



ORIGINAL ARTICLE

Critical appraisal (II): Application to scientific articles on therapeutic interventions, systematic reviews and meta-analyses



Álvaro Gimeno Díaz de Atauri^{a,*}, María Salomé Albi Rodríguez^a,
Elena Pérez González^b, Jaime Javier Cuervo Valdés^c, Juan Ruiz-Canela Cáceres^d,
Javier González de Dios^e, Comité de Trabajo de Pediatría Basada en la Evidencia¹

^a Servicio de Pediatría, Hospital Universitario 12 de Octubre, Madrid, Spain

^b Servicio de Pediatría, Hospital Universitario Virgen Macarena, Sevilla, Spain

^c Centro de Salud Ciudad Jardín, Badajoz, Spain

^d Pediatra, Sevilla, Spain

^e Servicio de Pediatría, Hospital General Universitario Dr. Balmis, Universidad Miguel Hernández, Instituto de Investigación Sanitaria y Biomédica de Alicante (ISABIAL), Alicante, Spain

Received 7 March 2025; accepted 24 April 2025

Available online 16 October 2025

KEYWORDS

Evidence-based
medicine;
Critical appraisal;
Systematic review;
Clinical trial;
Meta-analysis

Abstract In this article, we discuss the methodology for the critical appraisal of clinical trials (CTs) and systematic reviews (SRs). Studies of either type, when well designed and executed, provide the best evidence for clinical decision-making. Clinical trials are the primary source of evidence for evaluating therapeutic interventions, and SRs analyze and synthesize research of a certain quality to answer a specific clinical question. To assess their validity, it is essential that they aim to answer a specific question in terms of population, intervention and outcomes and that the risk of bias is reduced as much as possible during their development, from design to publication. The article addresses the steps to take to assess whether a CE or SR meets the criteria for its results to be considered reliable (valid). Once their validity is confirmed, the magnitude and clinical relevance of the results must be assessed, in addition to their applicability to our patients in particular.

© 2025 Asociación Española de Pediatría. Published by Elsevier España, S.L.U. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

DOI of original article: <https://doi.org/10.1016/j.anpedi.2025.503891>

* Corresponding author.

E-mail address: alvaro.gimeno@salud.madrid.org (Á. Gimeno Díaz de Atauri).

¹ The remaining members of the Evidence-Based Pediatrics Working Committee are presented in [Appendix 1](#).

PALABRAS CLAVE

Medicina basada en la evidencia;
Lectura crítica;
Revisión sistemática;
Ensayo clínico;
Metanálisis

Lectura crítica (II): aplicación en artículos científicos sobre intervenciones terapéuticas, revisiones sistemáticas y metanálisis

Resumen En este artículo se aborda la metodología para la lectura crítica de estudios relacionados con ensayos clínicos (EC) y revisiones sistemáticas (RS). Ambos estudios, cuando están bien diseñados y ejecutados, proporcionan la mejor evidencia para la toma de decisiones clínicas. Los EC son la fuente primaria para evaluar intervenciones terapéuticas y las RS analizan y sintetizan los estudios de calidad para responder a una pregunta clínica específica. Para evaluar su validez, es fundamental que pretendan responder a una pregunta concreta en cuanto a población, intervención y resultados que se van a evaluar y que en su desarrollo, desde el diseño hasta la publicación, se elimine al máximo el riesgo de sesgos. En este artículo se abordan los pasos para analizar si un EC o una RS cumplen criterios para que sus resultados sean fiables (v. Una vez confirmado este hecho, debemos analizar tanto su magnitud como su relevancia clínica, así como la aplicabilidad en nuestros pacientes concretos).

© 2025 Asociación Española de Pediatría. Publicado por Elsevier España, S.L.U. Este es un artículo Open Access bajo la CC BY-NC-ND licencia (<http://creativecommons.org/licencias/by-nc-nd/4.0/>).

