Journal of Clinical Epidemiology 190 (2026) 112083

ORIGINAL RESEARCH

Approaches for reporting and interpreting statistically nonsignificant findings in evidence syntheses: a systematic review

Amin Sharifan[a,*], Andreea Dobrescu[a], Curtis Harrod[b], Irma Klerings[a], Ariel Yuhan Ong[c,d], Etienne Ngeh[e], Yu-Tian Xiao[f], Gerald Gartlehner[a,g]

[a]*Department for Evidence-based Medicine and Evaluation, University for Continuing Education Krems, Krems, Austria*
[b]*American College of Physicians, Philadelphia, PA, USA*
[c]*Institute of Ophthalmology, University College London, London, UK*
[d]*Oxford Eye Hospital, Oxford University Hospitals NHS Foundation Trust, Oxford, UK*
[e]*School of Health and Society, University of Salford, Salford, UK*
[f]*Center for Cancer Research, Medical University of Vienna and Comprehensive Cancer Center, Vienna, Austria*
[g]*Center for Public Health Methods, RTI International, Research Triangle Park, NC, USA*

Accepted 17 November 2025; Published online 21 November 2025

Abstract

Objectives: To systematically review approaches for reporting and interpreting statistically nonsignificant findings with clinical relevance in evidence synthesis and to assess their methodological quality and the extent of their empirical validation.

Study Design and Setting: We searched Ovid MEDLINE ALL, Scopus, PsycInfo, Library of Guidance for Health Scientists, and MathSciNet for published studies in English from January 1, 2000, to January 30, 2025, for (1) best practices in guidance documents for evidence synthesis when interpreting clinically relevant nonsignificant findings, (2) statistical methods to support the interpretation, and (3) reporting practices. To identify relevant reporting guidelines, we also searched the Enhancing the QUAlity and Transparency Of health Research Network. The quality assessment applied the Mixed Methods Appraisal Tool, Appraisal tool for Cross-Sectional Studies, and checklists for expert opinion and systematic reviews from the Joanna Briggs Institute. At least two reviewers independently conducted all procedures, and a large language model facilitated data extraction and quality appraisal.

Results: Of the 5332 records, 37 were eligible for inclusion. Of these, 15 were editorials or opinion pieces, nine addressed methods, eight were cross-sectional or mixed-methods studies, four were journal guidance documents, and one was a systematic review. Twenty-seven records met the quality criteria of the appraisal tool relevant to their study design or publication type, while 10 records, comprising one systematic review, two editorials or opinion pieces, and seven cross-sectional studies, did not. Relevant methodological approaches to evidence synthesis included utilization of uncertainty intervals and their integration with various statistical measures (15 of 37, 41%), Bayes factors (six of 37, 16%), likelihood ratios (three of 37, 8%), effect conversion measures (two of 37, 5%), equivalence testing (two of 37, 5%), modified Fisher's test (one of 37, 3%), and reverse fragility index (one of 37, 3%). Reporting practices included problematic ''null acceptance'' language (14 of 37, 38%), with some records discouraging the inappropriate claim of no effect based on nonsignificant findings (nine of 37, 24%). None of the proposed methods were empirically tested with interest holders.

Conclusion: Although various approaches have been proposed to improve the presentation and interpretation of statistically nonsignificant findings, a widely accepted consensus has not emerged, as these approaches have yet to be systematically tested for their practicality and validity. This review provides a comprehensive review of available methodological approaches spanning both the frequentist and Bayesian statistical frameworks and identifies critical gaps in empirical validation of some approaches, namely the lack of thresholds to guide the interpretation of results. These findings highlight the need for systematic testing of proposed methods with interest holders and the development of evidence-based guidance to support appropriate interpretation of nonsignificant results in evidence synthesis.

Keywords: Evidence-based practice; Statistical data interpretation; Probability; Research design; Scholarly communication; Meta-research

Funding: This research was partly supported by internal funding from the Department for Evidence-Based Medicine and Evaluation at the University for Continuing Education Krems. In addition, the research was funded by the Gesellschaft für Forschungsförderung Niederösterreich (GFF) under the call RTI-Dissertations 2024. The content does not necessarily reflect the views of the Province of Lower Austria or Gesellschaft für Forschungsförderung Niederösterreich as the funding agency. Neither the Province of Lower Austria nor the funding agency can therefore be held responsible for the content.

Registration: CRD42025644578.

* Corresponding author. Department for Evidence-Based Medicine and Evaluation, University for Continuing Education Krems, Dr. Karl Dorrek-strasse 30, Krems an der Donau 3500, Lower Austria, Austria.

E-mail address: amin.sharifan@donau-uni.ac.at (A. Sharifan).

https://doi.org/10.1016/j.jclinepi.2025.112083

Plain Language Summary

This review looked at how to best report results that are not statistically significant because some of these findings can still be important to inform clinical care or health policy. We searched databases for studies published between 2000 and 2025. Out of more than 5000 records, 37 studies were relevant. These studies showed that there is no single best way to report nonsignificant findings.

1. Introduction

Uncertainty is inherent in medicine [1−4]. Statistical methods and methodological frameworks help quantify and interpret this uncertainty in clinical decision-making and medical research [5−7]. Among various statistical frameworks, the frequentist approach predominates the field [8], often leading researchers to interpret findings through a binary lens of statistical significance [9]. This approach has faced criticism for oversimplifying complex data and obscuring the clinical relevance of research findings [10]. Consequently, statistically significant (hereafter referred to as significant) findings may be overemphasized, regardless of their clinical importance, whereas statistically nonsignificant (hereafter referred to as nonsignificant) findings are disregarded, even when they have a signal of clinical relevance. This issue is particularly problematic in studies with rare events, small sample sizes, or high participant variability, where nonsignificant findings are common [11]. Selective reporting further aggravates such problems by favoring significant over nonsignificant findings [12−15]. This approach distorts the scientific literature through publication and data availability biases that may influence clinical guidelines, health-care policies, and research directions.

The challenges in interpreting nonsignificant findings extend beyond primary research into evidence synthesis and meta-analysis [16,17]. When findings are categorized merely as significant or not, readers can misinterpret findings in ways similar to those in primary studies [18]. The complexity of interpreting nonsignificant findings often leads to their undervaluation in evidence synthesis [19], risking a skewed understanding of evidence and undermining appropriate application to clinical practice. Furthermore, the interpretation and reporting of nonsignificant findings pose additional challenges by potentially introducing a distorted presentation of findings, which can mislead readers [20].

Even though nonsignificant findings may hold clinically important insights [21], no studies have comprehensively examined how they should be interpreted and reported in clinical medicine. This work aimed to systematically review approaches for reporting and interpreting statistically nonsignificant findings with clinical relevance in evidence synthesis and to assess their methodological quality and the extent of their empirical validation.

2. Methods

We sought to answer three key questions (KQ):

- KQ1: What are the current best practices and rationales for interpreting nonsignificant findings that suggest potential clinical effects in health-care research, including reporting considerations?
- KQ2: What statistical approaches are recommended for evaluating whether signals of clinical relevance warrant reporting despite not being significant?
- KQ3: What specific terminology and stylistic conventions are recommended for accurately and transparently communicating nonsignificant findings with potential clinical relevance?

