Decision Letter (JCQ-2026-0025)

From: [redacted]
To:
CC:
BCC:

Subject: Journal of Clinical Question - Decision on Manuscript ID JCQ-2026-0025

Body: 18-Aug-2026

Dear Dr Gong:

Manuscript ID JCQ-2026-0025 entitled "Heterogeneity in Meta-analysis: Concepts, Quantification, Exploration, and Clinical Interpretation" which you submitted to the Journal of Clinical Question, has been reviewed. The comments of the reviewer(s) are included at the bottom of this letter.

The reviewer(s) have recommended publication, but also suggest some revisions to your manuscript. Therefore, I invite you to respond to the reviewer(s)' comments and revise your manuscript.

To submit your revised manuscript, follow the below steps.
1. Access to https://mc.manuscriptcentral.com/jcq and log into the site.
2. Click "Author" at the top left of the screen at "Home" page.
3. Click "Manuscripts with Decisions" on Author Dashboard page.
4. Click "create a revision" on the manuscript you are going to revise.
5. The system creates your revision manuscript form. Provide your comments about the revised points at the response field, fill in on each field step by step as same as you did for the original submission, and submit your revised manuscript.
IMPORTANT: Your original files are available to you when you upload your revised manuscript. Please delete any redundant files before completing the submission.
Your manuscript number has been appended to denote a revision.

You will be unable to make your revisions on the originally submitted version of the manuscript. Instead, revise your manuscript using a word processing program and save it on your computer. Please also highlight the changes to your manuscript within the document by using the track changes mode in MS Word or by using bold or colored text.

Once the revised manuscript is prepared, you can upload it and submit it through your Author Center.

When submitting your revised manuscript, you will be able to respond to the comments made by the reviewer(s) in the space provided. You can use this space to document any changes you make to the original manuscript. In order to expedite the processing of the revised manuscript, please be as specific as possible in your response to the reviewer(s).

IMPORTANT: Your original files are available to you when you upload your revised manuscript. Please delete any redundant files before completing the submission.

Because we are trying to facilitate timely publication of manuscripts submitted to the Journal of Clinical Question, your revised manuscript should be uploaded as soon as possible. If it is not possible for you to submit your revision in a reasonable amount of time, we may have to consider your paper as a new submission.

Once again, thank you for submitting your manuscript to the Journal of Clinical Question and I look forward to receiving your revision.

Sincerely,
Richard S. Isaacs
Editor in Chief
Journal of Clinical Question

[Editor's Comments]
Associate Editor
Comments to the Author:
(There are no comments.)

[Reviewer(s)' Comments]
Reviewer: 1

Comments to the Author
1. The article cites recent review articles on same topic and essentially elaborates on same content on how to use and interpret heterogeneity measures. The same content is also available as chapter in cochrane manual and other sources. The essential question is "what does this article offer that is different than the articles from 2025, 2026 & cochrane manual ?" Without substantial novelty or new concept provided, and being a "method" paper without a literature review or analysis, this article seems to not add any new idea or a new perspective to already existing knowledge on the topic.
2. Declaration of Generative AI use: cites response to reviewers. But isn't the article being submitted here for the first time ?
3. References 14 and 16 are same.

Reviewer: 2

Comments to the Author
I read this paper carefully. It summarises a lot of conventional wisdom but is not ground breaking. I have no objections to its publication as a useful resource. A few points

1. Confidence intervals appear to use 0.975 quantiles and so should be described as 95% confidence intervals, not just confidence intervals.
2. For prediction intervals, study specific prediction intervals under the RE model have been proposed: Study specific prediction intervals for Random-effects meta-analysis: a tutorial: prediction intervals in meta-analysis. RCM van Aert, CH Schmid, D Svensson, D Jackson Research synthesis methods 12 (4), 429-447. Do the authors have any comments on this?
3. More discussion on I^2 and its limitations might be useful: Undue reliance on I 2 in assessing heterogeneity may mislead by Rucker et al
4. Meta-regression is fine for study level characteristics but to really understand heterogeneity I would argue that IPD (individual patient data) meta-analysis is really needed. Richard Riley and colleagues have done much work in that area.

These are really quite minor comments. This is a useful, albeit not very novel, contribution as I say, and I have rather few comments.

Date Sent: 18-Aug-2026

Comment form Reviewer 3

Title: Heterogeneity in Meta-analysis: Concepts, Quantification, Exploration, and Clinical Interpretation

Major revision

This manuscript provides a broad and useful discussion of heterogeneity in meta-analysis. Heterogeneity is a fundamental aspect of evidence synthesis, and the manuscript appropriately emphasises the importance of considering its underlying sources when interpreting differences between studies. However, several aspects would benefit from further clarification, greater methodological precision, and some restructuring. My detailed comments are provided below.

Major comments

Introduction

Lines 50–52: "Its objective is not merely to increase statistical precision, but to determine whether the available evidence supports a clinically meaningful and sufficiently consistent conclusion."

The description of the objective of meta-analysis could be refined. The statement that the objective is "not merely to increase statistical precision" may place too much emphasis on statistical precision as a defining feature of meta-analysis. More fundamentally, meta-analysis aims to quantitatively synthesise evidence from multiple studies to obtain an overall estimate of the intervention effect, while accounting for the characteristics and variability of the contributing studies. The authors may consider to revise this statement to distinguish the primary purpose of quantitative evidence synthesis from the potential gain in statistical precision.

