Archives of Academic Emergency Medicine. 2026; 14(1): e15

OPEN ACCESS

LETTER TO EDITOR

Concerns About Heterogeneity Assessment in Statistics of Meta-analyses; a Letter to the Editor

Manoochehr Karami[1,2], Mehrdad Farrokhi[2]*

1. Pandemic and Epidemic preparedness and Response Research Network, Research Institute for Health Sciences and Environment, Shahid Beheshti University of Medical Sciences, Tehran, Iran.

2. Department of Epidemiology, School of Public Health and Safety, Shahid Beheshti University of Medical Sciences, Tehran, Iran.

Received: February 2026; Accepted: March 2026; Published online: 19 May 2026

Cite this article as: Karami M, Farrokhi M. Concerns About Heterogeneity Assessment in Statistics of Meta-analyses; a Letter to the Editor. Arch Acad Emerg Med. 2026; 14(1): e15. https://doi.org/10.22037/aaem.v14i1.2722.

Dear Editor;

In recent years, the number of published meta-analyses has increased significantly. However, this rapid growth raises important considerations regarding the methodologies employed in these studies (1).

The statistical sections of most meta-analyses including diagnostic reviews typically assess heterogeneity using parameters such as the P-value from the Cochrane Q statistic test or the I^2 statistic (2-8). While these methods are standard, it is crucial to first evaluate heterogeneity based on the investigators' insights into the nature of the included studies (1, 9). Many meta-analyses incorporate studies that likely exhibit significant heterogeneity, particularly in areas such as epidemiological research. Therefore, before applying statistical methods to evaluate heterogeneity, investigators should make informed decisions about potential sources of heterogeneity and subsequently choose the appropriate approach to pool the effect measures using either fixed or random effects models (9).

Another important aspect of the statistical section in meta-analyses of diagnostic studies is the reporting of the specific models used. Many studies simply state that they employed a software or package without detailing the exact methodologies applied. Additionally, when utilizing software packages for meta-analysis, researchers may rely on default prompts without fully understanding their implications. This can lead to substantial discrepancies in results across different software platforms, often stemming from variations in statistical methods employed.

For example, consider analyzing the following data using the Meta-Disc software with the DerSimonian-Laird model (Random Effects Model) for pooling data, which is the default setting in this software (Table 1). The analysis yielded a sensitivity of 0.93, similar to the results obtained using the MIDAS package in Stata (bivariate random-effects model) (Figure 1). However, when estimating specificity with Meta-Disc, the result was 0.80, which is lower than the 0.88 obtained using the MIDAS package (Figure 1). In some cases, these differences can be more pronounced and may simultaneously affect both sensitivity and specificity. The comparison between DerSimonian-Laird model and bivariate random-effects model is provided in table 2.

Using a simulation approach, we generated 100 diagnostic studies, with 10 studies included in each of 10 meta-analyses. Five of these meta-analyses were designed to have low heterogeneity, and the remaining five had high heterogeneity. The findings showed that diagnostic estimates were approximately identical between Meta-DiSc and the MIDAS package in Stata under conditions of low heterogeneity (Table3). However, when meta-analyses were conducted on highly heterogeneous studies, noticeable differences emerged in diagnostic parameters, including sensitivity, specificity, and the AUC of the SROC curve. Because diagnostic test accuracy meta-analyses typically involve substantial clinical and methodological variability, such as differences in index tests, reference standards, or study conditions, high heterogeneity is common. As a result, differences between MIDAS and Meta-DiSc can be expected when their default models are used.

Checklist for Diagnostic Meta-analysis (Best Practices for Heterogeneity Assessment and Statistical Method Selection):

A. Assessing heterogeneity in conceptual and methodological steps

• Begin with clinical and methodological judgment, not only statistics, to determine whether heterogeneity is expected based on study design, populations, index tests, and reference standards.

• Identify potential sources of heterogeneity before performing any quantitative analyses based on factors such as study design, imaging modality, diagnostic thresholds, patient characteristics.

***Corresponding Author:** Mehrdad Farrokhi; Department of Epidemiology, School of Public Health and Safety, Shahid Beheshti University of Medical Sciences, Tehran, Iran. Email: drfarrokhi2024@gmail.com, Tel: +989365035851, ORCID: https://orcid.org/0000-0002-1559-2323.

