Diagnostic Test Accuracy Meta-Analysis: A Practical Guide to Hierarchical Models

Javier Arredondo Montero *

Posted Date: 30 June 2025

doi: 10.20944/preprints202506.2461.v1

Keywords: Diagnostic test accuracy; Meta-analysis; Bivariate; BRMA; Hierarchical model; HSROC; Threshold effect; Meta-regression; Fagan nomogram; Moses-Littenberg; STATA; MetaDisc; RevMan

 doi:10.20944/preprints202506.2461.v1

Article

Diagnostic Test Accuracy Meta-Analysis: A Practical Guide to Hierarchical Models

Javier Arredondo Montero

Department of Pediatric Surgery, Complejo Asistencial Universitario de León. c/Altos de Nava s/n, 24008 León, Castilla y León, Spain; jarredondo@saludcastillayleon.es or javier.montero.arredondo@gmail.com; Tel.: +34 987 23 74 00

Abstract

Background: Accurate evaluation of diagnostic tests is essential to guide clinical decision-making, particularly in surgical practice. Systematic reviews and meta-analyses of diagnostic test accuracy (DTA) are key for evidence synthesis; however, traditional approaches, including univariate pooling or simplified summary ROC (SROC) models such as the Moses–Littenberg method, often yield biased and clinically misleading estimates. **Methods**: An extensive literature review was conducted to synthesize current evidence on hierarchical random-effects models for DTA meta-analysis. Based on this framework, a simulated dataset was generated, and a comprehensive meta-analysis was performed with a didactic and explanatory focus. The analysis illustrates key methodological concepts, interpretation of model outputs, and the use of complementary tools, including likelihood ratios, scattergrams, meta-regression, publication bias assessment, and outlier detection. **Results**: The traditional meta-analysis performed with MetaDisc, applying the DerSimonian–Laird and Moses–Littenberg methods, produced separated sensitivity and specificity pooled estimates with artificially narrow confidence intervals and a symmetric, theoretical SROC curve extrapolated beyond the observed data range, ignoring threshold variability and underestimating between-study heterogeneity. In contrast, the hierarchical random-effects model provided more realistic and clinically interpretable estimates. Joint modeling of sensitivity and specificity revealed substantial between-study variability, a strong negative correlation consistent with a threshold effect, and wider, asymmetric confidence intervals that accurately reflected uncertainty. Influence diagnostics identified outliers and highly influential studies that distorted the traditional model results. **Conclusions**: Promoting the correct application and interpretation of hierarchical models in DTA meta-analyses is essential to ensure high-quality, reliable, and scientifically robust evidence.

Keywords: diagnostic test accuracy; meta-analysis; bivariate; BRMA; hierarchical model; HSROC; threshold effect; meta-regression; fagan nomogram; moses-Littenberg; STATA; MetaDisc; RevMan

Introduction

Accurate diagnostic information is fundamental to surgical decision-making. Surgeons routinely rely on diagnostic tests to guide critical aspects such as patient selection for liver transplantation using non-invasive fibrosis assessment, detection of anastomotic leakage after colorectal surgery using inflammatory biomarkers, or diagnosis of acute appendicitis through clinical scoring systems. In these scenarios, imaging modalities, laboratory tests, and risk scores directly influence surgical indications, timing, and perioperative management.

Systematic reviews and meta-analyses of diagnostic test accuracy (DTA) play a crucial role in summarizing available evidence, informing the development of clinical guidelines, and shaping surgical practice. However, despite their growing importance, the methodological quality of DTA meta-analyses in surgical research remains inconsistent. Many published reviews still rely on outdated or overly simplistic statistical approaches, such as the separate pooling of sensitivity and

 doi:10.20944/preprints202506.2461.v1

specificity using basic random-effects models [1,2]. These methods frequently ignore key factors, including threshold effects, between-study heterogeneity, and the inherent trade-off between sensitivity and specificity. Consequently, they often produce biased, overly precise, or clinically misleading estimates that fail to capture the true diagnostic performance of tests in complex surgical settings.

