusion into the systematic review (or otherwise). Here too, the PICOTS framework is very helpful. Each step is carefully recorded and reasons for exclusion are documented for the studies excluded in the third step. This is done to ensure transparency and objectivity in study selection. It is good practice to ensure that screening of titles, abstracts, and full text for potential inclusion, is done by more than one reviewer, working independently.

It is also helpful to prepare a flow diagram showing the results of the literature searches, exclusion of publications with reasons, and the pathway to final inclusion of eligible studies. This is similar to the flow-diagram of participant recruitment in trials.

APPRAISE (CRITICAL APPRAISAL OF INCLUDED STUDIES)

All SRs undertake critical appraisal of included studies for methodological quality. This refers to assessment of efforts made by investigators of primary studies to minimize bias during the conduct of their study. Bias or systematic error can creep into primary research studies with inappropriate study designs, and inappropriate study methods. The former includes choosing study designs that inherently have high(er) risk of bias, and insufficient precautions to address the common sources of bias within each study design. For example, in studies examining interventions, RCT is the ideal study design, and within RCT, sources of bias include selection bias, allocation bias, performance bias, and outcome reporting bias. Inappropriate study methods include using inappropriate tools for measuring outcomes, lack of calibration of instruments used to record outcomes, inappropriate recording methods, inappropriate/insufficient follow-up, etc.

Appraisal in SRs is generally restricted to examination of study design issues and efforts to minimize bias due to this. There are standard online tools available for each type of study design. The Cochrane Risk of Bias tool [15] is considered a standard tool for RCT and includes appraising the methods used (and adequacy thereof) for key design elements in intervention trials viz. random sequence generation, concealment of allocation, blinding of study participants, blinding of outcome assessors, incomplete outcome reporting, and selective outcome reporting. There is an additional element for appraising any other bias. Software tools for SRs, such as the Cochrane Review Manager or RevMan [16] have options for the pictorial representation of quality appraisal of included studies.

The Newcastle Ottawa Scale (NOS) is often used to assess the quality of non-randomized studies including case-control, cohort studies, and even qualitative studies [17]. The NOS contains eight items, categorized into three broad perspectives: selection of the study groups; comparability of the groups; and ascertainment of either the exposure or outcome of interest (for case-control or cohort studies, respectively). For each item, a star system is used to allow a semi-quantitative assessment of study quality. High-quality studies are defined by a score 6 or more of 9 total points [18].

Another popular tool for non-RCT studies is the Risk of Bias in Non-Randomized Studies – of Interventions tool abbreviated as ROBINS-I [19]. It includes assessments of bias in pre-intervention (biases due to confounding as well as participant selection), at intervention (bias in classification of interventions), and post-intervention (biases due to deviations from the intended interventions, missing data, measurement of outcomes, and selective reporting).

The QUADAS-2 tool [20] can be used to evaluate the risk of bias of diagnostic test accuracy studies. It examines the risk of bias in four broad domains viz. patient selection, index test, reference standard, and flow and timing. Among these, the first three are also evaluated in terms of applicability.

There are specific tools for assessing quality of environmental health studies. These include tools developed by the Office of Health Assessment and Translation (OHAT) and Integrated Risk Information System (IRIS) [21]. There are also additional tools specific for animal studies. For example, SYCRLE's tool is an adaptation of the Cochrane Risk of Bias tool, and is used to assess internal validity, addressing selection, performance, detection, attrition and reporting biases [22].

ANALYZE (DATA EXTRACTION AND ANALYSIS)

Systematic reviewers prepare data extraction forms (that are not published, although Cochrane reviews present these details) which include the following information from each included study: (*i*)

Identification characteristics (authors, source, year); (*ii*) Study characteristics (enrolment criteria, sample size, PICOTS information), (*iii*) Appraisal for bias (using standard tools/checklists), (*iv*) Data reflecting the outcomes specified in PICOTS, and (*v*) Additional notes, if any.

Data to be analyzed could include descriptive data and quantitative data. Narrative synthesis of the extracted data is helpful to understand the perspectives of the primary studies in terms of the PICO elements. A table highlighting the descriptive characteristics of the included studies is very helpful for readers. Quantitative data are extracted for each outcome measure (specified in the review protocol). Data extraction is also generally done independently by more than one reviewer, with provision to resolve discrepancies. Sometimes, published versions of individual studies lack pieces of data that are important for the review. In such situations, the systematic review authors correspond with study authors to obtain missing data (and record the process).

In intervention reviews, numerical data of outcome measures (from included studies) usually conform to either dichotomous data (expressed as proportions) or continuous data (expressed as mean with standard deviations, or variations of this). Other forms of presentation include median (with interquartile ranges). In diagnostic test reviews, each included study provides information on the number of true positive, false positive, true negative, and false negative test results.

The extracted data may be considered for pooled analysis if there is sufficient data (although there is no strict definition for this), and the data are in a format conducive for pooling. For example, data from a study presenting an outcome as mean (standard deviation) is not amenable for pooling with data from another study presenting the same outcome as median (IQR), unless mathematical conversion techniques are applied to convert medians to means. Likewise, in studies reporting diagnostic tests, if only data on sensitivity and specificity are reported without the numbers from which they are derived, it is difficult to pool them. Such problems can be resolved if systematic review authors have access to the raw data from primary studies, and/or are able to undertake individual patient meta-analysis [23].

Meta-Analysis

The statistical procedure for pooling data from individual studies is called meta-analysis. Meta-analysis presents the estimate of effect from each included study, relative weight of each study in the pool, and the pooled estimate of effect. The relative weight depends on the variance in the result which, is impacted by the sample size and width of the confidence interval of the effect. In general, studies with less variance (i.e., narrower confidence interval of the effect) have greater relative weight, and studies with large sample sizes and narrow interval have the greatest weight. Understanding the concept of study weights is important because the pooled estimate of effect is not a mathematical average of the data from individual studies, but a weighted average.