• Avoid relying solely on I^2 or the Q-test to determine the presence or relevance of heterogeneity.
• Use I^2, T^2, and Q-test results as supportive, not decisive, evidence when determining heterogeneity.
• Recognize that diagnostic accuracy meta-analyses almost always involve substantial heterogeneity due to differences in index tests, reference standards, and study conditions.

B. Choosing statistical models

• Select the effect-pooling model based on anticipated heterogeneity, not on statistical thresholds alone.
• Use random-effects models by default when clinical or methodological heterogeneity is expected.
• Justify the choice of fixed or random effects explicitly in the Methods section.
• Avoid blindly accepting software defaults; understand what statistical model the software applies.

C. Reporting statistical methods transparently

• Specify the exact statistical model used, not only the software name (e.g., DerSimonian–Laird, bivariate random effects, HSROC, GLMM).
• Document all software packages and versions (e.g., MetaDiSc vX.X, Stata MIDAS).
• State whether default settings were used or if parameters were manually modified.
• Report heterogeneity metrics (I^2, T^2, Q-test) for both sensitivity and specificity in diagnostic meta-analyses.

D. Addressing cross-software differences

• Recognize that different meta-analysis software may use different statistical models, leading to discrepancies in pooled results.
• Validate key findings across multiple methods or software platforms when feasible.
• Interpret differences between software outputs in the context of heterogeneity, especially in highly variable datasets.

1. Conclusion

In summary, clearly articulating the approach used to evaluate heterogeneity and specifying the exact statistical methods and software utilized in conducting meta-analyses can greatly enhance readers' ability to interpret and apply the findings effectively. By addressing these methodological concerns, we can improve the quality and reliability of meta-analyses in the scientific literature.

2. Declarations

2.1. Acknowledgments

Not applicable.

2.2. Authors' contributions

All authors contributed to conceptualization, writing the original draft, review and editing the final version of the manuscript.

2.3. Funding/Support

None.

2.4. Conflict of interest

None.

2.5. Data Availability

Not applicable.

2.6. Consent for Publication

All authors provided consent for the publication of this article.

2.7. Ethics Approval and Consent to Participate

Not applicable.

2.8. Using Artificial Intelligence Chatbots

We used AI-powered tools to assist with grammar checking and to enhance the academic quality of the manuscript. The text was primarily written by the authors, and the final version was thoroughly reviewed and approved by all authors.

References

1. Stogiannis D, Siannis F, Androulakis E. Heterogeneity in meta-analysis: a comprehensive overview. Int J Biostat. 2024;20(1):169-99.
2. Farrokhi M, Motavaselian M, Jafari Khouzani P, Moghadam Fard A, Daeizadeh F, Pourrahimi M, et al. Diagnostic Performance of Ultrasonography for Identification of Small Bowel Obstruction: A Systematic Review and Meta-analysis. Archives of Academic Emergency Medicine. 2024;12(1):e33.
3. Rahmani E, Farrokhi M, Farrokhi M, Nouri S, Moghadam Fard A, Hoorshad B, et al. Accuracy of Lung Ultrasonography for Diagnosis of Heart Failure; a Systematic Review and Meta-analysis. Archives of Academic Emergency Medicine. 2025;13(1):e33.
4. Goodarzy B, Rahmani E, Farrokhi M, Tavakoli R, Moghadam Fard A, Ranjbaran Ghaleh M, et al. Diagnostic Value of Chest Computed Tomography Scan for Identification of Foreign Body Aspiration in Children: A Systematic Review and Meta-analysis. Archives of Academic Emergency Medicine. 2024;13(1):e3.
5. Kazemi S-M, Khorram R, Fayyazishishavan E, Amani-Beni R, Haririan Y, Hosseini Khameneh SM, et al. Diagnostic Accuracy of Ottawa Knee Rule for Diagnosis of Fracture in Patients with Knee Trauma; a Systematic Review and Meta-analysis. Archives of Academic Emergency Medicine. 2023;11(1):e30.
6. Tabatabaei Nodoushan MH, Rahmani M, Cheraghi S, Farrokhi M. Role of D-dimer in the Identification of Periprosthetic Infections: A Systematic Review and Meta-analysis. Journal of Reviews in Medical Sciences. 2024;4(1):1-17.