Introduction

This article summarizes current methods for the critical appraisal of studies of therapeutic interventions and of systematic reviews (SRs). To complement this theoretical summary, numerous practical examples are provided in the journal of the Spanish Evidence-Based Pediatrics Working Group (<https://evidenciasenpediatria.es/>). A standardized approach to critical appraisal according to the recommendations of the international Evidence Based Medicine Working Group, as adapted by the Critical Appraisal Skills Program-Español (CASPe), involves three steps: assessing the validity, the importance and the applicability of the results.¹

Critical appraisal of studies of therapeutic interventions

The randomized controlled trial (RCT) is the gold standard for assessing the efficacy and safety of therapeutic interventions. It is a prospective experimental study in which the researchers control and manipulate the independent variables and patients are randomly allocated to the different interventions that are being compared. There are several types of clinical trial (Table 1), the development of which always involves seven key steps (Fig. 1). Critical appraisal of RCTs involves asking a series of questions²⁻⁷:

A Assess validity

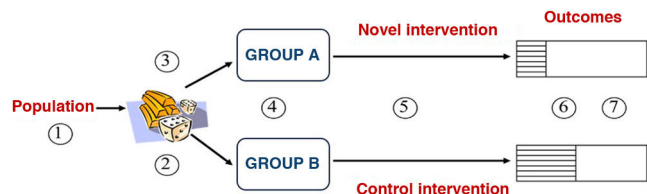
a) *Primary criteria*: if the study does not meet them, continuing to review it is probably not worth the effort.

* Did the study address a clearly formulated research question?

Randomized controlled trials should be designed to answer a single question. This question must be clearly structured in terms of the population under study, the implemented intervention and the outcomes of interest. They

Table 1 Main types of clinical trials.

Based on objectives	Explanatory Pragmatic
Based on participating researchers/centers	Single-center Multicentric
Based on methodology	Controlled Uncontrolled
Based on blinding	Open label (not blinded) Single blind Double blind Triple blind
Based on randomization	Randomized Nonrandomized
Based on design	Comparative Crossover Sequential Within-patient Matched Other



1. Structured clinical question
2. Randomization
3. Concealed randomization sequence
4. Blinding
5. Complete follow-up
6. Results: intention-to-treat analysis
7. Results: reporting

Figure 1 Basic elements and methodology of a randomized controlled trial. Seven steps.

must have only one primary outcome, and this outcome must be clinically relevant.

* Was the assignment of patients to treatment randomized?

Randomization is defined as the systematic and reproducible procedure by which participants are randomly assigned to different treatment groups. It is intended to ensure that there is an even distribution of all variables, both known and unknown, among the groups (in direct proportion to the sample size).

* Were all patients who entered the trial properly accounted for and attributed at its conclusion? Was follow-up complete?

It is important to assess whether the losses are qualitatively or quantitatively significant enough to invalidate the results. Pre-randomization losses fundamentally affect the ability to generalize the results (it is widely accepted that if the sample includes 80% to 90% of eligible subjects, it is representative of the population). Post-assignment losses can affect the certainty of the results (if attrition or losses to follow-up amount to more than 20% of the sample). It is important to assess for differences between groups in the number of losses.

* Were participants analyzed in the groups to which they were randomized (intention-to-treat analysis)?

Assess whether the analysis includes all subjects randomized to the treatment groups, regardless of whether they received any dose of the assigned treatment or whether there was any error in administration. In this case, the analysis is by intention to treat, as opposed to per-protocol analysis (according to the received treatment).

b) *Secondary criteria*: these allow three possible answers (*yes, no, can't tell*). They help detect the risk of bias, but them not being met does not completely invalidate the results.

* Were patients, their clinicians and study personnel "blind" to treatment allocation?

Blinding or masking refers to a series of precautions taken to ensure that the patient, researchers or both remain unaware of the treatment allocation. The purpose is to prevent cognitive bias in all involved parties, but especially participants (placebo effect, nocebo effect, attrition bias) and researchers (observer bias, attrition bias).