Conceptual framework of heterogeneity

Line 95: "Statistical heterogeneity may reflect genuine effect modification, methodological bias, random error, data extraction errors, or the inappropriate combination of clinically incompatible studies."

The relationship between the different forms of heterogeneity could be described more clearly. Clinical and methodological diversity represent underlying differences between studies, whereas statistical heterogeneity refers to variability in the observed study-specific effects beyond that expected from sampling error. Although these three forms of heterogeneity are conceptually distinct, statistical heterogeneity may reflect these underlying clinical and/or methodological study differences. However, statistical heterogeneity may also arise from other sources, such as bias, errors, or random variation. Clarifying this relationship would help avoid presenting the three concepts as entirely independent forms of heterogeneity.

Line 157: The normality assumption for the study-specific true effects in the conventional random-effects model should be presented more explicitly as a modelling choice rather than as an inherent property of random-effects meta-analysis. The normal distribution is a convenient and commonly used assumption, but alternative random-effects distributions have also been proposed, including t-distributions, skewed distributions and mixture of distribution. The authors may wish to mention this distinction, particularly given the focus of the manuscript on understanding and modelling heterogeneity.

In addition, the parameters in the random-effects model should be clearly interpreted. The parameter representing the centre of the random-effects distribution is typically interpreted as the mean treatment effect, μ, whereas the parameter representing the variance of this distribution corresponds to the

between-study variance, τ^2, and quantifies statistical heterogeneity. At present, this interpretation is not sufficiently clear when the model is first introduced. The connection between the mathematical notation and the subsequent terminology, especially for the "between-study variance" should be made explicit.

Line 168: "A statistically valid random-effects estimate…"

It is unclear what is meant by "a statistically valid random-effects estimate." Please clarify what criterion is being used to define validity in this context. If the authors are referring to appropriate estimation under the assumed model, this could be stated more precisely.

Quantifying statistical heterogeneity

The section could be strengthened by discussing visual assessment of heterogeneity. Forest plots provide an important graphical assessment of the magnitude, direction, and consistency of study-specific effects and should be introduced earlier in the manuscript rather than only in the reporting section.

Exploring sources of heterogeneity

The discussion of subgroup analyses and meta-regression would benefit from a stronger acknowledgement of their limitations. Both approaches can have limited statistical power to identify or explain between-study differences, particularly when the number of studies is small or when several covariates are considered simultaneously.

The distinction between subgroup analysis and meta-regression could also be described more precisely. Subgroup analysis is generally used for categorical study characteristics, whereas meta-regression can incorporate continuous as well as categorical study-level covariates.

I also suggest expanding this section to include flexible random-effects models. Recent methodological work has investigated models that relax the conventional normality assumption for the random-effects distribution, including skewed distributions, t-distributions, mixture models, and Dirichlet Process models. In particular, Panagiotopoulou et al. provide a systematic review and simulation study of such approaches. Their findings suggest that flexible random-effects models may be useful when substantial heterogeneity remains unexplained by available study-level covariates, particularly when the distribution of the underlying effects may depart substantially from normality. (Panagiotopoulou K, Evrenoglou T, Schmid CH, Metelli S, Chaimani A. Meta-analysis models relaxing the random-effects normality assumption: methodological systematic review and simulation study. BMC Med Res Methodol. 2025 Oct 16;25(1):231. doi: 10.1186/s12874-025-02658-3. PMID: 41102684; PMCID: PMC12532406.)

Line 263: The statement "…excluding influential outliers, and conducting leave-one-out analyses" may fit more naturally in the subsequent section on outliers and influence. Consider moving this sentence to improve the organisation of the manuscript.

Outliers, influence, and small-study effects

This section would benefit from a more explicit discussion of how outlying and influential studies can be identified. Outliers may be identified through visual inspection of forest plots, but statistical diagnostics can provide a more formal assessment. Viechtbauer and Cheung provide a comprehensive discussion of outlier and influence diagnostics specifically for meta-analysis. (Viechtbauer W, Cheung MW. Outlier and

influence diagnostics for meta-analysis. Res Synth Methods. 2010;1(2):112–125. doi:10.1002/jrsm.11. PMID:26061377)

It would also be useful to distinguish outliers from influential studies. A study can be statistically unusual without having a substantial influence on the pooled estimate, whereas a study may have considerable influence because of its precision or other characteristics without necessarily being an outlier. These concepts therefore should not be treated as interchangeable.

Recommended reporting of heterogeneity

Lines 286–287: This appears to be the first point at which the forest plot is introduced. Since the forest plot is one of the principal tools for visually assessing the consistency of study-specific estimates, I suggest introducing it earlier, in the section discussing the identification and quantification of statistical heterogeneity. The current Cochrane Handbook similarly emphasises visual inspection of study confidence intervals in forest plots when assessing statistical heterogeneity.

Practical framework for clinical meta-analysts

Lines 308–309: “Potential sources of heterogeneity should be investigated…”

This recommendation could be expanded. When heterogeneity cannot be adequately explained by available clinical or methodological characteristics through subgroup or meta-regression analyses, the authors could mention that flexible random-effects models may provide an alternative approach for modelling the remaining between-study variation. This should be presented as complementary to, rather than a replacement for, investigation of potential sources of heterogeneity.