We also conceptualized the following contextual questions (CQ):

- CQ1: What recommendations exist for reporting nonsignificant findings with a signal of clinical effect in the handbooks and guidance documents from reputable organizations or institutions?
- CQ2: What specific vocabulary, phrases, and stylistic guidelines do organizations or institutions suggest for accurately and transparently communicating nonsignificant findings that are clinically meaningful?

We conducted this review following the Cochrane guidelines for methodology reviews [22] and prospectively registered it with the International Prospective Register of Systematic Reviews (CRD42025644578). This report is in line with the Preferred Reporting Items for Systematic Reviews and Meta-analyses (PRISMA) 2020 statement [23].

2.1. Eligibility criteria

The inclusion criteria followed the Studies, Data, Methods, and Outcomes (SDMO) framework (Table S1) [24]. We included documents that guided the interpretation and reporting of nonsignificant findings in health care.

2.2. Information sources

We searched MEDLINE ALL (Ovid), Scopus (Elsevier), PsycInfo (Ebsco), Library of Guidance for Health Scientists (LIGHTS) [25], and MathSciNet (American Mathematical Society) to identify publications between January 1, 2000,

What is new?

Key findings

- Best practices have not been established for interpreting and reporting nonsignificant findings.
- The extent to which proposed methodologies for the interpretation of nonsignificant findings have been adopted in evidence synthesis is unclear.

What this adds to what is known?

- Several statistical approaches can help with the interpretation of such findings, but they expose a critical disconnect between methodological innovation and real-world adoption across the research community.
- There are knowledge gaps in the empirical validation of methods for interpreting nonsignificant findings and the absence of standardized interpretive thresholds, which highlight priority areas for methodological research.

What is the implication and what should change now?

- We advocate the need for coordinated efforts between evidence synthesis software developers, journal editors, and institutions involved in evidence synthesis to bridge the persistent evidence-practice gap in biomedical research.

and January 30, 2025 (see Supplementary Material for search strategies). This timeframe captured recent advancements in the field following the highlight of *P* value shortcomings and a call for evidence-based medical statistics [26]. A reviewer familiar with the topic (AS) developed the initial search strategy, which an information specialist (IK) refined. An information specialist reviewed the strategy using the Peer Review of Electronic Search Strategies (PRESS) guidelines [27]. We used the Deduplicator tool [28], and then EndNote, to remove duplicate search results, and we uploaded them to Covidence for screening.

We also searched the Enhancing the QUAlity and Transparency Of health Research Network (EQUATOR) network up to February 6, 2025, for relevant reporting guidelines by choosing ''statistical analysis plan-whole report'' and ''statistical methods and analyses'' under the filter ''section of the report'' within the EQUATOR library.

2.3. Selection process

The screening team (YTX, EN, AYO) completed calibration to standardize procedures for title/abstract screening (30 records) and full-text screening (five records) [29]. One reviewer (AS) developed the calibration, which a senior reviewer (GG) validated for accuracy. Following the pilots, four independent reviewers (AS, YTX, EN, AYO) screened the records. A senior reviewer (GG) resolved conflicts.

2.4. Data collection process

We extracted concepts aligned with our KQ, including proposed statistical methods for interpreting nonsignificant findings, approaches to reporting such findings, barriers associated with each method, and contextual applicability. One reviewer (AS) extracted data and verified accuracy using Claude 3.7 Sonnet configured in extended-thinking mode with a formal style. The choice of including artificial intelligence was driven by enhancing data extraction efficiency [30–32]. Two other reviewers (AYO, EN) cross-checked the large language model's output. There were no missing or unclear information issues during data extraction.

2.5. Methodological quality appraisal

We tailored our quality appraisal process to accommodate the diverse publications included in this review and prioritized tools designed to assess methodological quality rather than to judge risk of bias [33] since most did not have an experimental or epidemiological design. For cross-sectional studies, we used the Appraisal tool for Cross-Sectional Studies (AXIS) tool [34]; for opinion-based articles or editorials, we applied the Joanna Briggs Institute's (JBI) checklist for expert opinion [35]; for mixed-methods analyses, we used the Mixed Methods Appraisal Tool [36]; and for systematic reviews and research syntheses, we relied on the JBI checklist for systematic reviews [37]. We selected these tools based on their capacity to address methodological nuances specific to each study design. Risk of bias due to missing results and certainty of evidence evaluation did not apply to this study.

2.6. Data analysis and synthesis

Due to the inherent lack of quantifiable data within this methodological topic, we synthesized findings descriptively. We organized data using Microsoft Excel (version 2408, Microsoft Corp., Redmond, WA, USA), conducted the analyses using R (version 4.4.2, R Core Team, Vienna, Austria), and created the flow diagram with the PRISMA2020 Shiny app [38].

3. Results

3.1. Study characteristics

Of the 5332 records identified, 37 were included in the analysis (Fig). Among these, 15 were editorials or opinion articles [39–53], nine were method articles [54–62], seven were cross-sectional studies [63–69], four were journal guideline documents [70–73], one was a mixed-method

study [74], and one was a systematic review [75]. Details regarding excluded records in the full-text phase are available in Table S2.

Of the included studies, 27 (73%) met the quality criteria of the appraisal tool relevant to their study design or publication type [39−50,52−62,70,71,73,74], while 10 (27%) did not meet these criteria [51,63−69,72,75]. Among those that did not meet quality criteria, seven were cross-sectional studies, two were editorials or opinion pieces, and one was a systematic review. The most common (five of seven, 71%) methodological limitations among cross-sectional studies that failed to meet quality criteria included inadequate justification of sample size. For editorials and opinion pieces, limitations primarily related to insufficient acknowledgment of evidence sources. The systematic review that did not meet quality criteria lacked transparency in risk of bias assessment procedures. Despite these methodological shortcomings, we included all eligible studies to provide a comprehensive overview of the existing literature on this topic, while acknowledging that the quality of evidence supporting different approaches varies.

Other than editorials, opinion pieces, method articles, and journal recommendation papers, a small proportion of publications adhered to their respective reporting guidelines (three of 10, 30%) [66,74,75], namely the STrengthening the Reporting of OBservational studies in Epidemiology statement for cross-sectional studies (STROBE) and only one reported having a protocol (one of 10, 10%) [63]. No articles incorporated perspectives from interest holders, comprising individuals with interests in health-care issues [76]. Furthermore, the articles addressed nonsignificance in a range of study designs, including experimental and observational studies and systematic reviews. Among these, randomized controlled trials were the most frequently discussed (11 of 37, 30%) [41,45,49,57,60,62,65,66,68,69,75]. Further details regarding the characteristics of eligible publications and quality assessment can be found in Tables 1, S3−S8.

3.2. *KQ1: best practices for interpreting nonsignificant findings*

We did not identify any eligible publications that addressed best practices for interpreting nonsignificant findings.