* Were the groups similar at the start of the trial?

If there are baseline differences in some of the characteristics of the compared groups, the validity of the study may be compromised. This question is usually answered by analyzing the initial table outlining the characteristics of the intervention and control group.

* Aside from the experimental intervention, were the groups treated equally?

Assess whether interventions other than the treatment under study (cointerventions) have been applied differentially to the groups and whether this can affect the results.

B Assess the *importance* of the results

* How large was the treatment effect?

The main measure indicative of the validity of evidence is the strength of association, which can be assessed by

means of the relative risk (RR), chiefly used in RCTs and cohort studies, or the odds ratio (OR), chiefly used in case-control studies and meta-analyses (MAs). The RR is obtained by dividing the risk of a given event in the treatment group by the risk of the event in the control group. A RR value of 1 indicates the absence of an association, values greater than 1 a positive association and values of less than one a negative association. Point estimates offer little information and should always be accompanied by the corresponding confidence interval (CI).

The relative risk reduction (RRR) is calculated by subtracting the proportion of events in the treatment group from the proportion of events in the control group and dividing the difference by the proportion in the control group. This measure has a limitation: it cannot differentiate between large and small absolute effects of a treatment. This drawback can be addressed with a different measure, the absolute risk reduction, calculated by subtracting the proportion of events in the treatment group from the proportion of events in the control group, and has the advantage of taking into account the baseline susceptibility of the patients.

Once the clinician feels confident in regard to the validity and the strength of the association, the data need to be translated into a measure that clearly reflects the impact on clinical practice. The most useful measure is the number needed to treat (NNT), which is calculated as the inverse of the absolute risk reduction. The NNT measures the therapeutic "effort" that must be made to prevent one additional undesirable event. But how can one tell whether this value is large or small? The answer lies not in statistical significance, but in clinical significance (which depends on the importance of the event to be prevented, but also on the cost, feasibility of adherence and safety of the treatment). The CI of the NNT will be essential for understanding the precision of the results and their impact. [Table 2](#) provides a hypothetical example of the calculation of the effect measures of a therapeutic intervention.

* How precise was the estimate of the treatment effect?

The CI provides a range of values that is likely to contain the true population value based on the results obtained in the study. Therefore, it is a measure of the precision (or uncertainty) of the findings. The CI provides quantitative information (as opposed to the *P* value, which only reports the probability that the observed differences are due to chance [statistical significance], independently of their magnitude).

C Assess the *applicability* of the results of the RCT to clinical practice

* Will the results help me in caring for my patient?

This is not simply a matter of determining whether the characteristics of study participants and the clinician's patients differ, but also whether any such differences are likely to alter outcomes. Some of the aspects to consider are biological factors (eg, are there pathophysiological differences in the disease that could lead to differences in the response to treatment? are there differences in patient characteristics that could modify the response to treatment?), socioeconomic factors (eg, could there be differences in the adherence to treatment at the level of the patient or the

Table 2 Measures of the effect of a therapeutic intervention. Results of a hypothetical randomized clinical trial comparing a treatment for prevention of a disease with placebo.

	Diseased	Not diseased	
Treatment	25	225	250
	a	b	a + b
Placebo	50	200	250
	c	d	c + d
	75	425	500
	a + c	b + d	a + b + c + d

Risk (incidence) in treatment group:
 $I_t = a / (a + b) = 25 / 250 = 0.10$ (10%).
 Risk (incidence) in placebo group:
 $I_0 = c / (c + d) = 50 / 250 = 0.20$ (20%).
 Relative risk:
 $RR = I_t / I_0 = 0.10 / 0.20 = 0.50$.
 The risk of disease is 50% lower in treated patients (reduced from 20% to 10%).
 Relative risk reduction:
 $RRR = (I_0 - I_t) / I_0 = (0.20 - 0.10) / 0.20 = 0.50$ (50%).
 The treatment increases the probability of being cured by 50% compared to placebo.
 Absolute risk reduction:
 $ARR = I_0 - I_t = 0.20 - 0.10 = 0.10$ (10%).
 Out of 100 treated patients, 10 more patients will be cured compared to not having treated the patients.
 Number needed to treat:
 $NNT = 1 / ARR = 1 / 0.10 = 10$.
 Ten patients need to be treated to prevent disease in one of them.

health care system that could reduce the response to treatment?) and epidemiological factors (eg, does the patient have any concurrent health problems that could significantly affect the risks and benefits of treatment?).