To overcome these limitations, advanced statistical models have been developed over the past two decades and are now considered the methodological gold standard for DTA meta-analysis. Specifically, hierarchical models—namely, the hierarchical summary receiver operating characteristic (HSROC) model proposed by Rutter and Gatsonis [3], and the bivariate random-effects model (BRMA) introduced by Reitsma et al.[4]—offer a robust analytical framework. These models account for both within- and between-study variability, explicitly model the correlation between sensitivity and specificity, and incorporate threshold effects, providing more reliable and clinically interpretable estimates of diagnostic performance.

Despite their clear methodological advantages and endorsement by organizations such as the Cochrane Collaboration [5], the application of hierarchical models in surgical diagnostic research remains inconsistent. Barriers include limited statistical expertise, lack of familiarity with these methods, and restricted access to suitable software. As a result, diagnostic meta-analyses in surgery often produce oversimplified or misleading estimates, reducing their relevance for clinical decision-making.

The objective of this article is to provide surgeons and clinical researchers with a practical and accessible introduction to hierarchical models for DTA meta-analysis, with a focus on their application in surgical research. In addition to explaining model structures and interpretation, this review highlights complementary analytical tools such as Fagan nomograms, likelihood ratio scatterplots, and meta-regression to explore heterogeneity. The aim is to promote the correct application of hierarchical methods, improving the quality, precision, and clinical relevance of DTA meta-analyses in surgery.

Understanding Basic Diagnostic Test Accuracy Metrics: Beyond Sensitivity and Specificity

The diagnostic performance of a test is typically summarized using sensitivity and specificity. Sensitivity reflects the ability of a test to correctly identify individuals with the disease, while specificity reflects its ability to classify individuals without the disease correctly. Although these metrics are often presented as intrinsic test characteristics, they are directly influenced by the diagnostic threshold or cut-off applied to define a positive result [6,7].

In surgical research, diagnostic thresholds are commonly based on biomarker levels, imaging findings, or clinical scores. For example, in the diagnosis of acute appendicitis, tools such as the Alvarado score or the AIR score apply predefined cut-offs to stratify patients into low, intermediate, or high-risk groups. Similarly, in postoperative surveillance for colorectal cancer, carcinoembryonic antigen (CEA) thresholds guide follow-up imaging and trigger further investigation for suspected recurrence. Adjusting these thresholds inevitably affects diagnostic performance: lowering the threshold increases sensitivity at the expense of specificity, and vice versa.

A typical example in surgical research is the use of C-reactive protein (CRP) to detect postoperative complications, such as anastomotic leakage. Studies applying lower CRP thresholds often report high sensitivity (e.g., 90%) but low specificity (e.g., 50%), resulting in frequent false positives and unnecessary interventions. Conversely, studies using higher CRP thresholds may achieve greater specificity (e.g., 85%) but reduced sensitivity (e.g., 65%), increasing the risk of missed diagnoses. These trade-offs reflect the so-called threshold effect, an inverse relationship between sensitivity and specificity driven by differences in positivity thresholds across studies [8].

The receiver operating characteristic (ROC) analysis is a statistical tool used to evaluate the diagnostic performance of a test across all possible thresholds. It consists of plotting the true positive rate (sensitivity) on the y-axis against the false positive rate (1-specificity) on the x-axis for every

 doi:10.20944/preprints202506.2461.v1

potential cut-off value that defines a positive result. This produces a curve that visually illustrates the trade-off between sensitivity and specificity as the threshold varies.

The area under the ROC curve (AUC) provides a single, global summary measure of the test's overall ability to discriminate between individuals with and without the disease. An AUC of 1 represents a perfect test that correctly classifies all individuals, while an AUC of 0.5 reflects a non-informative test equivalent to random chance [9–11]. Although frequently reported, the AUC is a threshold-independent metric and does not reflect how the test performs at specific, clinically relevant decision points. For this reason, reporting sensitivity, specificity, and likelihood ratios at predefined thresholds remains essential to complement the information provided by the AUC.

Positive and negative predictive values (PPV and NPV) quantify the probability that a patient has or does not have the disease based on a positive or negative test result [12]. However, both are heavily dependent on disease prevalence in the evaluated population. Applying a diagnostic test indiscriminately to low-risk surgical populations may result in low PPV, generating false positives and unnecessary interventions. Conversely, restricting testing to high-risk populations, such as patients with clinical suspicion or elevated inflammatory markers, increases PPV and improves test utility.