3.3. *KQ2: approaches for interpreting nonsignificant findings*

Publications have emphasized that nonsignificant findings require careful interpretation to avoid misinforming research and clinical decisions. Key arguments raised concerns about the limitations of hypothesis testing in

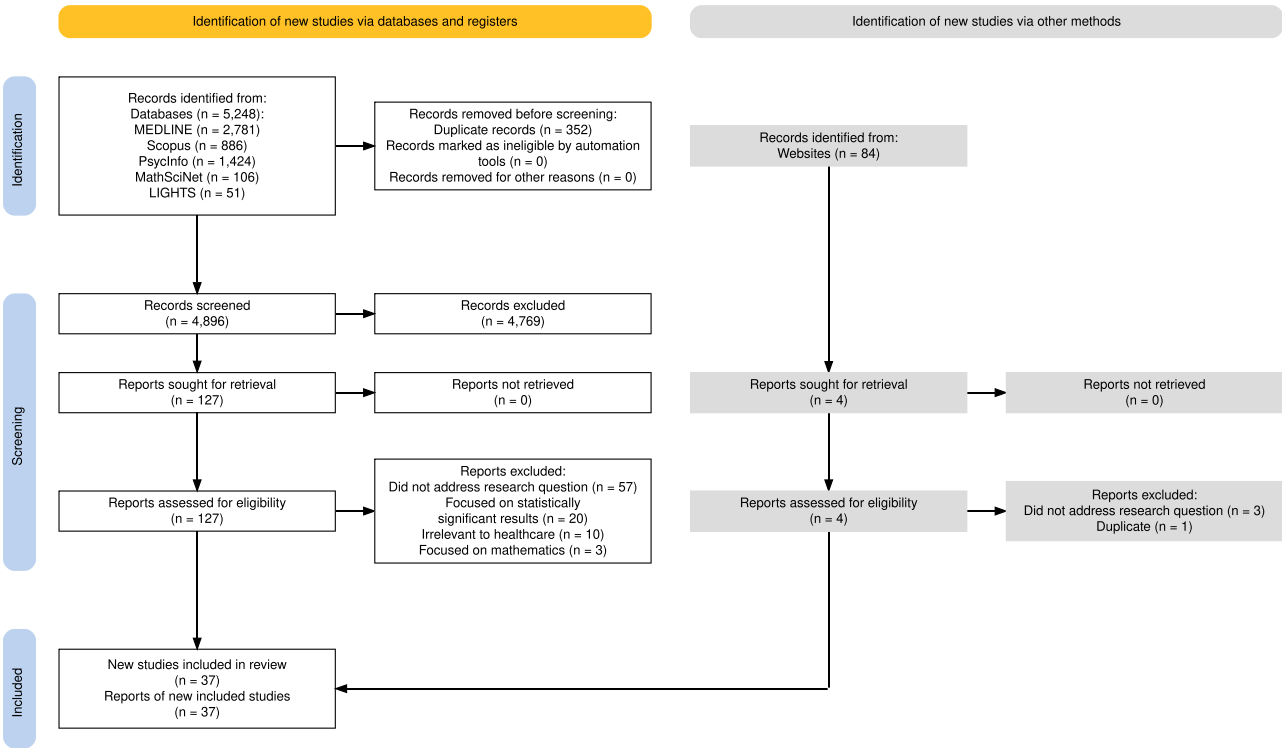

Figure 1. PRISMA Flow diagram. The website represents records identified from the EQUATOR Network.EQUATOR, Enhancing the Quality and Transparency of Health Research; LIGHTS, Library of guidance for health scientists; PRISMA, preferred reporting items for systematic reviews and meta-analyses.

Table 1. Study characteristics

Category	Characteristics	n/N (%)
Article type	Editorials and opinions	15/37[a] (41)
	Methods	9/37 (24)
	Cross-sectional and mixed methods	8/37 (22)
	Journal guide for authors	4/37 (11)
	Systematic review	1/37 (2)
Clinical area	General medicine	18/37 (49)
	Psychology	3/37 (8)
	Oncology	2/37 (5)
	Anesthesiology	2/37 (5)
	Cardiology	1/37 (2)
	Evidence-based medicine	1/37 (2)
	Biostatistics	1/37 (2)
	Gerontology	1/37 (2)
	Mental health	1/37 (2)
	Nicotine and tobacco research	1/37 (2)
	Nursing	1/37 (2)
	Orthopedic medicine	1/37 (2)
	Reproductive health	1/37 (2)
	Spinal cord medicine	1/37 (2)
	Sport medicine	1/37 (2)
	Urology	1/37 (2)
Funding	Not reported	21/37 (57)
	Government	9/37 (24)
	None	5/37 (14)
	Nonprofit organization	2/37 (5)
Geographic contributions	Europe	10/37 (27)
	North America	10/37 (27)
	Europe and North America	8/37 (22)
	Oceania	5/37 (14)
	Not reported	2/37 (5)
	Europe and Oceania	1/37 (2)
	Europe, North America, and South America	1/37 (2)
Quality assessment questions	Addressed	280/324[b] (86)
	Not addressed	14/324 (4)
	Unclear	2/324 (1)
	Not applicable	28/324 (8)

[a] The denominator represents the number of identified studies.
[b] The denominator represents the number of questions.

supporting the null hypothesis [50], cautioning against interpreting nonsignificant findings as proof of no effect [39,40,42,45,57,58], and criticizing the dichotomization of findings by statistical significance, which can overlook clinically important findings [48]. Additionally, some publications criticized the practice of labeling nonsignificant findings as ‘‘negative’’ or ‘‘useless.’’ [49,61] Given these challenges, the following approaches were proposed for the proper interpretation of nonsignificant findings.

3.3.1. Utilizing uncertainty intervals

Focusing on confidence interval interpretation was the most frequently recommended method for KQ2 (15/37, 41%), either alone (eight of 37, 22%) [39,44,45,48,52,71,72,75] or in combination with other approaches, including the use of a minimal clinically important difference (three of 37, 8%) [40,47,57], effect sizes (two of 37, 5%) [50,67], reverse Bayesian methods (one of 37, 3%) [59], or area under the curve (one of 37, 3%) [74]. Publications consistently endorsed confidence intervals but provided limited threshold guidance beyond prespecified minimal clinically important differences, which reflects that the smallest difference considered important from a clinical perspective. One source emphasized interpreting both the point estimates and the bounds of intervals [39], while another elaborated on the notion that the true effect is more likely around the point estimate as opposed to interval limits [45]. In the latter scenario, the line of no effect and the minimal clinically important difference may be used as the thresholds [40,42,48,75]

3.3.2. Bayesian methods

Bayesian methods and Bayes factors (BFs) were recommended for interpreting nonsignificant findings (six of 37, 16%) and demonstrated greater standardization compared to others [48,55,56,58,63,65].

3.3.3. Likelihood ratios

Three publications recommended likelihood ratios [49,60,69]. This approach quantifies the relative strength of evidence between competing hypotheses and may be preferable when uniform prior distributions render BFs inapplicable or undefined, as likelihood ratios are not bound to priors. The interpretation follows the thresholds corresponding to BFs [69].