The graphical output of meta-analysis is referred to as a forest plot. Although they may seem intimidating, a step-wise approach as shown in **Fig. 1** makes it easier to understand and interpret forest plots. Fig. 1 presents a meta-analysis (from a fictitious systematic review) of six hypothetical RCTs comparing Option A vs Option B for a clinical condition.

Step 1: What is the comparison? This is presented at the top of the forest plot and shows the interventions being compared as well as the outcome.

Step 2: What outcome measure is being compared? Each outcome can be represented by several measures. Each outcome measure is analyzed in a separate forest plot.

Step 3: How is the data presented? Dichotomous data are compared using odds ratio (OR), risk ratio or relative risk (RR), or risk difference (RD). All are valid measures. OR are mathematically purer, but RR are easier to understand. RD can be used to calculate the number needed to treat (NNT). Continuous data are presented as mean difference (MD), or weighted mean difference (WMD), or standardized mean difference (SMD). All measures are presented with confidence intervals (usually 95%, but modifiable).

Step 4: Which statistical model is used? There are two statistical models viz. fixed effect (FE) and random effects (RE). The FE model assumes that there is a single common estimate of effect, and all studies aim to estimate that common effect. In contrast, the RE model assumes that there is no single common effect, but a distribution of true effects, which varies from study to study [24]. This model considers heterogeneity among studies in terms of participants, biological characteristics, disease characteristics, measurement tools, etc. Thus, in the FE model, it is assumed that studies don’t estimate the true effect because of random error, whereas in the RE model, both random error and heterogeneity affect the pooled estimate of effect. **Web Fig. 1** presents the differences between FE and RE models of analysis, using the forest plot presented in **Fig. 1**.

Step 5: Examine individual studies. The forest plot shows the outcome data for each study, its effect (with confidence interval), relative weight in the pooled analysis, and a pictorial presentation of this data (which is usually a square whose position represents the effect, size represents the weight, and a horizontal line through the square represents the confidence interval).

Step 6: Examine pooled effect. The pooled effect is presented numerically as well as graphically. It represents a weighted average estimate of effect. The pictorial representation is with a diamond whose center corresponds to the pooled effect, and width represents the confidence interval.

A vertical line in the center of the forest plot represents the line of no effect. In the case of RR and OR, this corresponds to 1.0 and implies that the risk ratio (or odds ratio) is 1.0, confirming the absence of a difference between the groups. For mean differences, the line of no effect corresponds to zero, confirming that there is no difference between the groups. Therefore, it is obvious that confidence intervals whose bounds (limits) are on the same side of the line of no effect, suggest a statistically significant result, whereas confidence intervals crossing the line of no effect represent estimates that could lie on either side. No further tests of statistical significance are required, however some forest plots present additional tests for this. Similarly, narrower confidence intervals suggest more precise estimates, and vice versa.

Step 7: Examine and explore heterogeneity. Heterogeneity among studies refers to variation in the effect, which could be due to random chance or other factors. Random chance would be the only explanation for differences in estimates of effects if all studies were conducted in exactly the same way. In reality, studies are conducted somewhat differently, hence differences in effect result from random chance plus additional factors. This heterogeneity can be apparent by visual inspection of the pooled data wherein confidence intervals that fail to overlap suggest (but not confirm) the presence (but not the degree) of heterogeneity.

Currently, the Cochran statistic or more recently, the *I* square test (I^2) is used to mathematically calculate the degree of heterogeneity [25]. Currently, I^2 <50% is accepted as low degree of heterogeneity, I^2 between 50-75% as moderate degree, and I^2 >75% as high degree of heterogeneity. A *P* value of <0.10 suggests a statistically significant degree of heterogeneity, which should be explored to identify possible reasons. The RE model is generally preferred when there is significant heterogeneity among studies, for the reasons cited previously.

It may also be worth considering sub-group analysis when significant heterogeneity is evident. Here, studies sharing common characteristics are grouped together and pooled estimates of each sub-group are presented along with the overall estimate. **Web Fig. 2** presents an example wherein the studies presented in **Web Fig. 1** have been split into two sub-groups based on underlying disease severity. Please note that the outcome presented in **Web Fig. 2** is different from that in **Web Fig. 1**.

It should be remembered that studies could have significant heterogeneity if they were so different so as to be non-amenable to pooling in a meta-analysis in the first place.

Authors have the option of undertaking sensitivity analysis of the results of meta-analysis. Here, studies with low(er) methodological quality are excluded from the analysis, and the pooled estimates of effect of only the high-quality studies are examined. This helps to determine how 'sensitive' the pooled estimates are to the exclusion of methodologically lower quality studies. Lower quality studies are prone to higher risk of bias and tend to over-estimate the effect of interventions. Results that are not sensitive to the exclusion of lower quality studies (meaning that the overall effect remains unchanged, even if the magnitude changes) are expressed as robust results.

Step 8: Interpret the forest plot. The above steps facilitate interpretation of the pooled estimate of effect of the interventions being compared for one specific outcome, in terms of the parameter used to present the pooled estimate and the statistical model used to combine the data. Additionally, this is done considering the number of studies contributing to the pooled estimate, total number of participants, their individual characteristics and effects, methodological quality, and degree of heterogeneity.

Publication bias: Despite best efforts of systematic review authors to include all relevant studies addressing the research question, a review may be hampered by the non-availability of published primary studies. Generally, primary studies with positive results (i.e. showing evidence of efficacy of interventions) are more likely to be published than those showing negative results. This can result in publication bias, wherein the publication (or non-publication of some studies) determines the direction or strength of the overall evidence [26]. This is why high quality systematic reviews make tremendous efforts to search for unpublished literature.

There are several methods to assess the probability of publication bias in systematic reviews. Begg and Mazumdar rank correlation test [27] for publication bias correlates the ranks of effect sizes (of various studies in the meta-analysis) against the ranks of the variance in the treatment effect.