Likelihood ratios (LRs) offer a more robust, prevalence-independent measure of diagnostic performance [13,14]. The positive likelihood ratio (LR+) quantifies how much more likely a positive result is in a patient with the disease compared to one without it. The negative likelihood ratio (LR−) expresses how much less likely a negative result is in a diseased versus a non-diseased individual. Generally, an LR+ greater than 10 provides strong evidence to confirm disease, while an LR− below 0.1 provides strong evidence to exclude it. However, such extreme values are uncommon in surgical diagnostics.

For instance, clinical scores for acute appendicitis, such as the Alvarado score, often demonstrate high AUC values (e.g., 0.86). However, their LR estimates reveal only modest diagnostic performance. In pediatric populations, pooled analyses report an LR+ of approximately 2.4 and an LR− of around 0.28, indicating that while the test contributes to diagnostic reasoning, it cannot confidently rule in or rule out appendicitis [15].

The Diagnostic Odds Ratio: Interpretation and Limitations

The diagnostic odds ratio (DOR) is frequently reported in systematic reviews and meta-analyses as a global indicator of test performance. Mathematically, it represents the odds of a positive test result in individuals with the disease compared to those without it and is calculated as:

$$DOR = (TP \times TN) / (FP \times FN)$$

Higher DOR values indicate better overall discriminatory capacity, with a value of 1 reflecting no diagnostic ability beyond chance [16]. Despite its statistical appeal, the DOR has considerable limitations in clinical practice. First, it combines sensitivity and specificity into a single summary metric but does not differentiate their contributions, masking potential trade-offs relevant to clinical decision-making.

This limitation is particularly relevant in surgical contexts, where priorities often differ. For example, maximizing sensitivity is crucial when diagnosing postoperative complications, such as anastomotic leakage, where false negatives pose a significant clinical risk. Conversely, maximizing specificity is desirable when considering reoperation for suspected bile duct injury, where false positives may lead to unnecessary surgical interventions. The DOR, by providing a global estimate, does not capture these nuances or reflect threshold-dependent variations in test performance.

Furthermore, the DOR lacks intuitive clinical interpretation. Unlike sensitivity, specificity, or likelihood ratios, it does not directly translate into probabilities relevant for bedside decision-making. Relying exclusively on the DOR risks oversimplifying diagnostic accuracy, potentially leading to an overestimation of a test's clinical utility.

 doi:10.20944/preprints202506.2461.v1

It is essential to distinguish between the DOR and the conventional odds ratio (OR), which is often used in epidemiological studies. While both are mathematically similar in structure, their interpretation differs substantially. The OR is primarily a measure of the association between exposure and outcome, rather than diagnostic performance. As highlighted by Pepe et al., a statistically significant OR does not necessarily translate into clinically useful diagnostic accuracy [17]. Therefore, relying solely on OR-based associations can lead to overestimation of a test's real-world diagnostic utility.

While the DOR can complement other metrics in research settings, it should not be considered the primary indicator for evaluating diagnostic performance in surgical practice. Reporting sensitivity, specificity, likelihood ratios, and post-test probabilities remains essential to provide clinically meaningful, context-specific information.

Which Metrics Should Be Reported in Diagnostic Test Accuracy Meta-Analyses

Given these considerations, reporting standards for DTA meta-analyses should prioritize clinically interpretable and statistically robust parameters. The core recommended metrics include pooled sensitivity and specificity with their corresponding 95% confidence intervals, as well as the LR+ and LR− with 95% confidence intervals, all derived from appropriate hierarchical models. Additionally, the summary receiver operating characteristic (HSROC) curve and its 95% confidence and prediction regions should be reported to provide a comprehensive graphical overview of diagnostic performance across thresholds [5]. Although the DOR can be included as a secondary, global indicator of discrimination, it should never replace the separate reporting of sensitivity, specificity, or likelihood ratios, given its limitations for clinical interpretation. The area under the ROC curve (AUC) derived from HSROC modeling provides additional insight but requires cautious interpretation. Some methods estimate the AUC only within the observed range of specificities, potentially underestimating the true area. In contrast, others extrapolate beyond the available data (such as *midas*), which may lead to overestimating performance if the extrapolation is unreliable. Therefore, it is essential to report how the AUC was obtained and to avoid overinterpreting its clinical relevance.