Minor comments

Lines 22–23: “Heterogeneity … available evidence.”

Heterogeneity alone does not determine whether a pooled estimate is clinically meaningful. It is one of several factors that should be considered when interpreting a meta-analytic result, together with the magnitude and direction of the effect, uncertainty around the estimate, clinical relevance, risk of bias, and applicability of the evidence. Please consider rephrasing this statement.

Line 36: Statistical heterogeneity is a property of the results being quantitatively synthesised and therefore cannot be formally assessed before a quantitative synthesis has been performed.

Lines 64–65: Please add Reference 2.

Lines 104–106: “A useful distinction is between prognostic factors, which alter the underlying risk of an outcome, and effect modifiers, which alter the relative effect of an intervention.”

The use of “alter” may be misleading. Prognostic factors are characteristics associated with the outcome risk, whereas effect modifiers are characteristics for which the relative treatment effect differs across levels of the characteristic. Consider revising this wording to avoid implying a causal effect of prognostic factors on outcome risk.

Line 122: I suggest revising the statement that methodological heterogeneity "mimics" statistical heterogeneity. Methodological differences between studies may contribute to observed differences in effect estimates and therefore to statistical heterogeneity, but methodological heterogeneity and statistical heterogeneity represent conceptually distinct constructs.

Line 138: Please clarify that the within-study variances (or their estimates) are typically treated as known or fixed when fitting the conventional random-effects model.

Line 148: Please specify that the model assumes (k) studies.

Line 152: Instead of "sufficiently similar," it may be more appropriate to state that the studies "are considered homogeneous".

Line 157: I suggest replacing "Study-specific true effects are assumed to follow a distribution…" with a more precise formulation such as: "Study-specific true effects are assumed to be exchangeable draws from a common distribution…". This wording better reflects the underlying assumption of the conventional random-effects model.

Line 198: The DerSimonian-Laird estimator requires a reference.

Lines 200–201: Please provide a reference.

Line 213: If τ^2 is zero, the prediction interval reduces to the corresponding confidence interval. This point could be mentioned.

Lines 223–230: These sentences appear to summarise several approaches to detecting and quantifying heterogeneity rather than referring specifically to prediction intervals. Consider moving them to a separate paragraph to improve the logical structure.

Lines 233–234: Please provide a reference.

Lines 242–243: Please provide a reference.

Line 253: The equation should define the parameters $\beta_0, \beta_1, \beta_2, X_{i1}, X_{i2}$ and ε_i.

Table 2: If the authors incorporate flexible meta-analysis models into the manuscript, these approaches should also be included in Table 2 as potential analytical approaches for modelling unexplained heterogeneity.

Author's Response to Decision Letter for (JCQ-2026-0025)

Heterogeneity in Meta-analysis: Concepts, Quantification, Exploration, and Clinical Interpretation

Dear Editor and Reviewer,

We sincerely thank the reviewers for their careful evaluation of our manuscript and for their constructive comments. We have revised the manuscript substantially in response to these suggestions. In particular, we have clarified the intended contribution of the article, expanded the discussion of the limitations of I^2, added discussion of study-specific prediction intervals and individual participant data meta-analysis, corrected the terminology for confidence intervals, revised the Generative AI declaration, and removed the duplicate reference. Our point-by-point responses are provided below.

Reviewer 1
Major Comment 1: Lines 50–52: "Its objective is not merely to increase statistical precision, but to determine whether the available evidence supports a clinically meaningful and sufficiently consistent conclusion."
The description of the objective of meta-analysis could be refined. The statement that the objective is "not merely to increase statistical precision" may place too much emphasis on statistical precision as a defining feature of meta-analysis. More fundamentally, meta-analysis aims to quantitatively synthesise evidence from multiple studies to obtain an overall estimate of the intervention effect, while accounting for the characteristics and variability of the contributing studies. The authors may consider to revise this statement to distinguish the primary purpose of quantitative evidence synthesis from the potential gain in statistical precision.
Response: We agree and have revised the opening definition accordingly. The Introduction now identifies quantitative synthesis and estimation of an overall intervention effect as the primary purpose, while describing increased precision as a possible consequence rather than the defining objective.
Revised manuscript text: "Meta-analysis quantitatively synthesizes effect estimates from multiple studies to obtain an overall estimate of an intervention effect while considering the characteristics and variability of the contributing studies. Increased statistical precision may be a consequence of this synthesis, but it is not its defining purpose." (Lines 50–53)

Major Comment 2: Line 95: "Statistical heterogeneity may reflect genuine effect modification, methodological bias, random error, data extraction errors, or the inappropriate combination of clinically incompatible studies."
The relationship between the different forms of heterogeneity could be described more clearly. Clinical and methodological diversity represent underlying differences between studies, whereas statistical heterogeneity refers to variability in the observed study-specific effects beyond that expected from sampling error. Although these three forms of heterogeneity are conceptually distinct, statistical heterogeneity may reflect these underlying clinical and/or methodological study differences. However, statistical heterogeneity may also arise from other sources, such as bias, errors, or random variation. Clarifying this relationship would help avoid presenting the three concepts as entirely independent forms of heterogeneity.
Response: We revised the conceptual framework to make this relationship explicit. Clinical and methodological diversity are now described as underlying differences in study characteristics, whereas statistical heterogeneity is defined as observed variability beyond within-study sampling error. The text also states that statistical heterogeneity may reflect clinical or methodological differences or arise from bias, data or analysis errors, chance variation, or inappropriate pooling.
Revised manuscript text: "Clinical and methodological diversity refer to underlying differences in study characteristics. Statistical heterogeneity refers to variability in study-specific effect estimates beyond that expected from within-study sampling error. Although conceptually distinct, statistical heterogeneity may reflect underlying clinical or methodological differences between studies. It may also arise from bias, data extraction or analysis errors, residual random variation, or the inappropriate combination of clinically incompatible studies." (Lines 91–96)