One of the popular methods to assess publication bias, is using funnel plots. This refers to a scatter plot of all the studies in a meta-analysis with effect size on the x-axis and standard error on the *y*-axis. Ideally the plot also shows the estimated effect size (with confidence intervals) and the predicted effect size (with confidence intervals). The plot also shows a vertical line that runs through the (adjusted) combined effect and the corresponding lower and upper bounds of the confidence interval. Such a plot visually highlights whether there is asymmetry in the distribution of the included studies, which hints at publication bias. This approach works only where there are more than ten studies in the meta-analysis. Egger regression method shows “the degree of funnel plot asymmetry as measured by the intercept from regression of standard normal deviates against precision” [28].

When publication bias is suspected, systematic review authors should measure the impact of this on the estimated effect. This can be done using Duval and Tweedle trim and fill technique [29] which mathematically adjusts the pooled effect, accounting for funnel plot asymmetry.

In reviews showing efficacy of interventions with publication bias, Rosenthal analysis or the ‘fail-safe N method’ was used to try and identify the number of additional studies (with negative results) that would be needed to make the pooled estimate statistically insignificant [30]. Of course, this depends on making assumptions of data in unobserved/unpublished studies, hence is itself fraught with bias(es).

APPLY (CONSIDERATIONS ABOUT APPLICATION OF THE RESULTS OF SYSTEMATIC REVIEWS)

Both users and producers of systematic reviews have to make value-based judgements on three important issues viz, (*i*) What does the evidence (accessed, assimilated, appraised and analysed to answer the research question) show; (*ii*) What is the quality of the overall evidence and the level of confidence that can be placed in it; and (*iii*) Can the evidence be considered for use in clinical situations? Careful analysis of these three issues leads to the next and final step in evidence-informed healthcare practice viz. discussion of the evidence with individual patients by healthcare personnel with clinical expertise, to arrive at a shared decision.

Several new initiatives have been introduced to help systematic review users make better sense of the data presented. One of these is the Summary of Findings Table (SoFT) [31], that shows the absolute as well as relative effect of the intervention (including parameters like number needed to treat), the quantity of evidence, and the certainty of available evidence (which is an indirect measure of quality). SoFT are prepared for each of the key outcomes.

Another approach is to grade the evidence quality using an approach popularized by the acronym GRADE (Grading of Recommendations, Assessment, Development and Evaluation) [31]. This approach allows systematic review producers and users to apply semi-objective judgments on factors that may limit the quality of evidence in a SR. The key factors used are study limitations (viz. risk of bias), inconsistency (due to heterogeneity), indirectness, imprecision, and publication bias. A detailed explanation of the GRADE approach is outside the scope of this article.

Often the various analyses in systematic reviews do not point in the same direction. A common situation is one wherein some measures of efficacy favor one treatment, whereas other measures do not. Further, sometimes efficacious interventions may be less safe, or there is insufficient data to confirm safety. Therefore, the overall decision on whether to use the intervention may need more information than that reported in a systematic review.

It should be emphasized that evidence-based practice is not the mere application of systematic review findings to patients (healthcare consumers), but only a summary of the best research evidence that needs to be integrated with clinical expertise and patient values and preferences, to arrive at a shared decision (between the healthcare recipient and provider). Thus paradoxically, a shared decision to ignore the findings of a SR, on account of issues related to clinical expertise and/or patient values, is also well-aligned with the principles of evidence-based healthcare.

Strengths, limitations and challenges of systematic reviews: Systematic reviews of well-designed and well-conducted studies are the keystone of high-quality research evidence. The information from systematic review can be included in development of evidence-based guidelines and recommendations, health technology assessment, healthcare policy decisions, or health payment/reimbursement decisions. However, systematic review only provide research evidence on what works in research settings (referred to as efficacy), but not necessarily on what will work in real-world settings (referred to as effectiveness). The gap between efficacy versus effectiveness, and methods to plug it, are beyond the scope of this article. Second, users of systematic review look for answers to decision questions (exemplified by: Shall I use this intervention?) whereas producers of systematic review generate answers to research questions (exemplified by: Does this intervention work?). The difference between answers to research questions and decision questions needs to be clearly understood for appropriate use of systematic review in clinical practice.

Although systematic review include many methodological refinements to reduce bias, they are completely dependent on the quantity and quality of the primary studies available to answer the research question. This can lead to the piquant situation where an excellent systematic review finds limited (or no) evidence, and concludes the need of more research. Although this does not diminish the value of the SR, it may sometimes be unhelpful for decision-makers.

Despite attempts to minimize bias, certain forms of bias can creep into systematic reviews. These include publication bias, sponsorship bias (sponsored studies are published more often, especially when they show significant results), and intentional or unintentional emphasis of systematic review authors to highlight only some aspects of the systematic review [32]. Some of these anticipated biases can be addressed by ensuring that the conduct and reporting of systematic review conform to guidelines established for the purpose. These are exemplified by the PRISMA tool [33,34]. PRISMA is an acronym for 'Preferred Reporting Items for Systematic reviews and Meta-Analyses'. The checklist comprises 27 individual items that systematic review authors are expected to report. It also includes a flow chart summarizing the output of literature search in terms of studies identified, screened (after removal of duplicate publications), eligible for inclusion, those excluded, and those actually included. Extensions of the original PRISMA tool include PRISMA-P for systematic review protocols, PRISMA-IPD for reviews with individual patient data, and PRISMA-NMA for network meta-analyses.

Finally, users of systematic reviews should not blindly believe everything presented in the review, but learn to critically appraise systematic review for validity, significance and applicability. Standard tools and checklists available for the purpose can be very helpful [35]. Last but not the least, readers of Indian Pediatrics may benefit from the Journal Club section wherein SRs have been critically appraised from time to time.

Key Messages

- Systematic reviews involve the application of scientific methods to reduce bias in review of literature.
- The key components of systematic reviews can be summarized as: Ask, Access, Assimilate, Appraise, Analyze and Apply.
- Meta-analysis is a statistical tool that provides pooled estimates of effect from the data extracted from individual studies included in the review.