It is also important to consider that confidence intervals can vary depending on the statistical approach applied. Methods such as Wald or Clopper-Pearson differ in their assumptions and precision, with exact approaches typically yielding wider, more conservative intervals that better capture true uncertainty, especially in small-sample settings or when extreme proportions are present.

Why Hierarchical Models Are Needed in Diagnostic Test Accuracy Meta-Analyses

Diagnostic research, especially in surgical settings, is inherently heterogeneous. Patient populations differ in demographics, comorbidities, and disease prevalence, all of which influence diagnostic test performance. For example, the accuracy of appendicitis scoring systems varies substantially between pediatric and adult cohorts due to differences in symptom presentation and baseline risk. Similarly, the reference standards used to define disease presence or absence often differ across studies, introducing inconsistency in the classification of true positives and true negatives. A relevant example is the use of varying radiological criteria to confirm bile duct injury after cholecystectomy, contributing to discrepancies in reported test performance.

Diagnostic tools themselves—including imaging modalities, biomarker assays, and clinical scores—often differ in technical specifications, operator expertise, and interpretation criteria. For instance, the diagnostic performance of ultrasonography for acute cholecystitis is highly dependent on the operator's skill and the quality of the equipment, particularly for subtle findings such as gallbladder wall thickening or pericholecystic fluid. Likewise, the diagnostic utility of biomarkers such as C-reactive protein or procalcitonin for detecting postoperative complications varies

 doi:10.20944/preprints202506.2461.v1

depending on the laboratory methods used, including ELISA, high-sensitivity assays, chemiluminescence, or point-of-care devices.

Beyond these clinical and technical factors, methodological differences across studies—including study design (prospective versus retrospective), sample size, risk of bias, and the selection or post hoc adjustment of diagnostic thresholds—further amplify heterogeneity in observed sensitivity and specificity.

A critical contributor to heterogeneity is the threshold effect: variations in the positivity threshold directly alter the trade-off between sensitivity and specificity. Lowering the threshold increases sensitivity but reduces specificity, while more restrictive thresholds have the opposite effect. Conventional intervention-designed meta-analytical approaches, such as the DerSimonian-Laird method, pool sensitivity and specificity separately, ignoring their inverse relationship and failing to capture the complexity introduced by threshold effects.

Earlier DTA models, such as the Moses-Littenberg approach [18], introduced simplified summary ROC curves but lacked the statistical rigor to properly model threshold variability, between-study heterogeneity, or the intrinsic correlation between sensitivity and specificity. These models often produce biased, overly precise, or clinically misleading estimates that do not reflect real-world diagnostic scenarios. Despite these limitations, they remain in use, largely due to reliance on outdated software platforms with limited analytical capabilities.

To address these challenges, hierarchical models have been developed and are now recognized as the methodological gold standard for DTA meta-analysis. These models explicitly account for threshold variability, model the correlation between sensitivity and specificity, and incorporate both within- and between-study heterogeneity. Their application generates more reliable, clinically interpretable, and generalizable estimates of diagnostic performance—critical in complex and heterogeneous fields such as surgical research [19,20].

Hierarchical Models for DTA Meta-Analyses: Structure and Parameterization

To overcome the limitations of conventional DTA meta-analytical approaches, two hierarchical random-effects models have become the methodological standard: the HSROC model proposed by Rutter and Gatsonis [3], and the BRMA introduced by Reitsma et al. [4]. Both models share the same statistical foundation, enabling the joint synthesis of sensitivity and specificity while accurately modeling their correlation and between-study heterogeneity.

The HSROC model assumes that each included study reflects an underlying ROC curve, with variability in diagnostic thresholds across studies contributing to the observed heterogeneity. Using a specific parameterization—Λ (overall diagnostic accuracy), Θ (threshold parameter), β (slope of the curve), σ^2_a (between-study variance for accuracy), and σ^2_{th} (between-study variance for threshold)—the HSROC model estimates a global summary ROC curve that captures the trade-off between sensitivity and specificity across varying thresholds. The HSROC curve provides both a visual and quantitative synthesis of overall test performance, including confidence regions that reflect statistical uncertainty. This model is particularly appropriate when included studies use different positivity thresholds, making the estimation of a summary ROC curve more informative than a single point estimate.