Major Comment 3: Line 157: The normality assumption for the study-specific true effects in the conventional random-effects model should be presented more explicitly as a modelling choice rather than as an inherent property of random-effects meta-analysis. The normal distribution is a convenient and commonly used assumption, but alternative random-effects distributions have also been proposed, including t-distributions, skewed distributions and mixture of distribution. The authors may wish to mention this distinction, particularly given the focus of the manuscript on understanding and modelling heterogeneity.
In addition, the parameters in the random-effects model should be clearly interpreted. The parameter representing the centre of the random-effects distribution is typically interpreted as the mean treatment

effect, μ, whereas the parameter representing the variance of this distribution corresponds to the between-study variance, $\tau 2$, and quantifies statistical heterogeneity. At present, this interpretation is not sufficiently clear when the model is first introduced. The connection between the mathematical notation and the subsequent terminology, especially for the "between-study variance" should be made explicit.
Response: We agree. The random-effects section now states that study-specific true effects are exchangeable draws from a common distribution and that normality is a convenient modelling choice. Alternative t, skewed, mixture, and nonparametric distributions are noted. The text now explicitly defines μ as the center or mean treatment effect and τ² as the between-study variance. A separate subsection further discusses flexible random-effects models.
Revised manuscript text: "A random-effects model assumes that each of k studies estimates a different but related true effect. In the conventional model, the study-specific true effects are treated as exchangeable draws from a common distribution. A normal distribution is a convenient and commonly used modeling choice rather than an inherent property of random-effects meta-analysis; alternatives include t-distributions, skewed distributions, mixture distributions, and nonparametric distributions. Here, μ is the center of the random-effects distribution and is interpreted as the mean treatment effect, whereas τ² is the variance of that distribution and represents the between-study variance that quantifies statistical heterogeneity."(Lines 160–170 and 284–292)

Major Comment 4: Line 168: "A statistically valid random-effects estimate…"
It is unclear what is meant by "a statistically valid random-effects estimate." Please clarify what criterion is being used to define validity in this context. If the authors are referring to appropriate estimation under the assumed model, this could be stated more precisely.
Response: We replaced the ambiguous phrase with more precise wording. The revised text refers to an estimate 'appropriately obtained under the assumed random-effects model' and distinguishes model-based estimation from clinical appropriateness when studies differ substantially in direction or clinical implication.
Revised manuscript text: "Even an estimate appropriately estimated under the assumed random-effects model may be clinically inappropriate when studies differ substantially in direction or clinical implication." (Lines 177–180)

Major Comment 5: The section could be strengthened by discussing visual assessment of heterogeneity. Forest plots provide an important graphical assessment of the magnitude, direction, and consistency of study-specific effects and should be introduced earlier in the manuscript rather than only in the reporting section.
Response: We added a new opening paragraph to the quantification section stating that visual inspection of the forest plot is an essential first step and explaining how effect direction, magnitude, precision, confidence-interval overlap, and possible outlying results should be assessed.
Revised manuscript text: "Visual inspection of a forest plot is an essential first step in assessing statistical heterogeneity. The distribution of study-specific effect estimates and their confidence intervals, together with the magnitude, direction, and precision of the effects, can reveal inconsistency or possible outlying results that no single numerical measure fully captures." (Lines 181–185)

Major Comment 6: The discussion of subgroup analyses and meta-regression would benefit from a stronger acknowledgement of their limitations. Both approaches can have limited statistical power to identify or explain between-study differences, particularly when the number of studies is small or when several covariates are considered simultaneously.
Response: We expanded the discussion of limitations. The revised text notes limited power for subgroup comparisons, particularly with few studies or multiple subgroups, and explains that meta-regression power and model stability decline when few studies or several covariates are analyzed. Ecological bias and spurious study-level associations are also emphasized.
Revised manuscript text: "Their power to detect genuine between-subgroup differences is often limited, particularly when few studies are available or multiple subgroups are examined. However, both subgroup analyses and meta-regression have limited power when few studies are available. Power and model stability decline further when several covariates are examined simultaneously. Meta-regression is also vulnerable to ecological bias and spurious associations because study-level relationships may not reflect individual-level treatment effect modification." (Lines 263–268 and 278–283)

Major Comment 7: The distinction between subgroup analysis and meta-regression could also be described more precisely. Subgroup analysis is generally used for categorical study characteristics, whereas meta-regression can incorporate continuous as well as categorical study-level covariates.
Response: We revised the section to distinguish the methods explicitly. Subgroup analysis is now described as comparing pooled effects across predefined categories, whereas meta-regression is described as relating effect estimates to continuous or categorical study-level covariates.
Revised manuscript text: "Subgroup analyses generally compare pooled effects across predefined categories of study characteristics. Meta-regression extends subgroup analysis by relating effect estimates to one or more continuous or categorical study-level covariates." (Lines 263–270)