REFERENCES

1. Sackett D, Strauss S, Richardson W, et al. Evidence-Based Medicine: How to practice and teach EBM.2nd ed. Edinburghl Churchill Livingstone: 2000.
2. Cook DJ, Mulrow CD, Haynes RB. Systematic reviews: Synthesis of best evidence for clinical decisions. Ann Intern Med. 1997;126:376-80.
3. PennState Eberley College of Science. Lesson 4: Bias and Random Error. Accessed October 01, 2020. Available from: *https://online.stat.psu.edu/stat 509/node/26/*
4. Comprehensive Meta-analysis. Accessed October 01, 2020. Available from: *https://www.meta-analysis.com/pages/why_do.php?cart=*

5. National Institute for Health Research. PROSPERO International prospective register of systematic reviews. Accessed October 01, 2020. Available from: *https://utas.libguides.com/SystematicReviews/Protocol*
6. Mathew JL, Singh M. Evidence based child health: Fly but with feet on the ground! Indian Pediatr. 2008;45:95-8.
7. Virginia Commonwealth University. How to conduct a literature review (Health Sciences). Accessed October 01, 2020. Available from: *https://guides.library. vcu.edu/health-sciences-lit-review/question*
8. http://www.knowledge.scot.nhs.uk/k2atoolkit/source/identify-what-you-need-to-know/spice.aspx
9. Cooke A., Smith D, Booth A. Beyond PICO: The SPIDER tool for qualitative evidence synthesis. Qualitative Health Research. 2012;22:1435-43.
10. Booth A, Noyes J, Flemming K, et al. Formulating questions to explore complex interventions within qualitative evidence synthesis. Accessed October 01, 2020. Available from: *https://library.nd.edu.au/evidencebased practice/ask/question*
11. Infolibrarian. Bibliographic databases. Accessed October 01, 2020. Available from: *http://infolibrarian.com/edb.html*
12. E-Resources for China Studies. Accessed October 01, 2020. Available from: *http://www.wanfangdata.com*
13. LILACS, health information from Latin America and the Caribbean countries. Accessed October 01, 2020. Available from: *https://lilacs.bvsalud.org/en/*
14. OpenGrey. System for information on grey literature in Europe. Accessed October 01, 2020. Available from: *http://www.opengrey.eu*
15. Sterne JAC, Savović J, Page MJ, et al. RoB 2: A revised tool for assessing risk of bias in randomised trials. BMJ. 2019;366:l4898.
16. Cochrane Training. RevMan 5. Accessed October 01, 2020. Available from: *https://training.cochrane.org/online-learning/core-software-cochrane-reviews/revman/revman-5-download*.
17. Wells GA, Shea B, O'Connell D, et al. The Newcastle-Ottawa Scale (NOS) for assessing the quality of non-randomised studies in meta-analyses. Accessed October 02, 2020. Available from: *http://www.ohri.ca/programs/clinical -epidemiology/oxford.asp*
18. Stang A. Critical evaluation of the Newcastle-Ottawa scale for the assessment of the quality of nonrandomized studies in meta-analyses. Eur J Epidemiol. 2010;25:603-5.
19. Sterne JA, Hernan MA, Reeves BC, et al. ROBINS-I: A tool for assessing risk of bias in non-randomised studies of interventions BMJ. 2016;355:i4919.

20. University of Bristol. QUADAS-2. Accessed October 01, 2020. Available from: *https://www.bristol.ac.uk/population-health-sciences/projects/quadas/quadas-2/*
21. OHAT Risk of Bias Rating Tool for Human and Animal Studies. Accessed February 27, 2021. Available from: *https://ntp.niehs.nih.gov/ntp/ohat/pubs/ riskofbiastool_508.pdf*
22. Hooijmans CR, Rovers MM, De Vries RB, et al. SYRCLE's risk of bias tool for animal studies. BMC Med Res Meth. 2014;14:1-9.
23. Cochrane Methods. About IPD meta-analyses. Accessed October 01, 2020. Available from: *https://methods.cochrane.org/ipdma/about-ipd-meta-analyses*
24. Borenstein M, Hedges L, Rothstein H. Meta-analysis. Fixed effect vs. random effects. Accessed October 01, 2020. Available from: *https://www.meta-analysis.com/downloads/M-a_f_e_v_r_e_sv.pdf*
25. Heterogeneity in Meta-analysis. Accessed October 01, 2020. Available from: *https://www.statsdirect.com/help/meta_analysis/heterogeneity.htm*
26. Dalton JE, Bolen SD, Mascha EJ. Publication bias: The elephant in the review. Anesth Analg. 2016;123:812-3.
27. Begg CB, Mazumdar M. Operating characteristics of a rank correlation test for publication bias. Biometrics. 1994;50:1088-101.
28. Egger M, Smith GD, Schneider M, et al. Bias in meta-analysis detected by a simple, graphical test. BMJ. 1997;315:629-34.
29. Duval S, Tweedie R. Trim and fill: A simple funnel-plot–based method of testing and adjusting for publication bias in meta-analysis. Biometrics. 2000;56:455-63.
30. Rosenthal R. The file drawer problem and tolerance for null results. Psycholog Bulletin. 1979;86:638-41.
31. Schünemann HJ, Higgins JPT, Vist GE, et al. Chapter 14: Completing 'Summary of findings' tables and grading the certainty of the evidence. Available from: Completing 'Summary of findings' tables and grading the certainty of the evidence. Accessed October 01, 2020.
32. Drucker AM, Fleming P, Chan AW. Research techniques made simple: Assessing risk of bias in systematic reviews. J Invest Dermatol. 2016;136:e109-e14.
33. Preferred Reporting Items for Systematic reviews and Meta-Analyses (PRISMA). Accessed October 01, 2020. Available from: *http://www.prisma-statement.org*
34. Moher D, Liberati A, Tetzlaff J, Altman DG, The PRISMA Group. Preferred Reporting Items for Systematic reviews and Meta-Analyses: The PRISMA Statement. PLoS Med. 2009;6:e1000097.
35. Critical Appraisal Skills Programme. 10 questions to help you make sense of a Systematic review. Accessed October 01, 2020. Available from: *https://casp-uk.net/wp-content/uploads/2018/01/CASP-Systematic-Review-Checklist_2018.pdf*