3.3.4. Equivalence testing

Two publications recommended equivalence testing using predefined margins based on the minimal clinically meaningful differences [56,58]. The approach employs two one-sided tests or 90% confidence intervals to establish whether effects fall within equivalence bounds. However, acquiring evidence-based equivalence margins can be challenging, limiting the practical utility of evidence synthesis. Additionally, the available tool for equivalence testing in meta-analysis requires standardized mean difference metrics [77], which necessitates converting various statistical measures under the assumption of normally distributed data in evidence synthesis.

3.3.5. Effect size conversion methods

Absolute risk reduction and numbers needed to treat or harm were recommended as intuitive measures for dichotomous outcomes (two of 37, 5%) [51,61]. However, the latter might be perceived as unsuitable for nonsignificant findings or those with zero difference between groups, as such data fail to yield clinically meaningful values [51]. Numbers needed to treat or harm can be obtained from either absolute or relative effects, with the latter providing a more constant effect on various baseline risks across primary studies [61]. Both methods must be implemented on pooled effect estimates in meta-analyses.

3.3.6. Modified Fisher's test

One publication proposed a method for identifying potential false negatives among sets of nonsignificant findings [64]. The approach aggregates nonsignificant *P* values; if the resulting *P* value is lower than the threshold of 0.1, as proposed by the authors, at least one false negative is likely present. The method's reliability increases with larger result sets, although it cannot identify specific false negatives [64]. Therefore, its application in studies with a low number of endpoints and evidence synthesis might be limited.

3.3.7. Reverse fragility index

One publication recommended assessing robustness by determining the minimum events needed to shift nonsignificant findings to significant [66]. This method is applicable to dichotomous outcomes only. While providing a fragility measure for nonsignificant findings can help evaluate their robustness, the practical utility is uncertain. No specific thresholds have been proposed for this method; however, a similar approach using the reciprocal interpretation of values applied to the fragility index of meta-analyses [78] could be adapted here, although its applicability remains to be tested.

It is worth noting that among the identified approaches for the proper interpretation of nonsignificant findings, two were tailored for randomized controlled trials [41,59]. One approach, which focuses on power calculations and risk of type II errors [41], is not directly applicable to evidence synthesis. The second, Analysis of Credibility, is based on reverse Bayes methods and requires adaptation to account for between-study heterogeneity when applied to meta-analyses [59]. Further details on interpretation approaches are provided in Table S9.

3.4. KQ3: proposed terminologies for reporting nonsignificant findings

Publications identified two categories of problematic terminology: "Null acceptance" statements inappropriately supporting the null hypothesis, namely "there was no difference between the groups" (14 of 37, 38%) [39,42,44,45,47–49,53,55–58,63,70], and "trend" statements suggesting interpretation of *P* values near significance thresholds using qualifying terminologies, eg, "the results trended toward significance" (10 of 37, 27%) [40,43,46,53,62,64,67,68,70,73]. Recommended alternatives included precise language acknowledging uncertainty: "Study results do not support a recommendation in favor of the intervention" (one of 37, 3%) [41], and describing effect ranges compatible with confidence intervals (one of 37, 3%) [39]. Bayesian approaches offered additional precision with phrases like "data are insensitive" for inconclusive evidence (one of 37, 3%) [55]. Several publications recommended incorporating contextual considerations beyond statistical measures, including costs, interest holders' perspectives, and potential harms (nine of 37, 24%) [39–41,44,56,57,61,65,75]. On the other hand, while some publications stressed the importance of correctly reporting a nonsignificant *P* value as "statistically nonsignificant" (one of 37, 3%) [44], others discouraged highlighting the quoted term and *P* values altogether, advocating instead for the use of alternative metrics, namely confidence intervals, to convey the results more meaningfully (two of 37, 5%) [39,57]. Further details on encouraged and discouraged reporting of nonsignificant results are available in Tables 2 and S10.

Table 2. Encouraged and discouraged language in the literature for reporting statistically nonsignificant findings when using the frequentist approach

Category	Characteristics
Encouraged phrasing	Results are most compatible with no important effect [39]
	We were unable to demonstrate a difference between the groups [70]
	Promising [52]
	Interesting hint [62]
	Suggestive [52]
Discouraged phrasing	No effect [39,42,45,48,53,55,56,58,63]
	Trend toward significance/nonsignificant trend/trend [40,43,46,53,62,67,68,70,73]
	Almost statistically significant [43,53,62,67]
	No difference [39,44,57,70]
	Approaching statistical significance [46,62,70]
	Ineffective [49,70]
	Marginal significance [64,68]
	Lack of evidence [49]
	Near statistical significance [40]
	No association [47]
	Borderline significance [40]

3.5. Exploratory questions

3.5.1. Ethical considerations

Some publications emphasized the ethical importance of reporting nonsignificant findings, highlighting how publication and reporting biases can undermine research integrity and lead to research waste (seven of 37, 19%) [40,53−56,63,67]. One publication highlighted the critical consequences of false negatives in diseases with high mortality or limited treatment options [66].

3.5.2. Barriers to adopting alternative statistical methods

Most publications suggested that the frequentist approach remains deeply ingrained in the field, creating barriers to adopting alternative statistical methods (14 of 37, 38%) [39,40,42,47−50,53,55,61,66−68,74]. They also noted that alternative statistical methods often require additional training and familiarity among researchers (five of 37, 14%) [48,56,58,60,63]. Nevertheless, most sources (28 of 37, 76%) showed unclear adoption status, suggesting limited systematic implementation across the biomedical literature. Further details are available in Table S11.

3.6. Contextual questions

Major institutions involved with clinical trials or evidence syntheses provided minimal to no guidance on interpreting and reporting nonsignificant findings (Table S12).

4. Discussion

4.1. Overview

This systematic review identified multiple approaches for interpreting and reporting nonsignificant findings, but no single evidence-based approach with broad agreement among researchers, statisticians, methodologists, and journal editors exists. While the presentation and interpretation of confidence intervals emerged as the most endorsed approach for incorporating nonsignificant results into evidence synthesis, substantial heterogeneity exists in implementation recommendations, and most methods lack empirical validation or clear adoption pathways.

4.2. Nonsignificant findings in meta-analysis and evidence synthesis

Meta-analysis constitutes the cornerstone of quantitative evidence synthesis and provides a framework for combining results from multiple methodologically similar individual studies with appropriate statistical heterogeneity [17,79]. The output of a meta-analysis is a pooled effect estimate along with associated measures of precision, such as confidence intervals in frequentist frameworks or credible intervals in Bayesian approaches [80]. When pooled effect estimates are nonsignificant, the interpretation becomes challenging. The Cochrane Handbook emphasizes interpreting the magnitude and precision of pooled effects rather than relying on binary significance testing but lacks concrete guidance for authors [81]. Importantly, nonsignificant results in evidence synthesis reflect uncertainty rather than the absence of effect [82]. Prediction intervals, which capture effect size variability [83], provide additional context for interpreting nonsignificant pooled estimates by illustrating the scope of uncertainty in the evidence base. Despite these established principles, our review demonstrates that standardized implementation approaches for these interpretive principles in evidence synthesis remain lacking.