Major Comment 8 I also suggest expanding this section to include flexible random-effects models. Recent methodological work has investigated models that relax the conventional normality assumption for the random-effects distribution, including skewed distributions, t-distributions, mixture models, and Dirichlet Process models. In particular, Panagiotopoulou et al. provide a systematic review and simulation study of such approaches.
Their findings suggest that flexible random-effects models may be useful when substantial heterogeneity remains unexplained by available study-level covariates, particularly when the distribution of the

underlying effects may depart substantially from normality. (Panagiotopoulou K, Evrenoglou T, SchmidCH, Metelli S, Chaimani A. Meta-analysis models relaxing the random-effects normality assumption: methodological systematic review and simulation study. BMC Med Res Methodol. 2025 Oct 16;25(1):231. doi: 10.1186/s12874-025-02658-3. PMID: 41102684; PMCID: PMC12532406.)
Response: We added a dedicated subsection on flexible random-effects models. It describes when these approaches may be considered, outlines the requested distributional alternatives, explains that they complement rather than replace investigation of clinical and methodological sources, and notes assumptions, data requirements, convergence, and sensitivity. The Panagiotopoulou et al. study was added as Reference 37, and flexible models were also added to Table 2.
Revised manuscript text: "When substantial heterogeneity remains unexplained by available study-level covariates, flexible random-effects models can be considered to relax the conventional normality assumption. Approaches based on t-distributions, skewed distributions, finite-mixture models, or Dirichlet process models may better represent heavy tails, asymmetry, or latent clusters in the distribution of true effects. Such models are complementary to, rather than a replacement for, investigation of clinical and methodological sources of heterogeneity; their assumptions, data requirements, convergence, and sensitivity to model specification should be examined carefully." (Lines 284–292, Table 2 Page 23)

Major Comment 9: Line 263: The statement "…excluding influential outliers, and conducting leave-one-out analyses" may fit more naturally in the subsequent section on outliers and influence. Consider moving this sentence to improve the organisation of the manuscript.
Response: We reorganized the material. The general sensitivity-analysis paragraph now focuses on analytical decisions, whereas outlier screening, leave-one-out analysis, and influence diagnostics are presented together in the dedicated outlier and influence section.
Revised manuscript text: "Sensitivity analyses examine whether conclusions are robust to analytical decisions. Relevant approaches include excluding studies at high risk of bias, restricting analyses to randomized controlled trials, comparing fixed-effect and random-effects estimates, using alternative estimators of τ^2, and applying Hartung-Knapp confidence intervals. Potential outliers can be screened visually in forest plots and evaluated using standardized or studentized residuals. Influence can be assessed using leave-one-out analyses…" (Lines 312–320)

Major Comment 10: This section would benefit from a more explicit discussion of how outlying and influential studies can be identified. Outliers may be identified through visual inspection of forest plots, but statistical diagnostics can provide a more formal assessment. Viechtbauer and Cheung provide a comprehensive discussion of outlier and influence diagnostics specifically for meta-analysis. (Viechtbauer W, Cheung MW. Outlier and influence diagnostics for meta-analysis. Res Synth Methods. 2010;1(2):112–125. doi:10.1002/jrsm.11. PMID:26061377)
Response: We added practical diagnostic guidance. Potential outliers may be screened in forest plots and assessed with standardized or studentized residuals. Influence may be examined using leave-one-out analyses, Cook's distances, DFBETAS, covariance ratios, hat values, and changes in the pooled effect or τ^2. Viechtbauer and Cheung's article was added as Reference 41.
Revised manuscript text: "Potential outliers can be screened visually in forest plots and evaluated using standardized or studentized residuals. Influence can be assessed using leave-one-out analyses, Cook's distances, DFBETAS, covariance ratios, hat values, and changes in the pooled effect or τ^2 after omitting each study. These diagnostics should prompt data verification and substantive investigation rather than automatic study exclusion." (Lines 325–329).

Major Comment 11: It would also be useful to distinguish outliers from influential studies. A study can be statistically unusual without having a substantial influence on the pooled estimate, whereas a study may have considerable influence because of its precision or other characteristics without necessarily being an outlier. These concepts therefore should not be treated as interchangeable.
Response: We added explicit definitions and examples distinguishing the two concepts. The revised text also cautions that a study should not be excluded merely because it changes the pooled result and recommends investigating substantive explanations first.
Revised manuscript text: "An outlier is a study whose observed effect is statistically unusual relative to the fitted model, whereas an influential study is one whose inclusion materially changes the pooled estimate, estimated heterogeneity, or model coefficients. These concepts are not interchangeable: an outlier may have little influence, and a highly precise study may be influential without being an outlier." (Lines 307–314).