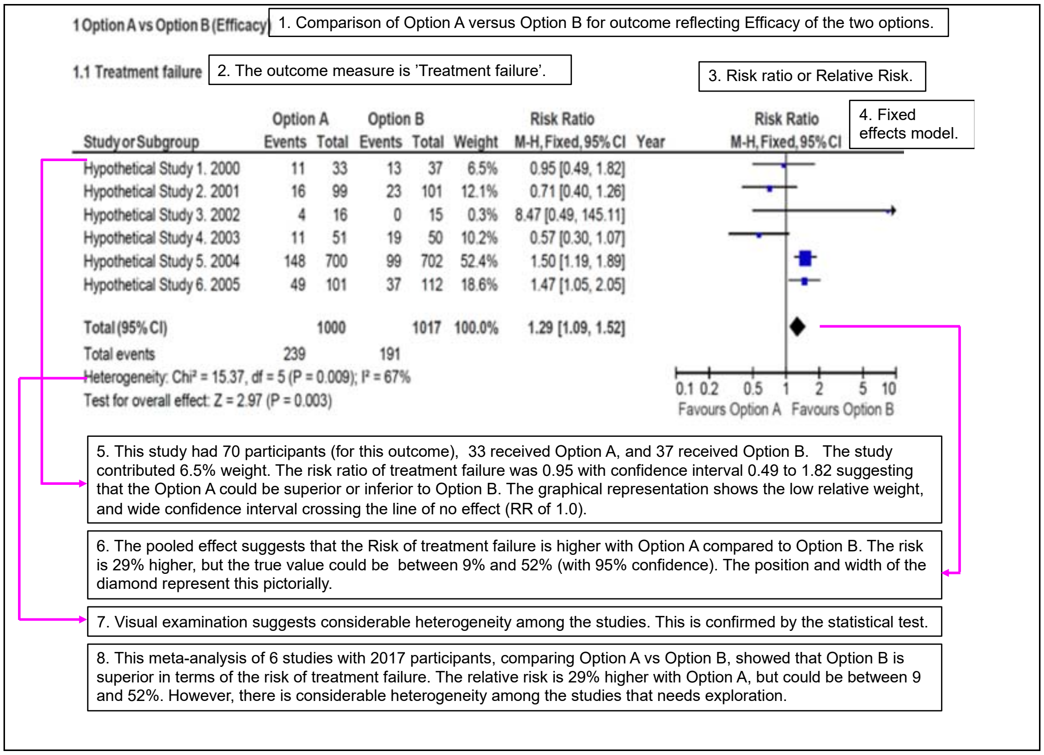
Table I Key Aspects of Systematic reviews

Key principle	*Interpretation*	*Remarks*
Ask	What is the specific research question 'asked' or addressed in the SR?	The entire methodology of a SR, interpretation of findings, and conclusions, depend on this.
Access	What literature sources were accessed (or searched) to identify the primary search studies to be included in the SR? What was the 'search strategy'?	The focus is to ensure that no study that can potentially answer the research question, gets missed.
Assimilate	What strategies were used to assimilate or synthesize or 'put together' the primary research studies?	In order to minimize bias, most systematic review prudently limit the included studies to those conforming to the best, or sometimes most appropriate study designs that can answer the research question.
Appraise	How were the included studies critically appraised for methodological quality?	This is to estimate the risk of bias in the primary studies, and the potential impact on the systematic review results and conclusions.
Analyze	What data was extracted from each primary study for synthesis? How were the data analyzed? What are the main findings? What is the level of confidence in these findings, based on the methodological aspects of the included studies?	The data extracted from the primary studies could be examined with a combination of qualitative and quantitative methods. Meta-analysis helps to obtain a pooled estimate of the included data.
Apply	Can the findings of the systematic review be applied in the patient or population of your interest?	Conclusions of a systematic review have to be integrated with clinical expertise and patient preferences/values for a truly evidence-informed healthcare decision.

Table II Applicability of PICOTS to Systematic reviews Addressing Various Types of Research Questions

Research question	*Intervention*	*Diagnosis*	*Prognosis*	*Prevalence/Incidence*	*Association*
Example	Is plasma exchange therapy beneficial in COVID-19?	Can 'loss of smell' be used to diagnose COVID-19?	Do people with COVID-19 having co-existing diabetes or hypertension, fare worse?	What proportion of patients with COVID-19, have or develop acute respiratory distress syndrome?	Does international travel result in COVID-19?
P=Patient/ Population	People with severe COVID-19	People with suspected COVID-19	People with confirmed COVID-19	People with COVID-19	Indian citizens, residing in the country.
I = Intervention or Exposure	Plasma exchange therapy	Confirmation of 'loss of smell'	(Controlled and uncontrolled) Diabetes, or Hypertension		International travel (within the preceding 21 days)
C = Comparison	No plasma exchange	Reverse transcriptase PCR for novel Coronavirus	None of the above		No international travel (within the preceding 21 days)
O = Outcomes	Mortality, Need for invasive ventilation, Side effects, Cost	Diagnostic accuracy, Cost	Disease severity, Need for intensive care, Mortality	Acute respiratory distress syndrome (ARDS)	Development of COVID-19
T = Time-frame	Within 30 days of treatment (for all outcomes)	Not applicable*	From diagnosis to recovery or discharge or death.	From diagnosis to recovery or discharge or death.	Within 28 days of the date of conclusion of the travel.
S = Study design	RCT	Diagnostic test study	Cohort study with comparison group	Cross-sectional study (for prevalence) Cohort study (for incidence)	Case-control study.