4.3. Clinical and research implications

The variety of approaches to interpreting nonsignificant findings highlights the absence of a single best practice and reflects the complexity of the problem. Misinterpretation of nonsignificant findings, whether through inappropriate acceptance of the null hypothesis or misleading trend language, can lead to suboptimal clinical decisions and research waste. The common recommendation to focus on the interpretation of confidence intervals suggests emerging consensus; however, researchers sometimes misinterpret confidence intervals similar to *P* values, incorrectly claiming ‘‘no difference’’ when intervals include the value of no effect rather than acknowledging the full range of plausible values [84].

In response to this issue, recent proposals for decision threshold guidance in interpreting the results in evidence synthesis may enhance the practical implementation of confidence intervals [85,86]. On the other hand, Bayesian approaches, with the established threshold classification schemes [87], offer a promising standardized framework that could facilitate adoption in clinical guideline development [88]. Multiple sources referenced Jeffreys' classification scheme for the BF_{10}, which quantifies the relative evidence for the alternative hypothesis (H_1) vs the null hypothesis (H_0). In this scheme, BF_{10} values between 1/3 and 3 are considered anecdotal evidence and are generally regarded as barely worth mentioning. Moderate or substantial evidence in favor of the alternative hypothesis is indicated by values between 3 and 10, while values from 10 to 30 suggest strong evidence. Values between 30 and 100 are interpreted as very strong evidence, and those exceeding 100 provide decisive support for the alternative hypothesis. The same strength of evidence in favor of the null hypothesis is given by the reciprocal of these thresholds (ie, a BF_{10} below 1/10 for strong evidence, below 1/30 for very strong, and below 1/100 for decisive evidence) [55,58,63]. However, the additional statistical expertise required may constrain widespread implementation. Alternatively, methods such as the reverse fragility index may help decision makers more objectively assess the imprecision of findings.

These considerations are particularly relevant to evidence synthesis involving safety outcomes, where reporting concerns are well-documented [89−91]. Furthermore, systematic application of these approaches could help address publication bias in evidence syntheses across scientific disciplines [92,93].

4.4. Implementation in existing reporting guidelines

To translate these clinical and research insights into practice, it is important to consider how they align with and inform existing reporting guidelines. Current reporting guidelines, including PRISMA, provide limited direction for interpreting nonsignificant findings beyond basic effect size and confidence interval reporting. Our findings demonstrate that available guidance resources are insufficient, given the complexity of interpretation challenges identified in contemporary research.

Integration of methods that quantify evidence in favor of the null hypothesis, such as Bayesian inferences, represents a reasonable improvement for reporting standards concerned with testing hypotheses. These methodological advances warrant consideration in future updates to major reporting guidelines and their extensions, so researchers have more comprehensive frameworks for interpreting and communicating nonsignificant findings.

4.5. Limitations

Most identified approaches lacked empirical testing with relevant interest holders, particularly clinicians, patients, and health-care policymakers, who must ultimately interpret such findings. The unclear adoption status of most recommendations suggests limited real-world validation. Furthermore, our study focused on English-language literature published between 2000 and 2025, which may exclude non-English publications and earlier foundational works. While the date frame captures recent advancements, it risks overlooking historical perspectives. Finally, the adoption of methods proposed for primary studies in the context of evidence synthesis should be approached with caution, as various factors, namely heterogeneity between studies, must be considered.

5. Conclusion

This systematic review addresses a critical gap by comprehensively evaluating available approaches for interpreting nonsignificant findings in evidence synthesis. To further advance the field, it is important to move beyond replacing one statistical tool with another and, instead, embrace a more holistic approach to interpretation and reporting of findings. Rather than emphasizing distinctions between Bayesian and frequentist paradigms, integrating these frameworks enables more nuanced evidence synthesis, which could enhance clinical decision-making, though this requires empirical validation. Authors are encouraged to adopt the methods described in this article to contextualize their findings, and journals should be more receptive to these approaches, especially for findings with potential clinical relevance. Institutions involved in evidence synthesis may also consider incorporating the methodologies identified here into their platforms, ensuring a more comprehensive, evidence-based interpretation of available data for the scientific community. Nevertheless, systematic evaluation of implementation barriers and user comprehension will be essential before broader institutional adoption of these methodologies. Finally, the formation of expert panels and guideline groups may facilitate the interpretation of findings for statistical methods without readily available thresholds.

Ethics statement

This study used only publicly available data from the scientific literature. As the research did not involve any participants or identifiable personal data, ethical approval was not applicable.

Patient and public involvement

This study did not engage patients or the public in its design, implementation, reporting, or dissemination.

Declaration of generative AI in scientific writing

During the preparation of this work, the authors used Claude Pro to quality-check the data extraction and quality appraisals. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

CRediT authorship contribution statement

Amin Sharifan: Writing − review & editing, Writing − original draft, Visualization, Software, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Andreea Dobrescu:** Writing − review & editing, Supervision, Conceptualization. **Curtis Harrod:** Writing − review & editing, Supervision, Conceptualization. **Irma Klerings:** Writing − review & editing, Investigation. **Ariel Yuhan Ong:** Writing − review & editing, Investigation. **Etienne Ngeh:** Writing − review & editing, Investigation, Conceptualization. **Yu-Tian Xiao:** Writing − review & editing, Investigation. **Gerald Gartlehner:** Writing − review & editing, Validation, Supervision, Methodology, Investigation, Formal analysis, Conceptualization.

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.

Acknowledgments

We would like to thank Chris Cooper for reviewing this work's search strategy.

Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.jclinepi.2025.112083.

Data availability

The authors confirm that the data supporting the findings of this study are available within the article and its supplementary materials.

References

[1] Simpkin AL, Schwartzstein RM. Tolerating uncertainty — the next medical revolution? N Engl J Med 2016;375(18):1713−5. https://doi.org/10.1056/NEJMp1606402.

[2] Dahm MR, Crock C. Understanding and communicating uncertainty in achieving diagnostic excellence. JAMA 2022;327(12):1127. https://doi.org/10.1001/jama.2022.2141.

[3] Helou MA, DiazGranados D, Ryan MS, Cyrus JW. Uncertainty in decision making in medicine: a scoping review and thematic analysis of conceptual models. Acad Med 2020;95(1):157−65. https://doi.org/10.1097/ACM.0000000000002902.

[4] Riley RD, Collins GS, Kirton L, Snell KI, Ensor J, Whittle R, et al. Uncertainty of risk estimates from clinical prediction models: rationale, challenges, and approaches. BMJ 2025;388:e080749. https://doi.org/10.1136/bmj-2024-080749.

[5] Guyatt G, Vandvik PO, Iorio A, Agarwal A, Yao L, Eachempati P, et al. Core GRADE 7: principles for moving from evidence to recommendations and decisions. BMJ 2025;389:e083867. https://doi.org/10.1136/bmj-2024-083867.

[6] Zhang S, Heck PR, Meyer MN, Chabris CF, Goldstein DG, Hofman JM. An illusion of predictability in scientific results: even experts confuse inferential uncertainty and outcome variability. Proc Natl Acad Sci USA 2023;120(33):e2302491120. https://doi.org/10.1073/pnas.2302491120.

[7] Tsaneva-Atanasova K, Pederzanil G, Laviola M. Decoding uncertainty for clinical decision-making. Philos Trans A Math Phys Eng Sci 2025;383(2292):20240207. https://doi.org/10.1098/rsta.2024.0207.