* What would be the potential benefit of applying the intervention to this particular patient?

In deciding whether to implement a particular intervention, it is important to analyze the potential benefit to the patient. For example, caution should be exerted when studies use an intermediate variable to evaluate the effect of the intervention rather than a clinical event of greater direct relevance to the patient (eg, a decrease in cholesterol versus a decrease in ischemic events).

* Were all clinically important outcomes considered?

All clinically relevant outcomes that may result from treatment need to be analyzed, such as adverse events or the impact on quality of life.

* Are the likely treatment benefits worth the potential harms and costs?

When it comes to a simple, risk-free health care intervention that is highly effective in reducing mortality or preventing serious disease, the decision is easy. On other occasions, a detailed review of all aspects related to the intervention, its potential beneficial and harmful effects, implementation, costs, feasibility, degree of adherence and accessibility is required.

* Do the intervention and its consequences satisfy the values and preferences of the patient?

Clinicians need to explain to the patient what they have concluded regarding the potential results of implementing the intervention in the patient's particular case. In making the final decision, clinicians must consider the opinion of the patient and support them in making responsible decisions about their health with respect and empathy.

Critical appraisal of reports of systematic reviews and meta-analyses

Systematic reviews and meta-analyses, methods for summarizing qualitative and quantitative data, respectively, are considered some of the best sources of scientific evidence, although, as any other retrospective research method, they have limitations that chiefly stem from the quality of the available evidence and its reproducibility. The appraisal of these sources requires steps similar to those employed in the appraisal of RCTs, assessing their validity, importance and applicability.^{8,9}

A Assess validity

a) Primary criteria:

* Did the SR address a clearly focused question?

A specific and relevant clinical question must be asked in terms of the condition of interest (disease under study), population, exposure (to a treatment, diagnostic test, potential adverse effect, etc) and one or more outcomes of interest.

* Did the authors perform a comprehensive search for appropriate study designs?

The criteria used to identify and select articles eligible for inclusion must be explicitly defined and appropriate for the question at hand. The best studies are those that address the question being reviewed and have the appropriate design for answering it. The eligibility criteria applied for their selection should be based on the customary validity criteria (specific to the different types of scientific articles).

In addition to the study design, the reader must consider the characteristics of study subjects, the setting of the evaluated intervention and any comparative intervention, the control groups and the clinical outcomes of interest, so that studies are relatively homogeneous in terms of the health problem of interest and the methodological approach used to study it. Even when the same clinical problem is being investigated, differences in the essential elements of a study (patients, interventions, and outcomes) can lead to different conclusions.

b) Secondary criteria:

* Are all the important, relevant studies likely to have been included?

The validity of a SR depends on whether all available scientific evidence has been evaluated and no relevant studies have been left out. A comprehensive and unbiased search is one of the key differences between a SR and a narrative review. The articles retrieved from electronic bibliographic databases can vary greatly depending on the implemented search strategy, so the authors need to give specific information on the search strategy and the criteria used for selecting articles, as well as the consulted databases. Authors should