**Diagnostic test studies are cross-sectional in the sense that the index test (confirmation of loss of smell) and reference test (RT-PCR) should ideally be performed at the same time, or if that is not feasible, within a narrow interval, during which there is no probability of a change in the diagnostic status of a given patient (from negative to positive, or vice versa). Similarly, the gap between the index test and diagnostic test should not be such that people who receive one test may get cured, or drop-out, or die before the other test is administered.*



1 Option A vs Option B (Efficacy)

1. Comparison of Option A versus Option B for outcome reflecting Efficacy of the two options.

1.1 Treatment failure

2. The outcome measure is 'Treatment failure'.

3. Risk ratio or Relative Risk.

4. Fixed effects model.

	Option A		Option B			Risk Ratio	
Study or Subgroup	Events	Total	Events	Total	Weight	M-H, Fixed, 95% CI	Year
Hypothetical Study 1. 2000	11	33	13	37	6.5%	0.95 [0.49, 1.82]	
Hypothetical Study 2. 2001	16	99	23	101	12.1%	0.71 [0.40, 1.26]	
Hypothetical Study 3. 2002	4	16	0	15	0.3%	8.47 [0.49, 145.11]	
Hypothetical Study 4. 2003	11	51	19	50	10.2%	0.57 [0.30, 1.07]	
Hypothetical Study 5. 2004	148	700	99	702	52.4%	1.50 [1.19, 1.89]	
Hypothetical Study 6. 2005	49	101	37	112	18.6%	1.47 [1.05, 2.05]	
Total (95% CI)		**1000**		**1017**	**100.0%**	**1.29 [1.09, 1.52]**	
Total events	239		191				

Heterogeneity: Chi² = 15.37, df = 5 (P = 0.009); I² = 67%

Test for overall effect: Z = 2.97 (P = 0.003)

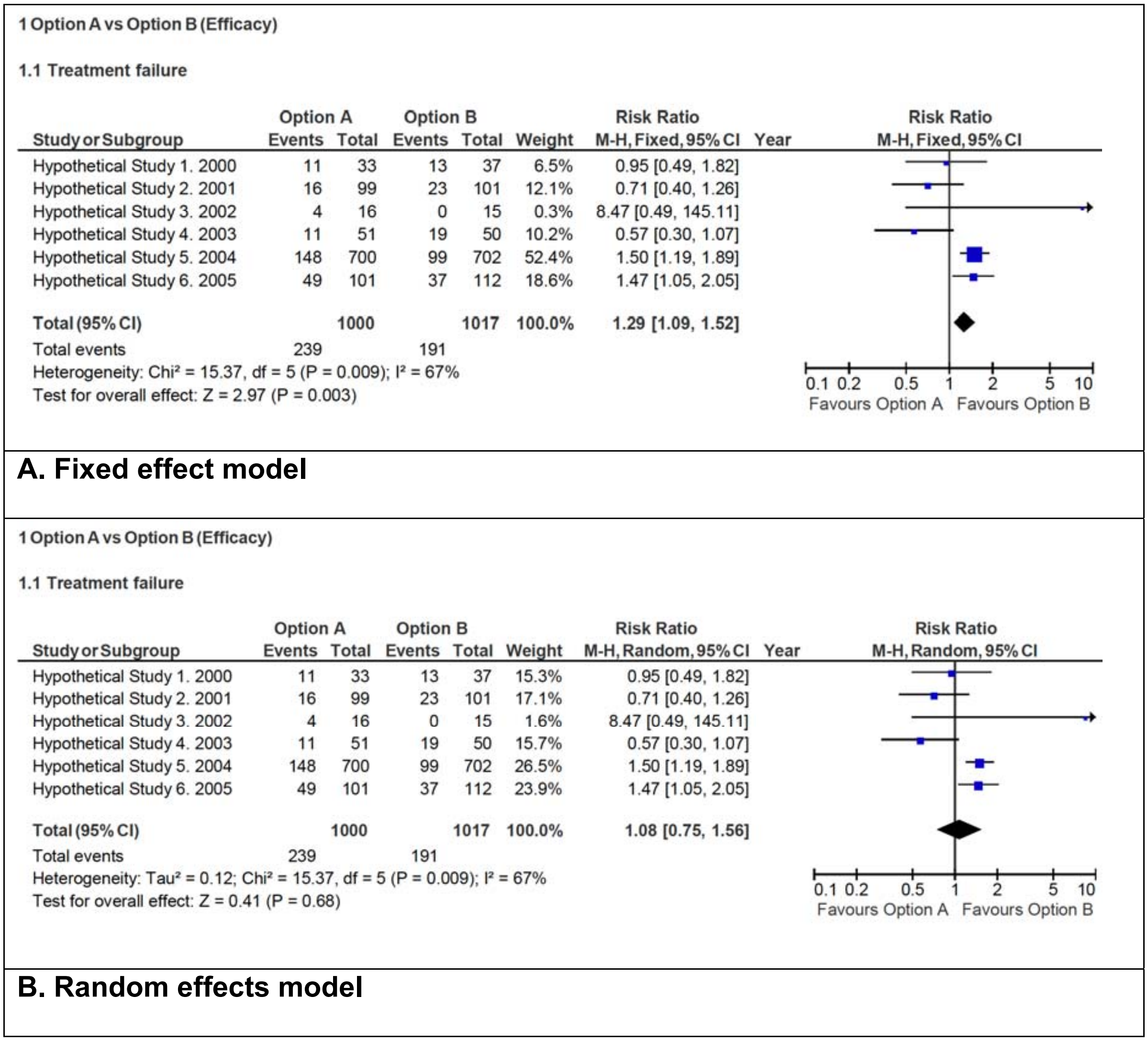
5. This study had 70 participants (for this outcome), 33 received Option A, and 37 received Option B. The study contributed 6.5% weight. The risk ratio of treatment failure was 0.95 with confidence interval 0.49 to 1.82 suggesting that the Option A could be superior or inferior to Option B. The graphical representation shows the low relative weight, and wide confidence interval crossing the line of no effect (RR of 1.0).

6. The pooled effect suggests that the Risk of treatment failure is higher with Option A compared to Option B. The risk is 29% higher, but the true value could be between 9% and 52% (with 95% confidence). The position and width of the diamond represent this pictorially.

7. Visual examination suggests considerable heterogeneity among the studies. This is confirmed by the statistical test.

8. This meta-analysis of 6 studies with 2017 participants, comparing Option A vs Option B, showed that Option B is superior in terms of the risk of treatment failure. The relative risk is 29% higher with Option A, but could be between 9 and 52%. However, there is considerable heterogeneity among the studies that needs exploration.