[8] Goligher EC, Heath A, Harhay MO. Bayesian statistics for clinical research. Lancet 2024;404(10457):1067−76. https://doi.org/10.1016/S0140-6736(24)01295-9.

[9] van Zwet E, Gelman A, Greenland S, Imbens G, Schwab S, Goodman SN. A new look at P values for randomized clinical trials. NEJM Evid 2023;3(1):EVIDoa2300003. https://doi.org/10.1056/EVIDoa2300003.

[10] Wasserstein RL, Lazar NA. The ASA statement on P-values: context, process, and purpose. Am Stat 2016;70(2):129−33. https://doi.org/10.1080/00031305.2016.1154108.

[11] Savitz DA, Wise LA, Bond JC, Hatch EE, Ncube CN, Wesselink AK, et al. Responding to reviewers and editors about statistical significance testing. Ann Intern Med 2024;177(3):385−6. https://doi.org/10.7326/M23-2430.

[12] Fanelli D, Costas R, Ioannidis JPA. Meta-assessment of bias in science. Proc Natl Acad Sci USA 2017;114(14):3714−9. https://doi.org/10.1073/pnas.1618569114.

[13] Kicinski M. How does under-reporting of negative and inconclusive results affect the false-positive rate in meta-analysis? A simulation study. BMJ Open 2014;4(8):e004831. https://doi.org/10.1136/bmjopen-2014-004831.

[14] Showell MG, Cole S, Clarke MJ, DeVito NJ, Farquhar C, Jordan V. Time to publication for results of clinical trials. Cochrane Database Syst Rev 2024;11(11):MR000011. https://doi.org/10.1002/14651858.MR000011.pub3.

[15] Guyatt G, Wang Y, Eachempati P, Iorio A, Murad MH, Hultcrantz M, et al. Core GRADE 4: rating certainty of evidence—risk of bias, publication bias, and reasons for rating up certainty. BMJ 2025;389:e083864. https://doi.org/10.1136/bmj-2024-083864.

[16] Gates S, Ealing E. Reporting and interpretation of results from clinical trials that did not claim a treatment difference: survey of four general medical journals. BMJ Open 2019;9(9):e024785. https://doi.org/10.1136/bmjopen-2018-024785.

[17] 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.

[18] Boutron I, Ravaud P. Misrepresentation and distortion of research in biomedical literature. Proc Natl Acad Sci USA 2018;115(11):2613−9. https://doi.org/10.1073/pnas.1710755115.

[19] Ahmed I, Sutton AJ, Riley RD. Assessment of publication bias, selection bias, and unavailable data in meta-analyses using individual participant data: a database survey. BMJ 2012;344:d7762. https://doi.org/10.1136/bmj.d7762.

[20] Boutron I, Dutton S, Ravaud P, Altman DG. Reporting and interpretation of randomized controlled trials with statistically nonsignificant results for primary outcomes. JAMA 2010;303(20):2058−64. https://doi.org/10.1001/jama.2010.651.

[21] Young PJ, Nickson CP, Perner A. When should clinicians act on non−statistically significant results from clinical trials? JAMA 2020;323(22):2256−7. https://doi.org/10.1001/jama.2020.3508.

[22] Clarke M. Guide to the contents of a Cochrane Methodology protocol and review. Cochrane; 2020. Available at: https://methodology.cochrane.org/sites/methodology.cochrane.org/files/uploads/guide_to_the_contents_of_a_cochrane_methodology_protocol_and_review.pdf. Accessed July 29, 2025.

[23] 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. https://doi.org/10.1136/bmj.n71.

[24] 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.

[25] Hirt J, Schönenberger CM, Ewald H, Lawson DO, Papola D, Rohner R, et al. Introducing the library of guidance for Health Scientists (LIGHTS): a living database for methods guidance. JAMA Netw Open 2023;6(2):e2253198. https://doi.org/10.1001/jamanetworkopen.2022.53198.

[26] Goodman SN. Toward evidence-based medical statistics. 1: the P value fallacy. Ann Intern Med 1999;130(12):995−1004. https://doi.org/10.7326/0003-4819-130-12-199906150-00008.

[27] McGowan J, Sampson M, Salzwedel DM, Cogo E, Foerster V, Lefebvre C. PRESS peer review of electronic search strategies: 2015 guideline statement. J Clin Epidemiol 2016;75:40−6. https://doi.org/10.1016/j.jclinepi.2016.01.021.

[28] Forbes C, Greenwood H, Carter M, Clark J. Automation of duplicate record detection for systematic reviews: deduplicator. Syst Rev 2024;13(1):206. https://doi.org/10.1186/s13643-024-02619-9.

[29] Polanin JR, Pigott TD, Espelage DL, Grotpeter JK. Best practice guidelines for abstract screening large-evidence systematic reviews and meta-analyses. Res Synth Methods 2019;10(3):330−42. https://doi.org/10.1002/jrsm.1354.

[30] Siemens W, von Elm E, Binder H, Böhringer D, Eisele-Metzger A, Gartlehner G, et al. Opportunities, challenges and risks of using artificial intelligence for evidence synthesis. BMJ Evid Based Med 2025;30:381−4. https://doi.org/10.1136/bmjebm-2024-113320.

[31] Gartlehner G, Kahwati L, Hilscher R, Thomas I, Kugley S, Crotty K, et al. Data extraction for evidence synthesis using a large language model: a proof-of-concept study. Res Synth Methods 2024;15(4):576−89. https://doi.org/10.1002/jrsm.1710.

[32] Polak MP, Morgan D. Extracting accurate materials data from research papers with conversational language models and prompt engineering. Nat Commun 2024;15(1):1569. https://doi.org/10.1038/s41467-024-45914-8.

[33] Stone JC, Barker TH, Aromataris E, Ritskes-Hoitinga M, Sears K, Klugar M, et al. From critical appraisal to risk of bias assessment: clarifying the terminology for study evaluation in JBI systematic reviews. JBI Evid Synth 2023;21(3):472−7. https://doi.org/10.11124/JBIES-22-00434.

[34] Downes MJ, Brennan ML, Williams HC, Dean RS. Development of a critical appraisal tool to assess the quality of cross-sectional studies (AXIS). BMJ Open 2016;6(12):e011458. https://doi.org/10.1136/bmjopen-2016-011458.

[35] Pearson A, Jordan Z, McArthur A, Florescu S, Cooper A, Yan H, et al. Systematic reviews of textual evidence: narrative, expert opinion or policy. In: Aromataris E, Lockwood C, Porritt K, Pilla B, Jordan Z, editors. JBI manual for evidence synthesis. JBI; Available at: https://synthesismanual.jbi.global. https://doi.org/10.46658/JBIMES-24-04. Accessed December 5, 2025.

[36] Hong QN, Fàbregues S, Bartlett G, et al. The Mixed Methods Appraisal Tool (MMAT) version 2018 for information professionals and researchers. Educ Inf 2018;34(4):285−91. https://doi.org/10.3233/EFI-180221.