Major Comment 12: Lines 286–287: This appears to be the first point at which the forest plot is introduced. Since the forest plot is one of the principal tools for visually assessing the consistency of study-specific estimates, I suggest introducing it earlier, in the section discussing the identification and quantification of statistical heterogeneity. The current Cochrane Handbook similarly emphasises visual inspection of study confidence intervals in forest plots when assessing statistical heterogeneity.
Response: We introduced forest-plot assessment at the beginning of the quantification section and cited the Cochrane Handbook. The reporting section now reinforces that the forest plot should be interpreted alongside numerical heterogeneity measures rather than separately.
Revised manuscript text: "Visual inspection of a forest plot is an essential first step in assessing statistical heterogeneity… The forest plot should be interpreted alongside heterogeneity measures, not separately. Visual assessment may reveal differences in direction of effect, study precision, or outlying results that are not captured by a single numerical statistic." (Lines 182–185 and 334–339)

Major Comment 13: Lines 308–309: "Potential sources of heterogeneity should be investigated…"
This recommendation could be expanded. When heterogeneity cannot be adequately explained by available clinical or methodological characteristics through subgroup or meta-regression analyses, the

authors could mention that flexible random-effects models may provide an alternative approach for modelling the remaining between-study variation. This should be presented as complementary to, rather than a replacement for, investigation of potential sources of heterogeneity.
Response: We expanded the practical framework accordingly. It now recommends prespecified subgroup or meta-regression analyses supported by sensitivity and influence analyses and notes that flexible random-effects models may complement these investigations when heterogeneity remains unexplained.
Revised manuscript text: "Potential sources should be explored through prespecified subgroup or meta-regression analyses, supported by sensitivity and influence analyses; when heterogeneity remains unexplained, flexible random-effects models may complement these investigations." (Lines 354–361)

Minor comment 1: Lines 22–23: "Heterogeneity … available evidence."
Heterogeneity alone does not determine whether a pooled estimate is clinically meaningful. It is one of several factors that should be considered when interpreting a meta-analytic result, together with the magnitude and direction of the effect, uncertainty around the estimate, clinical relevance, risk of bias, and applicability of the evidence. Please consider rephrasing this statement.
Response: We revised the Clinical Question Box to state that heterogeneity is one of several considerations. Interpretation now integrates effect magnitude and direction, uncertainty, clinical relevance, risk of bias, applicability, and the three domains of heterogeneity.
Revised manuscript text: "Heterogeneity is one of several factors that determine whether a pooled meta-analytic estimate is clinically meaningful. Interpretation should integrate the magnitude and direction of the effect, uncertainty around the estimate, clinical relevance, risk of bias, applicability, and clinical, methodological, and statistical heterogeneity." (Lines 21–30)

Minor comment 2: Line 36: Statistical heterogeneity is a property of the results being quantitatively synthesised and therefore cannot be formally assessed before a quantitative synthesis has been performed.
Response: We revised the Abstract to distinguish the timing of assessment: clinical and methodological diversity should be assessed before pooling, whereas statistical heterogeneity is evaluated from study-specific results during quantitative synthesis.
Revised manuscript text: "Clinical and methodological diversity should be assessed before pooling, whereas statistical heterogeneity is evaluated from the study-specific results during quantitative synthesis." (Lines 36–38)

Minor comment 3: Lines 64–65: Please add Reference 2.
Response: Reference 2 has been added to support the statement describing application of Cochrane guidance in the review approach.
Revised manuscript text: "Relevant reporting principles, together with guidance from the Cochrane Handbook and the Grading of Recommendations Assessment, Development and Evaluation (GRADE) framework, were applied where appropriate to promote transparent reporting and interpretation of heterogeneity." (Lines 83–86)

Minor comment 4: Lines 104–106: "A useful distinction is between prognostic factors, which alter the underlying risk of an outcome, and effect modifiers, which alter the relative effect of an intervention."
The use of "alter" may be misleading. Prognostic factors are characteristics associated with the outcome risk, whereas effect modifiers are characteristics for which the relative treatment effect differs acrosslevels of the characteristic. Consider revising this wording to avoid implying a causal effect of prognostic factors on outcome risk.
Response: We revised the definitions to avoid implying causality. Prognostic factors are now described as characteristics associated with underlying outcome risk, and effect modifiers as characteristics across whose levels the relative intervention effect differs.
Revised manuscript text: "Prognostic factors are characteristics associated with the underlying risk of an outcome, whereas effect modifiers are characteristics for which the relative intervention effect differs across levels of the characteristic." (Lines 104–106)

Minor comment 5: Line 122: I suggest revising the statement that methodological heterogeneity "mimics" statistical heterogeneity. Methodological differences between studies may contribute to observed differences in effect estimates and therefore to statistical heterogeneity, but methodological heterogeneity and statistical heterogeneity represent conceptually distinct constructs.
Response: We removed 'mimics' and now state that methodological differences may contribute to observed differences in effect estimates and thus to statistical heterogeneity, while emphasizing that methodological and statistical heterogeneity remain conceptually distinct.
Revised manuscript text: "Methodological differences may contribute to observed differences in effect estimates and therefore to statistical heterogeneity, but methodological and statistical heterogeneity remain conceptually distinct." (Lines 116–122)

Minor comment 6: Line 138: Please clarify that the within-study variances (or their estimates) are typically treated as known or fixed when fitting the conventional random-effects model.
Response: This clarification has been added immediately after the notation for the within-study variance.
Revised manuscript text: "The within-study variances, or their estimates, are typically treated as known or fixed when fitting the conventional random-effects model." (Lines 139–143)

Minor comment 7: Line 148: Please specify that the model assumes (k) studies.
Response: We now define the framework for k studies at the outset and repeat the k-study context when introducing the random-effects model.
Revised manuscript text: "For k studies (i = 1, ..., k), let Y_i denote the estimated intervention effect in

study i, with within-study variance v_i." (Lines 139–140 and 160–161)