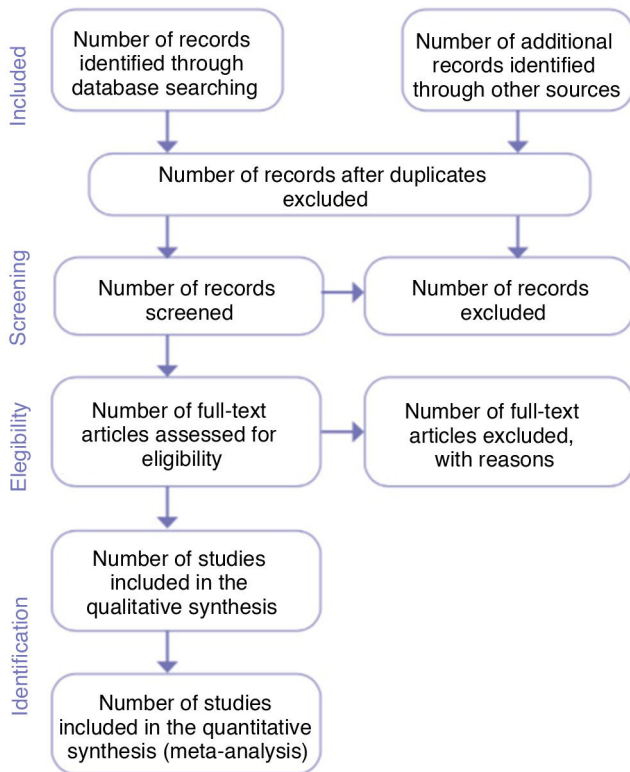


Figure 2 Flow of information through the different phases of a systematic review.

also have performed backward citation searching and, if necessary, contacted experts to identify unpublished studies as well. This reduces the risk of publication bias (higher probability of publication of studies with positive results) and, therefore, of overestimating efficacy. Ideally, the article selection process is summarized graphically in a flow diagram. Fig. 2 presents the flow diagram proposed in 2009 in the Preferred Reporting Items for Systematic reviews and Meta-Analyses (PRISMA) statement,¹⁰ widely used in the SRs published to date. The PRISMA guideline has been updated recently to include more detailed information.¹¹

* Did the authors assess the quality of included studies thoroughly and rigorously enough?

Peer review does not guarantee the validity of published research. Methodological differences between studies may explain significant differences in outcomes (less rigorous studies tend to overestimate the effectiveness of therapeutic interventions). The criteria for assessing validity must be explicit. This assessment, which is always subject to variability (through both random and systematic errors), must have been carried out independently by more than one reviewer, and the degree of agreement on which studies to include and which data to extract from them should have been assessed and reported. One of the tools that is useful for assessment of SRs on health care interventions is the AMSTAR 2 (A MeaSurement Tool to Assess systematic Reviews, version 2).¹²

* If the results of the studies in the systematic review were combined, was it reasonable to do so?

The derivation of an overall estimate of the effect of interest from various studies relies on the assumption that

the differences between them are due exclusively to chance or, in other words, that the results of the different studies are homogeneous.

Most reviews report important differences in the patients, exposures, outcome measures and methods between studies. The reader must determine whether the authors of the review assessed this heterogeneity. The most widely used method is the funnel plot. Some of the statistical methods commonly used to assess heterogeneity are the Cochran Q (significant heterogeneity if $P < 0.05$) and the I^2 , based on which heterogeneity is classified as low based on value thresholds of 25%, 50% and 75%, respectively. If there is significant heterogeneity in the primary sources, the authors have several options: (1) to refrain from combining the results and limit the review to a qualitative synthesis of their results; (2) to carry out a stratified analysis according to the variable that is the source of the heterogeneity, if it can be identified, or (3) resort to meta-regression, a more complex approach.

B Assess the *importance* of the results

* What were the overall findings of the review? Were they interpreted appropriately?

The goal is to estimate the overall effect of an intervention based on a weighted average of the pooled results of all available high-quality studies. Usually, the results of each study are weighted by the inverse of the study's variance, so that greater weight is given to larger studies, which provide more precise estimates. In some cases, the methodological quality of the study is taken into account in the weighting method. Results are usually expressed in terms of a relative measure (OR, RR or RRR). In addition to synthesizing quantitative data, the authors should have correctly tabulated relevant information regarding the included primary studies (inclusion criteria, sample/group sizes, patient characteristics, key study characteristics, primary and secondary outcomes). Meta-analysis results are usually displayed graphically, and the standard means for doing so is currently the forest plot (Fig. 3).