[37] Aromataris E, Fernandez R, Godfrey CM, Holly C, Khalil H, Tungpunkom P. Summarizing systematic reviews: methodological development, conduct and reporting of an umbrella review approach. JBI Evid Implement 2015;13(3):132−40. https://doi.org/10.1097/XEB.0000000000000055.

[38] Haddaway NR, Page MJ, Pritchard CC, McGuinness LA. PRISMA2020: an R package and shiny app for producing PRISMA 2020-compliant flow diagrams, with interactivity for optimised digital transparency and open synthesis. Campbell Syst Rev 2022;18(2):e1230. https://doi.org/10.1002/cl2.1230.

[39] Amrhein V, Greenland S, McShane B. Scientists rise up against statistical significance. Nature 2019;567(7748):305−7. https://doi.org/10.1038/d41586-019-00857-9.

[40] Ciapponi A, Belizán JM, Piaggio G, Yaya S. There is life beyond the statistical significance. Reprod Health 2021;18(1):80. https://doi.org/10.1186/s12978-021-01131-w.

[41] Derr J, Goldsmith LJ. How to report nonsignificant results planning to make the best use of statistical power calculations. J Orthop Sports Phys Ther 2003;33(6):303−6. https://doi.org/10.2519/jospt.2003.33.6.303.

[42] Dushoff J, Kain MP, Bolker BM. I can see clearly now: reinterpreting statistical significance. Methods Ecol Evol 2019;10(6):756−9. https://doi.org/10.1111/2041-210X.13159.

[43] Gibbs NM, Gibbs SV. Misuse of ‘‘trend’’ to describe ‘‘almost significant’’ differences in anaesthesia research. Br J Anaesth 2015;115(3):337−9. https://doi.org/10.1093/bja/aev149.

[44] Gibbs NM. Errors in the interpretation of ‘‘No Statistically Significant Difference.’’. Anaesth Intensive Care 2013;41(2):151−3. https://doi.org/10.1177/0310057X1304100203.

[45] Hackshaw A, Kirkwood A. Interpreting and reporting clinical trials with results of borderline significance. BMJ 2011;343:d3340. https://doi.org/10.1136/bmj.d3340.

[46] Harvey LA. Nearly significant if only…. Spinal Cord 2018;56(11):1017. https://doi.org/10.1038/s41393-018-0214-8.

[47] Hawkins AT, Samuels LR. Use of confidence intervals in interpreting nonstatistically significant results. JAMA 2021;326(20):2068−9. https://doi.org/10.1001/jama.2021.16172.

[48] Hopkins WG. Replacing statistical significance and non-significance with better approaches to sampling uncertainty. Front Physiol 2022;13:962132. https://doi.org/10.3389/fphys.2022.962132.

[49] Khan MS, Jamil A, Januzzi JL, Shakoor M, Bennett MM, Vanzyl JS, et al. POINT: statistical non-significance, likelihood ratio, and the interpretation of clinical trial evidence: insights from heart failure randomized trials. J Card Fail 2024;30(12):1629−32. https://doi.org/10.1016/j.cardfail.2024.07.026.

[50] Munafò MR, Wileyto EP. Guidelines on statistical reporting at nicotine & tobacco research. Nicotine Tob Res 2015;17(11):1295−6. https://doi.org/10.1093/ntr/ntv131.

[51] Muthu V. The number needed to treat: problemsdescribing non-significant results. Evid Based Ment Health 2003;6(3):72. https://doi.org/10.1136/ebmh.6.3.72.

[52] Steyerberg EW, Van Calster B. Redefining significance and reproducibility for medical research: a plea for higher P-value thresholds for diagnostic and prognostic models. Eur J Clin Invest 2020;50(5):e13229. https://doi.org/10.1111/eci.13229.

[53] Visentin DC, Cleary M, Hunt GE. The earnestness of being important: reporting non-significant statistical results. J Adv Nurs 2020;76(4):917−9. https://doi.org/10.1111/jan.14283.

[54] Cummins KM, Marks C. Farewell to bright-line: a guide to reporting quantitative results without the S-Word. Front Psychol 2020;11:815. https://doi.org/10.3389/fpsyg.2020.00815.

[55] Dienes Z. Using bayes to get the Most out of non-significant results. Front Psychol 2014;5. https://doi.org/10.3389/fpsyg.2014.00781.

[56] Harms C, Lakens D. Making ‘‘null effects’’ informative: statistical techniques and inferential frameworks. J Clin Transl Res 2018;3(Suppl 2):382−93. https://doi.org/10.18053/jctres.03.2017S2.007.

[57] Hemming K, Taljaard M. Why proper understanding of confidence intervals and statistical significance is important. Med J Aust 2021;214(3):116−8. https://doi.org/10.5694/mja2.50926.

[58] Lakens D, McLatchie N, Isager PM, Scheel AM, Dienes Z. Improving inferences about null effects with bayes factors and equivalence tests. J Gerontol B Psychol Sci Soc Sci 2020;75(1):45−57. https://doi.org/10.1093/geronb/gby065.

[59] Matthews RAJ. Beyond ‘‘significance’’: principles and practice of the analysis of credibility. R Soc Open Sci 2018;5(1):171047. https://doi.org/10.1098/rsos.171047.

[60] Perneger T. How to use likelihood ratios to interpret evidence from randomized trials. J Clin Epidemiol 2021;136:235−42. https://doi.org/10.1016/j.jclinepi.2021.04.010.

[61] Tello M, Zaiem F, Tolcher MC, Murad MH. Do not throw the baby out with the bath water: a guide for using non-significant results in practice. BMJ Evid Based Med 2016;21(5):161. https://doi.org/10.1136/ebmed-2016-110510.

[62] Wood J, Freemantle N, King M, Nazareth I. Trap of trends to statistical significance: likelihood of near significant P value becoming more significant with extra data. BMJ 2014;348:g2215. https://doi.org/10.1136/bmj.g2215.

[63] Aczel B, Palfi B, Szollosi A, Kovacs M, Szaszi B, Szecsi P, et al. Quantifying support for the null hypothesis in psychology: an empirical investigation. Adv Methods Pract Psychol Sci 2018;1(3):357−66. https://doi.org/10.1177/2515245918773742.

[64] Hartgerink CHJ, Wicherts JM, van Assen MALM. Too good to be false: nonsignificant results revisited. Collabra Psychol 2017;3(1):9. https://doi.org/10.1525/collabra.71.

[65] Hemming K, Javid I, Taljaard M. A review of high impact journals found that misinterpretation of non-statistically significant results from randomized trials was common. J Clin Epidemiol 2022;145:112−20. https://doi.org/10.1016/j.jclinepi.2022.01.014.

[66] Khan MS, Fonarow GC, Friede T, Lateef N, Khan SU, Anker SD, et al. Application of the reverse fragility index to statistically nonsignificant randomized clinical trial results. JAMA Netw Open 2020;3(8):e2012469. https://doi.org/10.1001/jamanetworkopen.2020.12469.

[67] Nead KT, Wehner MR, Mitra N. The use of ''Trend'' statements to describe statistically nonsignificant results in the oncology literature. JAMA Oncol 2018;4(12):1778−9. https://doi.org/10.1001/jamaoncol.2018.4524.