Minor comment 8: Line 152: Instead of "sufficiently similar," it may be more appropriate to state that the studies "are considered homogeneous".
Response: The wording has been revised as suggested, while retaining the additional requirement that a common underlying effect be clinically defensible.
Revised manuscript text: "The fixed-effect model may be appropriate when the studies are considered homogeneous and when the assumption of a common underlying effect is clinically defensible." (Lines 157–159)

Minor comment 9: Line 157: I suggest replacing "Study-specific true effects are assumed to follow a distribution…" with a more precise formulation such as: "Study-specific true effects are assumed to be exchangeable draws from a common distribution…". This wording better reflects the underlying assumption of the conventional random-effects model.
Response: We adopted the suggested formulation and immediately clarified that the conventional normal distribution is a modeling choice rather than an inherent feature of random-effects meta-analysis.
Revised manuscript text: "In the conventional model, the study-specific true effects are treated as exchangeable draws from a common distribution." (Lines 160–165)

Minor comment 10: Line 198: The DerSimonian-Laird estimator requires a reference.
Response: The original DerSimonian and Laird article is now cited at the first mention of the estimator as Reference 18. (Lines 210–214).

Minor comment 11: Lines 200–201: Please provide a reference.
Response: We added Reference 26 to support the discussion of restricted maximum likelihood, Paule–Mandel, and other estimators of between-study variance. (Lines 215–218).

Minor comment 12: Line 213: If $\tau 2$ is zero, the prediction interval reduces to the corresponding confidence interval. This point could be mentioned.
Response: This point has been added explicitly at the end of the initial prediction-interval paragraph.
Revised manuscript text: "If τ^2 is zero, the prediction interval reduces to the corresponding confidence interval." (Lines 235–236)

Minor comment 13: Lines 223–230: These sentences appear to summarise several approaches to detecting and quantifying heterogeneity rather than referring specifically to prediction intervals. Consider moving them to a separate paragraph to improve the logical structure.
Response: We moved and consolidated this material into a separate paragraph that contrasts the distinct roles of Q, I^2, τ^2, and the prediction interval and recommends their complementary interpretation.
Revised manuscript text: "Cochran's Q assesses whether the observed dispersion is greater than would be expected from sampling error alone; I^2 describes the proportion of observed variation attributable to between-study heterogeneity rather than sampling error; τ^2 estimates the absolute between-study variance; and the prediction interval provides a range within which the true effect in a comparable future setting is expected to lie. These measures address distinct but complementary questions and should therefore be interpreted together rather than used interchangeably." (Lines 242–248)

Minor comment 14: Lines 233–234: Please provide a reference.
Response: We added References 31 and 32 to support the discussion of the Hartung–Knapp–Sidik–Jonkman approach and its use when few studies are available. (Lines 247–251)

Minor comment 15: Lines 242–243: Please provide a reference.
Response: We added Reference 33 to support the statement regarding modified Hartung–Knapp procedures in selected circumstances. (Lines 267–270).

Minor comment 16: Line 253: The equation should define the parameters $\beta 0$, $\beta 1$, $\beta 2$, $Xi1$, $Xi2$ and εi.
Response: We added definitions for all requested parameters and also defined the residual random effect u_i and its variance τ^2.
Revised manuscript text: "In this model, β_0 is the expected effect when the covariates equal zero; β_1 and β_2 are study-level regression coefficients; X_{i1} and X_{i2} are the values of the two study-level covariates; u_i is the residual between-study random effect with variance τ^2; and ε_i is the within-study sampling error, typically assumed to have mean zero and variance v_i." (Lines 281–286)

Minor comment 17: Table 2: If the authors incorporate flexible meta-analysis models into the manuscript, these approaches should also be included in Table 2 as potential analytical approaches for modelling unexplained heterogeneity.
Response: We added a dedicated row describing flexible random-effects models, including the distributional assumption, rationale for use, model diagnostics, convergence, and sensitivity to the conventional normal random-effects model.
Revised manuscript text: "Flexible random-effects models — Distributional assumption (e.g., t, skewed, mixture, or Dirichlet process), rationale for use, model diagnostics, convergence, and sensitivity to the conventional normal random-effects model." (Table 2).

Reviewer 2
Comment 1: The article cites recent review articles on same topic and essentially elaborates on same

content on how to use and interpret heterogeneity measures. The same content is also available as chapter in Cochrane manual and other sources. The essential question is "what does this article offer that is different than the articles from 2025, 2026 & Cochrane manual?" Without substantial novelty or new concept provided, and being a "method" paper without a literature review or analysis, this article seems to not add any new idea or a new perspective to already existing knowledge on the topic.
Response: Thank you for this important comment. We agree that the manuscript should not simply reproduce established descriptions of heterogeneity measures already available in the Cochrane Handbook and recent reviews. We also acknowledge that the article does not introduce a new statistical estimator or theory. Accordingly, the manuscript has been substantially revised to clarify its scope and contribution.
First, the article is now explicitly presented as a narrative methodological review, and a new section, "Review Approach and Literature Selection," describes the targeted, non-exhaustive literature selection process and clarifies that no de novo quantitative synthesis was performed.
Second, the revised manuscript emphasizes an integrated clinical framework in which clinical and methodological heterogeneity are assessed before statistical heterogeneity is quantified. Statistical heterogeneity is therefore treated as a signal requiring explanation rather than as a diagnosis of its underlying cause.
Third, the manuscript extends beyond conventional discussion of Q and I^2 by integrating τ^2, prediction intervals, Hartung–Knapp methods, subgroup analysis, meta-regression, influence diagnostics, and flexible random-effects models, while highlighting their complementary roles and limitations.
Finally, the revised article links heterogeneity assessment with clinical applicability, GRADE certainty assessment, and practical reporting. Table 2 and Figure 1 provide an integrated reporting and decision framework that brings together PRISMA, Cochrane, and GRADE principles and guides readers from assessment of clinical comparability through statistical exploration to the final judgment of whether a pooled estimate remains clinically interpretable.
The added value of the revised article is not a new heterogeneity statistic, but a clinically focused and contemporary synthesis that integrates statistical measures, clinical and methodological assessment, certainty of evidence, and practical reporting into a coherent workflow for clinicians and meta-analysts.