In contrast, the BRMA model focuses on directly modeling sensitivity and specificity by jointly analyzing their logit-transformed values. The logit (log-odds) transformation stabilizes variance and linearizes proportions, ensuring coherent modeling of outcomes constrained between 0 and 1. By modeling logit-sensitivity and logit-specificity together, the BRMA framework preserves their natural correlation and accounts for both within-study variability and between-study heterogeneity. This structure inherently reflects the sensitivity–specificity trade-off and quantifies their statistical correlation ($\varrho_{a\beta}$), which is typically negative, consistent with the inverse relationship between these parameters. The parameterization of the BRMA model includes μ_a (mean logit-sensitivity), μ_β (mean logit-specificity), σ^2_a (between-study variance for sensitivity), σ^2_β (between-study variance for specificity), and $\varrho_{a\beta}$ (correlation between sensitivity and specificity). The BRMA model is particularly

 doi:10.20944/preprints202506.2461.v1

suitable when all included studies apply the same diagnostic threshold, making the estimation of a single pooled summary point both feasible and clinically meaningful. In such contexts, the summary point provides a concise and interpretable estimate of average test performance that is more informative than a global ROC curve, especially when threshold variability is minimal or absent.

Both HSROC and BRMA models are statistically robust, mathematically coherent, and produce comparable results under most conditions, particularly in the absence of covariates or meta-regression. Their use is endorsed by leading methodological authorities, including the Cochrane Handbook for DTA Reviews [5], as essential tools for generating valid, reliable, and clinically interpretable estimates of diagnostic accuracy.

Despite their shared statistical foundation, confusion in terminology persists within the literature. Terms such as "bivariate model," "hierarchical model," and "HSROC model" are frequently used interchangeably, often without clarifying their conceptual or methodological distinctions. Although both models derive from the same hierarchical structure and jointly account for sensitivity, specificity, and heterogeneity, they differ in parameterization and interpretative focus. Recognizing these differences is essential to select the most appropriate model based on the characteristics of the included studies and the specific objectives of the meta-analysis.

Selecting the Appropriate Statistical Software for Diagnostic Accuracy Meta-Analysis

The correct application of hierarchical models in DTA meta-analysis depends not only on conceptual understanding but also on the availability of appropriate statistical software. Although HSROC and BRMA models are considered the methodological gold standard, many commonly used platforms either lack support for these models or present substantial limitations regarding functionality, graphical outputs, or ease of use [21].

Several software solutions are available for DTA meta-analysis, including Meta-DiSc [22], RevMan [23], R [24], SAS [25], and Stata [26]. These platforms differ significantly in licensing costs, user accessibility, and the level of statistical expertise required. Meta-DiSc (version 1.4) [22] and RevMan (version 5.4) [23] remain widely adopted due to their free availability and intuitive interfaces. However, they do not support hierarchical modeling. Consequently, they rely on outdated approaches such as separate pooling of sensitivity and specificity with classic random-effects models (e.g., DerSimonian-Laird) or simplified SROC curves based on the Moses-Littenberg method. Although the Moses-Littenberg model partially accounts for the sensitivity–specificity correlation, it employs a basic linear regression framework that neither models threshold variability nor adequately reflects between-study heterogeneity, resulting in biased or overly optimistic results.

In contrast, advanced platforms such as Stata and R [24,26] facilitate the correct application of hierarchical models, offering more rigorous, clinically meaningful analyses. R, as an open-source statistical environment, provides extensive resources for DTA meta-analysis, including BRMA and HSROC models, meta-regression, and advanced visualizations. However, its implementation requires intermediate to advanced programming skills, which can limit accessibility for researchers without formal statistical training.