Fig. 1 Step-wise Interpretation of a Forest Plot



1 Option A vs Option B (Efficacy)

1.1 Treatment failure

Study or Subgroup	Option A Events	Option A Total	Option B Events	Option B Total	Weight	Risk Ratio M-H, Fixed, 95% CI	Year
Hypothetical Study 1. 2000	11	33	13	37	6.5%	0.95 [0.49, 1.82]	
Hypothetical Study 2. 2001	16	99	23	101	12.1%	0.71 [0.40, 1.26]	
Hypothetical Study 3. 2002	4	16	0	15	0.3%	8.47 [0.49, 145.11]	
Hypothetical Study 4. 2003	11	51	19	50	10.2%	0.57 [0.30, 1.07]	
Hypothetical Study 5. 2004	148	700	99	702	52.4%	1.50 [1.19, 1.89]	
Hypothetical Study 6. 2005	49	101	37	112	18.6%	1.47 [1.05, 2.05]	
Total (95% CI)		**1000**		**1017**	**100.0%**	**1.29 [1.09, 1.52]**	
Total events	239		191				

Heterogeneity: Chi² = 15.37, df = 5 (P = 0.009); I² = 67%
Test for overall effect: Z = 2.97 (P = 0.003)

A. Fixed effect model

1 Option A vs Option B (Efficacy)

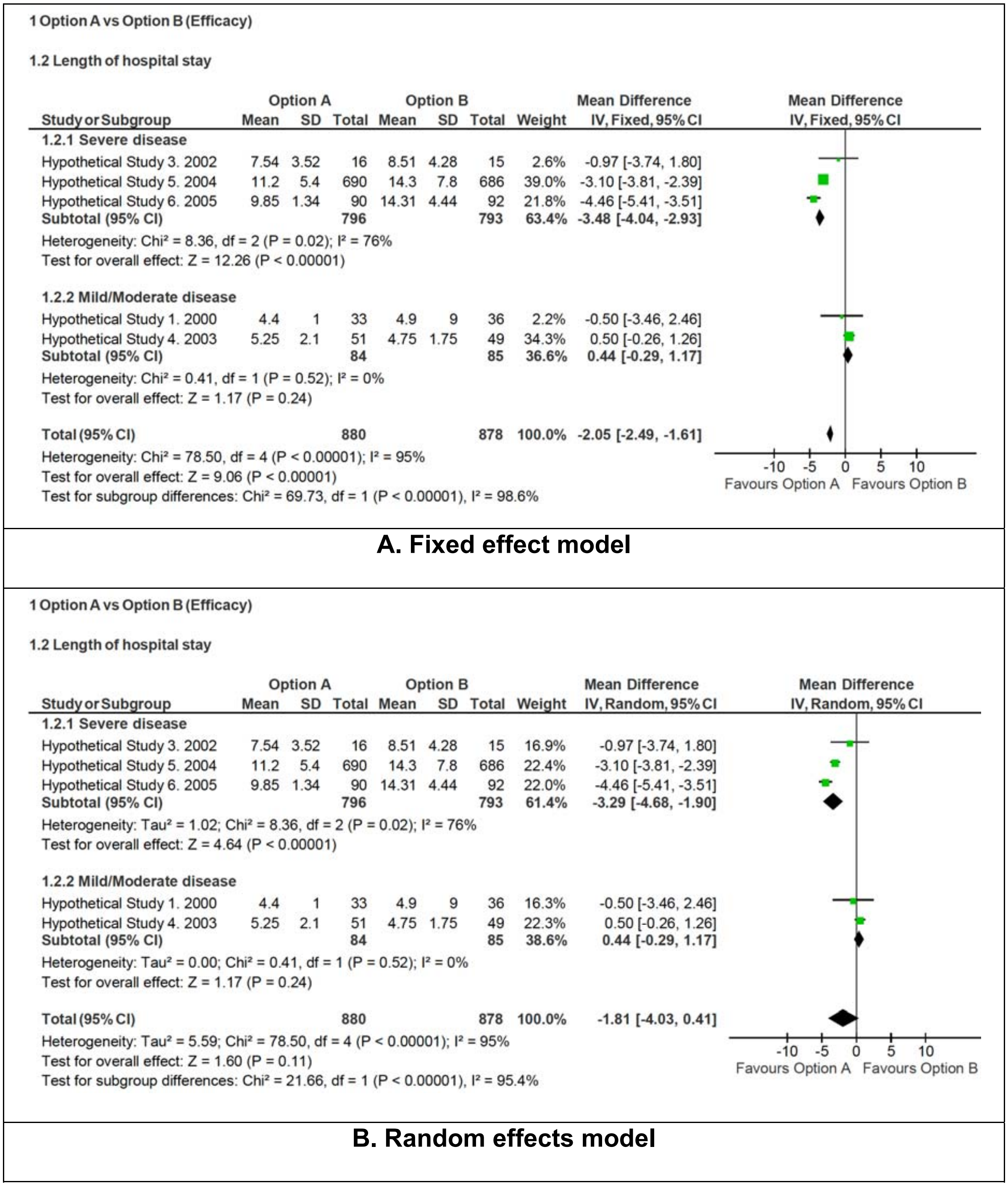
1.1 Treatment failure

Study or Subgroup	Option A Events	Option A Total	Option B Events	Option B Total	Weight	Risk Ratio M-H, Random, 95% CI	Year
Hypothetical Study 1. 2000	11	33	13	37	15.3%	0.95 [0.49, 1.82]	
Hypothetical Study 2. 2001	16	99	23	101	17.1%	0.71 [0.40, 1.26]	
Hypothetical Study 3. 2002	4	16	0	15	1.6%	8.47 [0.49, 145.11]	
Hypothetical Study 4. 2003	11	51	19	50	15.7%	0.57 [0.30, 1.07]	
Hypothetical Study 5. 2004	148	700	99	702	26.5%	1.50 [1.19, 1.89]	
Hypothetical Study 6. 2005	49	101	37	112	23.9%	1.47 [1.05, 2.05]	
Total (95% CI)		**1000**		**1017**	**100.0%**	**1.08 [0.75, 1.56]**	
Total events	239		191				