* How precise are the results?

The size of the effect observed in a study is but a point estimate of the true size of the effect or strength of the association of interest. If a different sample of individuals had been studied (or a different collection of studies, in the case of SRs and/or meta-analyses), the result could have differed, although, presumably, not by much. It is important to determine the precision of the estimate by means of the CI. The 95% CI is the range of values that one can be 95% confident contains the true effect size. When the CI includes the null value (0 in the case of measuring the effect of a difference and 1 in the case of relative measures calculated as quotients, such as the RR or OR), the result is not statistically significant.

C Assess the *applicability* to clinical practice of systematic review or meta-analysis reports

* Can the results be applied to the local population or context?

The external validity of the studies included in the review must be assessed. Does the report describe the character-

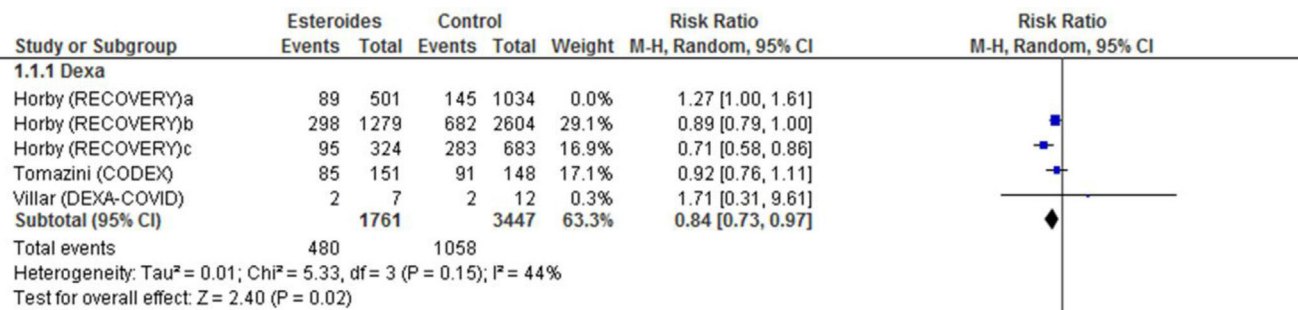


Figure 3 Example of a forest plot presenting meta-analysis results, used with the authorization of the authors (Committee/Group on Evidence-Based Pediatrics of the AEP and AEPap. *COVID-19 en Pediatría: Valoración Crítica de la Evidencia*. 2nd Ed. Updated February 28, 2022). Structured in five columns:

1. Primary studies: (author and year) with the groups or subgroups of patients included in the meta-analysis.
2. Results of each study: intervention group and control group.
3. Amount of weight awarded to each study, as a proportion.
4. Estimates of the effect with their confidence intervals.
5. Representation of the forest plot: the result of each study is a horizontal line whose width represents the 95% CI (precision); the vertical line is the “no effect” reference and corresponds to an OR = 1 or a difference in means = 0 (if the study line crosses the no effect line, it means that there are no significant differences between the treatments or that the sample size is insufficient to detect such differences); the diamond corresponds to the pooled results of all RCTs, providing a new CI that is much narrower, that is, more precise.

istics of the patients included in the primary studies? Were those patients similar to the local population? There may be doubts concerning applicability due to the specific characteristics of the local patient population (such as age, disease severity or comorbidities), which may lead to the additional challenges of subgroup analysis; in any case, detailed guidelines are available to help determine whether subgroup analyses are trustworthy.

* Did the authors consider all outcomes that are relevant for clinical decision-making?

Although searching for SRs focused on a specific clinical question (primary outcome) is generally recommended, as it is more likely to yield valid results, this does not mean that other outcomes of interest (secondary outcomes) should be neglected.

* Did the authors consider all possible clinically relevant outcomes?