7. Amiri T, Farrokhi M. Efficacy of Fetal MRI to Detect Cleft Palate: A Systematic Review and Meta-Analysis. Journal of Reviews in Medical Sciences. 2021;1(3):152-62.
8. Hosseini Khameneh SM, Amani-Beni R, Ahadiat S-A, Kahrizi MS, Jafari S, Seyedinnavade S, et al. Diagnostic Accuracy of Ultrasonography for Identification of Elbow Fractures in Children; a Systematic Review and Meta-analysis. Archives of Academic Emergency Medicine. 2023;11(1):e62.
9. Borenstein M. Introduction to meta-analysis. Chichester, U.K.: John Wiley & Sons; 2009. xxviii, 421 p. p.
10. Motavaselian M, Farrokhi M, Jafari Khouzani P, Moghadam Fard A, Daeizadeh F, Pourrahimi M, et al. Diagnostic Performance of Ultrasonography for Identification of Small Bowel Obstruction; a Systematic Review and Meta-analysis. Arch Acad Emerg Med. 2024;12(1):e33.

Table 1: Extracted Data for Diagnostic Analysis from a Meta-Analysis Previously Published in the Archives of Academic Emergency Medicine (10)

Study ID	True Positive	False Positive	False Negative	True Negative
Study1	17	4	1	2
Study2	25	0	5	2
Study3	15	6	2	12
Study4	24	1	2	16
Study5	23	0	5	16
Study6	21	1	3	25
Study7	44	3	1	12
Study8	33	1	2	20
Study9	51	0	4	15
Study10	12	0	5	27
Study11	60	4	7	6
Study12	18	1	1	76
Study13	26	6	3	66
Study14	77	5	4	23
Study15	17	36	4	83
Study16	64	26	0	43
Study17	82	11	1	56
Study18	84	6	2	76
Study19	82	57	11	67
Study20	147	10	5	67

Data are presented as number.

Table 2: Comparison of DerSimonian-Laird model (Random Effects Model) and the bivariate random-effects model for pooling data

Feature	DerSimonian-Laird (DL)	Bivariate Random-Effects (BRM)
Outcomes Analyzed	Single effect measure (univariate), e.g., an Odds Ratio for a treatment effect.	Two correlated effect measures (bivariate), e.g., sensitivity and specificity of a diagnostic test.
Core Principle	A Method of Moments approach to estimate the between-study variance (T^2). Non-iterative.	A Hierarchical/Multivariate model, estimated via iterative methods like Maximum Likelihood (ML) or Restricted Maximum Likelihood (REML).
T2 Estimation & CI	DL $T\hat{2}$ is Downwardly Biased (inefficient), especially with a small number of studies. This underestimation of heterogeneity can lead to anti-conservative (too narrow) confidence intervals for the pooled effect.	Generally, provides less biased and more efficient estimates of the full variance-covariance structure, leading to appropriately-sized confidence intervals.
Pooled Effect Estimation	Asymptotically consistent for the pooled effect ($\hat{\mu}$). It is computationally simple and widely implemented.	Asymptotically consistent for the two pooled effects. It leverages the correlation between outcomes, often leading to better precision for the joint estimates compared to two separate univariate models.
Variance-Covariance Structure	Ignores any correlation between multiple outcomes from the same study (cannot estimate it). Assumes independence between studies.	Robustly estimates the full between-study variance-covariance matrix, including two heterogeneity terms $T\_1\hat{2}$,$T\_2\hat{2}$ and the inter-outcome correlation ().
Application	General meta-analysis of a single, well-defined effect size.	Primarily used for Diagnostic Test Accuracy (DTA) meta-analysis (e.g., using the Hierarchical Summary ROC (HSROC) or BRM framework).
Statistical Model	Random-effects model: Yi=μ+ui+i where uiN(0,T2).	Random-effects model: (Y1iY2i) = (μ1μ2) + (u1iu2i) + (1i2i) where the random effects ui are correlated.