[68] Otte WM, Vinkers CH, Habets PC, van IJzendoorn DGP, Tijdink JK. Analysis of 567,758 randomized controlled trials published over 30 years reveals trends in phrases used to discuss results that do not reach statistical significance. Plos Biol 2022;20(2):e3001562. https://doi.org/10.1371/journal.pbio.3001562.

[69] Perneger T, Gayet-Ageron A. Evidence of lack of treatment efficacy derived from statistically nonsignificant results of randomized clinical trials. JAMA 2023;329(23):2050−6. https://doi.org/10.1001/ama.2023.8549.

[70] Assel M, Sjoberg D, Elders A, Wang X, Huo D, Botchway A, et al. Guidelines for reporting of statistics for clinical research in urology. Eur Urol 2019;75(3):358−67. https://doi.org/10.1016/j.eururo.2018.12.014.

[71] Lang TA, Altman DG. Basic statistical reporting for articles published in Biomedical Journals: the ''Statistical Analyses and Methods in the Published Literature'' or the SAMPL guidelines. Int J Nurs Stud 2015;52(1):5−9. https://doi.org/10.1016/j.ijnurstu.2014.09.006.

[72] Statistical guidelines for authors. J Med Screen 2008;15(1):51. https://doi.org/10.1258/jms.2008.008gos.

[73] Ou FS, Le-Rademacher JG, Ballman KV, Adjei AA, Mandrekar SJ. Guidelines for statistical reporting in medical journals. J Thorac Oncol 2020;15(11):1722−6. https://doi.org/10.1016/j.jtho.2020.08.019.

[74] Gartlehner G, Persad E, Ledinger D, Chapman A, Gadinger A, Feyertag J, et al. Beyond statistical significance: nuanced interpretations of statistically nonsignificant results were rare in cochrane reviews − a metaepidemiological study. J Clin Epidemiol 2023;160:46−53. https://doi.org/10.1016/j.jclinepi.2023.06.007.

[75] Gewandter JS, McDermott MP, Kitt RA, Chaudari J, Koch JG, Evans SR, et al. Interpretation of CIs in clinical trials with nonsignificant results: systematic review and recommendations. BMJ Open 2017;7(7):e017288. https://doi.org/10.1136/bmjopen-2017-017288.

[76] Akl EA, Khabsa J, Petkovic J, Magwood O, Lytvyn L, Motilall A, et al. ''Interest-holders'': a new term to replace ''stakeholders'' in the context of health research and policy. Cochrane Evid Synth Methods 2024;2(11):e70007. https://doi.org/10.1002/cesm.70007.

[77] Lakens D. Equivalence tests: a practical primer for t tests, correlations, and meta-analyses. Soc Psychol Personal Sci 2017;8(4):355−62. https://doi.org/10.1177/1948550617697177.

[78] Mun KT, Bonomo JB, Liebeskind DS, Saver JL. Fragility index meta-analysis of randomized controlled trials shows highly robust evidential strength for benefit of <3 hour intravenous alteplase. Stroke 2022;53(6):2069−74. https://doi.org/10.1161/STROKEAHA.121.038153.

[79] Gough D, Davies P, Jamtvedt G, Langlois E, Littell J, Lotfi T, et al. Evidence synthesis international (ESI): position statement. Syst Rev 2020;9(1):155. https://doi.org/10.1186/s13643-020-01415-5.

[80] Jackson D, Turner R, Rhodes K, Viechtbauer W. Methods for calculating confidence and credible intervals for the residual between-study variance in random effects meta-regression models. BMC Med Res Methodol 2014;14(1):103. https://doi.org/10.1186/1471-2288-14-103.

[81] Schünemann H, Vist G, Higgins J, et al. Chapter 15: interpreting results and drawing conclusions. Cochrane handbook for systematic reviews of interventions version 65. 2024. Available at: https://www.cochrane.org/authors/handbooks-and-manuals/handbook/current/chapter-15. Accessed September 12, 2025.

[82] Marson Smith PR, Ware L, Adams C, Chalmers I. Claims of ''no difference'' or ''no effect'' in cochrane and other systematic reviews. BMJ Evid Based Med 2021;26(3):118. https://doi.org/10.1136/bmjebm-2019-111257.

[83] Borenstein M. In a meta-analysis, the I-squared statistic does not tell us how much the effect size varies. J Clin Epidemiol 2022;152:281−4. https://doi.org/10.1016/j.jclinepi.2022.10.003.

[84] Greenland S, Senn SJ, Rothman KJ, Carlin JB, Poole C, Goodman SN, et al. Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. Eur J Epidemiol 2016;31(4):337−50. https://doi.org/10.1007/s10654-016-0149-3.

[85] Sousa-Pinto B, Neumann I, Vieira RJ, Bognanni A, Marques-Cruz M, Gil-Mata S, et al. Quantitative assessment of inconsistency in meta-analysis using decision thresholds with two new indices. J Clin Epidemiol 2025;181:111725. https://doi.org/10.1016/j.jclinepi.2025.111725.

[86] Morgano GP, Wiercioch W, Piovani D, Neumann I, Nieuwlaat R, Piggott T, et al. Defining decision thresholds for judgments on health benefits and harms using the grading of recommendations assessment, development, and evaluation (GRADE) Evidence to Decision (EtD) frameworks: a randomized methodological study (GRADE-THRESHOLD). J Clin Epidemiol 2025;179:111639. https://doi.org/10.1016/j.jclinepi.2024.111639.

[87] Johnson VE. Revised standards for statistical evidence. Proc Natl Acad Sci U S A 2013;110(48):19313−7. https://doi.org/10.1073/pnas.1313476110.

[88] Diamond G, Kaul S. Bayesian classification of clinical practice guidelines. Arch Intern Med 2009;169(15):1431−5. https://doi.org/10.1001/archinternmed.2009.235.

[89] Pitrou I, Boutron I, Ahmad N, Ravaud P. Reporting of safety results in published reports of randomized controlled trials. Arch Intern Med 2009;169(19):1756−61. https://doi.org/10.1001/archinternmed.2009.306.

[90] Golder S, Loke YK, Wright K, Norman G. Reporting of adverse events in published and unpublished studies of health care interventions: a systematic review. Plos Med 2016;13(9):e1002127. https://doi.org/10.1371/journal.pmed.1002127.

[91] Junqueira DR, Phillips R, Zorzela L, Golder S, Loke Y, Moher D, et al. Time to improve the reporting of harms in randomized controlled trials. J Clin Epidemiol 2021;136:216−20. https://doi.org/10.1016/j.jclinepi.2021.04.020.

[92] Bartoš F, Maier M, Wagenmakers EJ, Nippold F, Doucouliagos H, Ioannidis JPA, et al. Footprint of publication selection bias on meta-analyses in medicine, environmental sciences, psychology, and economics. Res Synth Methods 2024;15(3):500−11. https://doi.org/10.1002/jrsm.1703.

[93] DeVito NJ, Goldacre B. Catalogue of bias: publication bias. BMJ Evid Based Med 2019;24(2):53. https://doi.org/10.1136/bmjebm-2018-111107.