It is important to assess not only the outcome of interest, but any other clinically relevant responses or effects that may result from treatment, such as the impact on patient quality of life.

* Are the benefits worth the harms and costs?

The final clinical decision should be made with the conviction that the benefits outweigh the potential risks and costs, always taking into account the individual characteristics and preferences of each particular patient.

In conclusion, RCTs and SRs can provide the highest-quality evidence for clinical decision-making in everyday clinical practice. However, not all published studies yield reliable and relevant information to answer our clinical questions. Therefore, it is essential to always analyze the validity, importance and applicability of the results before applying the conclusions of any study to patient care.

Appendix A. Working Group on Evidence-Based Pediatrics (rest of members)

Pilar Aizpurua Galdeano, María Aparicio Rodrigo, Nieves Balado Insunza, Albert Balaguer Santamaría, Carolina Blanco Rodríguez, Laura Cabrera Morente, Fernando Carvajal Encina, Eduardo Cuestas Montañés, M. Jesús Esparza Olcina, Sergio Flores Villar, Garazi Frailea Astorga, Paz González Rodríguez, Rafael Martín Masot, M. Victoria Martínez Rubio, Manuel Molina Arias, Eduardo Ortega Páez, Carlos Ochoa Sangrador, Begoña Pérez-Moneo Agapito and M. José Rivero Martín.

Declaration of competing interest

The authors have no conflicts of interest to declare.

References

1. Programa de habilidades en lectura crítica español (CASPe) [Internet]. Alicante: CASPe: 2022 [Accessed 11 February 2025]. Available from: <https://redcaspe.org/materiales/>.
2. Gonzalez de Dios J, Albi Rodríguez MS. Lectura crítica de artículos científicos sobre intervenciones terapéuticas... en búsqueda del número necesario de pacientes a tratar. In: *Medicina basada en la evidencia. Lo que siempre quise saber sobre la evidencia aplicada a la práctica clínica sin morir en el intento*. Madrid: Lúa Ediciones; 2024. p. 421–32.
3. Molina Arias M, Ochoa Sangrador C. Ensayo clínico (I). Definición. Tipos. Estudios cuasiexperimentales. *Evid Pediatr*. 2014;10:52.
4. Molina Arias M, Ochoa Sangrador C. Ensayo clínico (II). Fases. Aspectos legales. *Evid Pediatr*. 2014;10:71.

5. Molina Arias M, Ochoa Sangrador C. Ensayo clínico (III). Aleatorización. Enmascaramiento. *Evid Pediatr.* 2015;11:15.
6. Molina Arias M, Ochoa Sangrador C. Ensayo clínico (IV). Resultados. Variables y medidas de efecto e impacto. *Evid Pediatr.* 2015;11:33.
7. Molina Arias M, Ochoa Sangrador C. Ensayo clínico (V). Estrategias de análisis. Pérdidas de información. Análisis por intención de tratar. *Evid Pediatr.* 2015;11:52.
8. González de Dios J, Ruiz-Canela Cáceres J, Balaguer Santamaría A. Lectura crítica de artículos de revisión sistemática y metanálisis. In: *Medicina basada en la evidencia. Lo que siempre quiso saber sobre la evidencia aplicada a la práctica clínica sin morir en el intento.* Madrid: Lúa Ediciones; 2024. p. 461–9.
9. Available from: Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, Welch VA, editors. *Cochrane Handbook for Systematic Reviews of Interventions* version 6.5 (updated August 2024). Cochrane; 2024. www.training.cochrane.org/handbook
10. Moher D, Liberati A, Tetzlaff J, Altman DG, PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *Ann Intern Med.* 2009;151:264–9. W64.
11. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ.* 2021;372:n71.
12. Shea BJ, Reeves BC, Wells G, Thuku M, Hamel C, Moran J, et al. AMSTAR 2: a critical appraisal tool for systematic reviews that include randomised or non-randomised studies of healthcare interventions, or both. *BMJ.* 2017;358:j4008.