Comment 2: Declaration of Generative AI use: cites response to reviewers. But isn't the article being submitted here for the first time?
Response: Thank you for pointing this out. The reference to the preparation of a "response to reviewers" was inadvertently retained from a template used for a previous manuscript and was inappropriate for this initial submission. We have corrected the Generative AI declaration by removing this wording. The revised statement now accurately describes the use of generative AI in the preparation of the manuscript and is consistent with the journal's policy.

Comment 3: References 14 and 16 are same.
Response: Thank you for identifying this duplication. We have replaced the duplicated reference 16 with a more appropriate recent methodological reference supporting sensitivity analyses and meta-regression. The corresponding citation has been updated in the revised manuscript (Lines 130–133), and the reference list has been corrected accordingly.

Reviewer 3
We sincerely thank the reviewer for the careful assessment of our manuscript and for the constructive suggestions. The comments have helped us improve the methodological precision and scope of the discussion.
Comment 1: Confidence intervals appear to use 0.975 quantiles and so should be described as 95% confidence intervals, not just confidence intervals.
Response: We agree. We have revised the manuscript to explicitly describe intervals based on the 0.975 quantile as 95% confidence intervals (95% CIs). The terminology has been updated in the discussion of forest plots, prediction intervals, and confidence intervals under random-effects models to avoid ambiguity regarding the confidence level. Location: Lines 182–183, 225–235, and 256–270.
Revised manuscript text: "A 95% confidence interval around the pooled effect quantifies uncertainty in the mean intervention effect." "The 95% confidence interval is: …"

Comment 2: For prediction intervals, study specific prediction intervals under the RE model have been proposed: Study specific prediction intervals for Random-effects meta-analysis: a tutorial: prediction intervals in meta-analysis. RCM van Aert, CH Schmid, D Svensson, D Jackson Research synthesis methods 12 (4), 429-447. Do the authors have any comments on this?
Response: Thank you for highlighting this important development. We have expanded the Prediction Intervals section to distinguish the conventional prediction interval for the true effect in a new study from study-specific prediction intervals under random-effects models.
The revised text explains that study-specific prediction intervals can accompany empirical Bayes estimates, or equivalently best linear unbiased predictions, of study-specific true effects. We also note their potential limitations and emphasize that they should complement, rather than replace, the conventional prediction interval and investigation of clinical and methodological sources of heterogeneity.
The tutorial by van Aert and colleagues has been added as Reference 30. (Lines 243–250)

Comment 3: More discussion on I^2 and its limitations might be useful: Undue reliance on I^2 in assessing heterogeneity may mislead by Rücker et al.
Response: We agree and have expanded the discussion of I^2 and its limitations, including the concern raised by Rücker and colleagues.
The revised manuscript emphasizes that I^2 is a relative measure of inconsistency that depends on within-study precision and does not directly quantify the absolute magnitude or clinical importance of heterogeneity. We now state that I^2 may increase as studies become more precise even when the

underlying between-study variance remains similar, and we recommend interpreting I^2 together with τ^2, study-specific effects, 95% confidence intervals, prediction intervals, and the clinical and methodological context. Rücker et al. has been added as Reference 25. (Lines 200–210)

Comment 4: Meta-regression is fine for study level characteristics but to really understand heterogeneity I would argue that IPD (individual patient data) meta-analysis is really needed. Richard Riley and colleagues have done much work in that area.
Response: We agree with this important distinction. We have expanded the meta-regression section to clarify that aggregate-data meta-regression is limited to study-level characteristics and is vulnerable to ecological bias. We now explicitly discuss individual participant data (IPD) meta-analysis when heterogeneity is hypothesized to arise from participant-level effect modifiers.
The revised text explains that IPD meta-analysis permits treatment–covariate interactions to be examined directly using participant-level information and can therefore better evaluate individual-level effect modification when within-study and between-study information are appropriately separated. The methodological work of Riley and colleagues and a recent systematic review of effect-modification analyses in IPD meta-analyses are cited as References 39 and 40. (Lines 287–298; Table 2, p. 24.)

We again thank the reviewers for their thoughtful and constructive comments. These revisions have helped us clarify the scope and contribution of the manuscript and strengthen the discussion of heterogeneity beyond the conventional reporting of statistical measures.

Best Regards
Hao Chen
Department of Pulmonology, Yokohama City University, Yokohama, Japan
Tel: +81 045-787-2800
Email: chinsmd@gmail.com