Heterogeneity: Tau² = 0.12; Chi² = 15.37, df = 5 (P = 0.009); I² = 67%
Test for overall effect: Z = 0.41 (P = 0.68)

B. Random effects model

Web Fig. 1. *Panels showing data analyzed using the FE model (Panel A) and RE model (Panel B). Although the same data from each study were analyzed, the relative weights became different. In the FE model, studies with larger sample sizes have disproportionately larger weight compared to the RE model. This results in a change in the pooled estimate of effect, though the effect of each study remains unchanged. In this analysis, the FE model showed a statistically significant pooled effect favouring Option B (Relative risk of treatment failure was 29% higher with Option A, and the limits of the 95% confidence interval remained on the same side of 1.0, confirming a statistically significant effect). In contrast, the RE model provided a more conservative estimate of the pooled effect (RR 1.08 compared to 1.29 in the FE model). In this analysis, the 95% confidence interval limits crossed 1.0, suggesting the absence of a statistically significant pooled effect.*



1 Option A vs Option B (Efficacy)

1.2 Length of hospital stay

Study or Subgroup	Option A Mean	Option A SD	Option A Total	Option B Mean	Option B SD	Option B Total	Weight	Mean Difference IV, Fixed, 95% CI
1.2.1 Severe disease								
Hypothetical Study 3. 2002	7.54	3.52	16	8.51	4.28	15	2.6%	-0.97 [-3.74, 1.80]
Hypothetical Study 5. 2004	11.2	5.4	690	14.3	7.8	686	39.0%	-3.10 [-3.81, -2.39]
Hypothetical Study 6. 2005	9.85	1.34	90	14.31	4.44	92	21.8%	-4.46 [-5.41, -3.51]
Subtotal (95% CI)			**796**			**793**	**63.4%**	**-3.48 [-4.04, -2.93]**
Heterogeneity: Chi² = 8.36, df = 2 (P = 0.02); I² = 76%								
Test for overall effect: Z = 12.26 (P < 0.00001)								
1.2.2 Mild/Moderate disease								
Hypothetical Study 1. 2000	4.4	1	33	4.9	9	36	2.2%	-0.50 [-3.46, 2.46]
Hypothetical Study 4. 2003	5.25	2.1	51	4.75	1.75	49	34.3%	0.50 [-0.26, 1.26]
Subtotal (95% CI)			**84**			**85**	**36.6%**	**0.44 [-0.29, 1.17]**
Heterogeneity: Chi² = 0.41, df = 1 (P = 0.52); I² = 0%								
Test for overall effect: Z = 1.17 (P = 0.24)								
Total (95% CI)			**880**			**878**	**100.0%**	**-2.05 [-2.49, -1.61]**
Heterogeneity: Chi² = 78.50, df = 4 (P < 0.00001); I² = 95%								
Test for overall effect: Z = 9.06 (P < 0.00001)								
Test for subgroup differences: Chi² = 69.73, df = 1 (P < 0.00001), I² = 98.6%								

A. Fixed effect model

1 Option A vs Option B (Efficacy)

1.2 Length of hospital stay

Study or Subgroup	Option A Mean	Option A SD	Option A Total	Option B Mean	Option B SD	Option B Total	Weight	Mean Difference IV, Random, 95% CI
1.2.1 Severe disease								
Hypothetical Study 3. 2002	7.54	3.52	16	8.51	4.28	15	16.9%	-0.97 [-3.74, 1.80]
Hypothetical Study 5. 2004	11.2	5.4	690	14.3	7.8	686	22.4%	-3.10 [-3.81, -2.39]
Hypothetical Study 6. 2005	9.85	1.34	90	14.31	4.44	92	22.0%	-4.46 [-5.41, -3.51]
Subtotal (95% CI)			**796**			**793**	**61.4%**	**-3.29 [-4.68, -1.90]**
Heterogeneity: Tau² = 1.02; Chi² = 8.36, df = 2 (P = 0.02); I² = 76%								
Test for overall effect: Z = 4.64 (P < 0.00001)								
1.2.2 Mild/Moderate disease								
Hypothetical Study 1. 2000	4.4	1	33	4.9	9	36	16.3%	-0.50 [-3.46, 2.46]
Hypothetical Study 4. 2003	5.25	2.1	51	4.75	1.75	49	22.3%	0.50 [-0.26, 1.26]
Subtotal (95% CI)			**84**			**85**	**38.6%**	**0.44 [-0.29, 1.17]**
Heterogeneity: Tau² = 0.00; Chi² = 0.41, df = 1 (P = 0.52); I² = 0%								
Test for overall effect: Z = 1.17 (P = 0.24)								
Total (95% CI)			**880**			**878**	**100.0%**	**-1.81 [-4.03, 0.41]**
Heterogeneity: Tau² = 5.59; Chi² = 78.50, df = 4 (P < 0.00001); I² = 95%								
Test for overall effect: Z = 1.60 (P = 0.11)								
Test for subgroup differences: Chi² = 21.66, df = 1 (P < 0.00001), I² = 95.4%								

B. Random effects model

Web Fig. 2 *Panels showing comparison of length of hospital stay (as an efficacy outcome). Panel A shows data analyzed using the FE model, and Panel B the RE model. Mean difference in length of stay (days) was used to compare the groups. In this example, only 5 of the 6 studies presented in Figure 1 reported the outcome, hence the total sample size is smaller than in Figure 1. It is evident that both analysis models identified a statistically significant reduction in hospital stay among studies enrolling participants with severe disease, whereas there was no such effect among participants with mild or moderate disease severity. The pooled estimate appears to be influenced by the greater weight of the studies in participants with severe disease. Here also, the difference in the pooled estimate, by the method of analysis is evident. While the FE model produced a statistically significant pooled estimate, the RE model did not. Since there is great heterogeneity among the studies (I^2 95%), in this situation, the RE model would be more appropriate*.