Table 3: Diagnostic results of simulation of 10 meta-analyses

Meta-analysis	Result*	Sensitivity		Specificity		AUC-SROC	
		MetaDisc	Stata	MetaDisc	Stata	MetaDisc	Stata
1	Low	0.86(0.84-0.88)	0.86(0.84-0.88)	0.88(0.86-0.90)	0.89(0.86-0.90)	0.93	0.93(0.90-0.95)
2	Low	0.85(0.82-0.87)	0.85(0.82-0.87)	0.91(0.89-0.92)	0.91(0.89-0.92)	0.92	0.94(0.92-0.96)
3	Low	0.86(0.84-0.88)	0.86(0.84-0.88)	0.90(0.88-0.92)	0.90(0.88-0.92)	0.94	0.94(0.92-0.96)
4	Low	0.84(0.82-0.86)	0.84(0.82-0.86)	0.89(0.87-0.91)	0.89(0.86-0.91)	0.91	0.89(0.86-0.91)
5	Low	0.86(0.83-0.88)	0.86(0.83-0.88)	0.89(0.87-0.91)	0.89(0.87-0.91)	94	0.94(0.91-0.96)
Mean	Low	0.854	0.854	0.894	0.896	92.8	92.8
6	High	0.84(0.82-0.86)	0.86(0.78-0.91)	0.88(0.85-0.90)	0.88(0.85-0.91)	0.94	0.93(0.90-0.95)
7	High	0.83(0.81-0.85)	0.83(0.80-0.87)	0.90(0.88-0.92)	0.93(0.87-0.96)	0.92	0.90(0.88-0.93)
8	High	0.84(0.81-0.86)	0.88(0.81-0.87)	0.88(0.85-0.90)	0.88(0.85-0.91)	0.93	0.90(0.87-0.92)
9	High	0.79(0.76-0.82)	0.84(0.71-0.92)	0.90(0.88-0.92)	0.92(0.86-0.96)	0.95	0.95(0.93-0.97)
10	High	0.83(0.80-0.85)	0.83(0.78-0.87)	0.85(0.83-0.87)	0.88(0.82-0.93)	0.91	0.91(0.89-0.94)
Mean	High	0.826	0.848	0.882	0.898	93	91.8

Data are presented with 95% confidence interval. *: Heterogeneity.

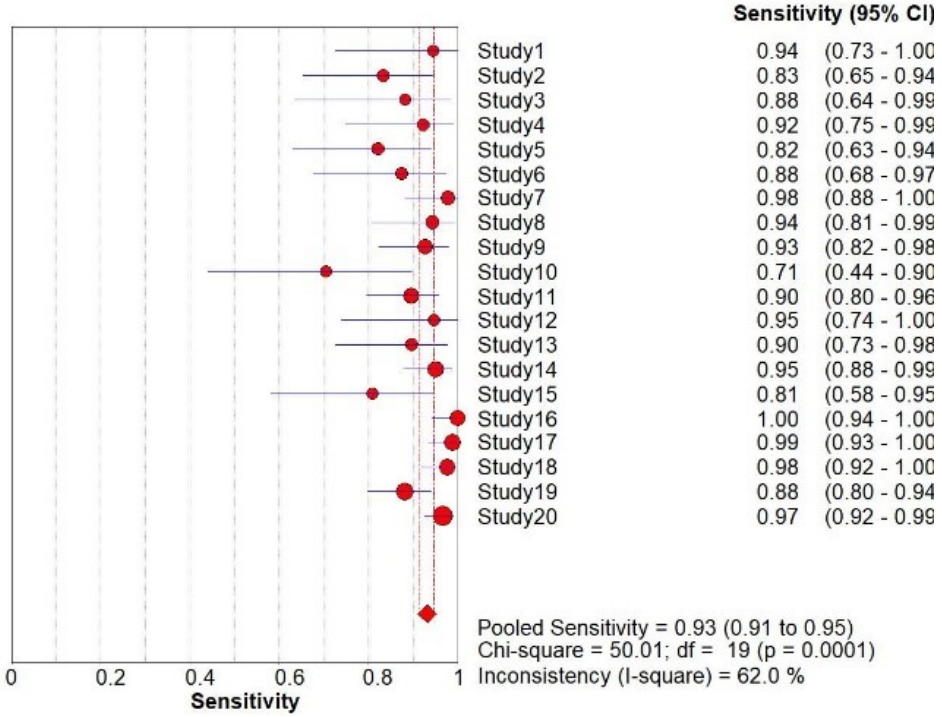

Figure 1: Estimation of sensitivity (left column) and specificity (right column) using DerSimonian-Laird model of Meta-Disc software and MIDAS package of STATA software (bivariate random-effects model).