Stata (StataCorp LLC, College Station, TX) [26] provides a balanced alternative, combining extensive analytical capabilities with greater user accessibility. Within Stata, three commands—*metandi*, *midas*, and *metadta* [27–30]—have been developed for DTA meta-analysis, each implementing the BRMA model with varying functionalities:

- *metandi* [27] applies both the HSROC and BRMA models explicitly, generating a summary HSROC curve, pooled sensitivity and specificity estimates, and associated confidence and prediction regions. However, as the earliest Stata command developed for this purpose, *metandi* presents analytical and graphical limitations. For example, it does not support meta-regression and offers fewer options for graphical customization.
- *Midas [28]*, although more limited in some aspects, remains a complementary option for hierarchical modeling. It is particularly noted for its user-friendly integration of exploratory

 doi:10.20944/preprints202506.2461.v1

tools, including heterogeneity plots, goodness-of-fit assessments, Fagan nomograms, likelihood ratio scattergrams, and publication bias evaluation via Deeks' regression test. Unlike *metandi* and *metadta*, which generate the HSROC curve based on the Rutter and Gatsonis parameterization, *midas* derives its ROC curve indirectly from the BRMA output, providing a graphical approximation with reduced methodological rigor. It also estimates the area under the curve (AUC) with its 95% confidence interval. However, empirical evidence suggests that these AUC estimates may be biased, particularly when not derived from a formally parameterized HSROC model. Additionally, *midas* allows for meta-regression, although limited to univariable analyses.

- *metadta* [29,30] is the most modern and versatile Stata command for DTA meta-analysis. Based on the BRMA framework, it offers extensive functionalities, including bivariate I^2 estimation (Zhou et al.), advanced meta-regression capabilities, and highly customizable graphical outputs. It is currently the preferred tool for conducting methodologically rigorous DTA meta-analyses within the Stata environment.

In summary, while multiple software options exist for DTA meta-analysis, only a subset enables appropriate hierarchical modeling aligned with current methodological standards. Stata provides a robust, accessible environment, with *metadta* [29,30] representing the most comprehensive option for producing clinically meaningful, transparent, and statistically sound syntheses of diagnostic performance.

Heterogeneity in Diagnostic Test Accuracy Meta-Analyses

The assessment of heterogeneity is a critical component of DTA meta-analysis, yet it presents unique methodological challenges compared to intervention studies. It is essential to differentiate between heterogeneity attributable to threshold effects—where variations in positivity thresholds across studies directly modify the sensitivity–specificity balance—and heterogeneity unrelated to threshold, which reflects genuine differences in diagnostic performance due to factors such as population characteristics, disease spectrum, or methodological variability.

Traditional heterogeneity metrics widely used in intervention meta-analyses, such as the I^2 statistic proposed by Higgins and Thompson or Cochran's Q test [31], are inadequate in DTA settings [5]. These approaches fail to account for the intrinsic correlation between sensitivity and specificity driven by threshold effects, often resulting in exaggerated or misleading estimates of between-study variability. To address this limitation, Zhou et al. proposed a bivariate I^2 statistic [32], which jointly considers the variability and correlation of sensitivity and specificity within the hierarchical framework. This approach provides a more accurate and interpretable estimation of residual heterogeneity and is implemented in metadta [29,30], currently considered an accepted standard for quantifying heterogeneity in DTA meta-analyses.

As a preliminary exploratory step, many DTA meta-analyses report the Spearman correlation coefficient (ϱ), which evaluates the relationship between sensitivity and specificity across studies. A strong negative correlation typically suggests the presence of threshold effects, whereby differences in positivity thresholds impact the trade-off between sensitivity and specificity [5]. While informative as an initial indicator, Spearman's ϱ provides only a unidimensional approximation and does not capture the complex, multidimensional structure of heterogeneity inherent to hierarchical models.

In addition to global heterogeneity estimates, BRMA models report variance components for sensitivity (σ^2_a) and specificity (σ^2_β) individually. These parameters quantify the between-study variability for each measure that is not explained by within-study precision, offering insight into the consistency and reproducibility of diagnostic performance. Larger variance estimates reflect greater heterogeneity and should be considered when interpreting the reliability of results.

Finally, prediction regions derived from BRMA models offer a practical and graphical representation of heterogeneity. These regions define the expected range within which true sensitivity and specificity values from future studies are likely to fall, complementing confidence intervals and offering an intuitive visualization of between-study variability.

 doi:10.20944/preprints202506.2461.v1

Meta-Regression in Diagnostic Test Accuracy Meta-Analyses

Beyond global summary estimates, exploring sources of heterogeneity is essential to enhance the clinical interpretability and methodological rigor of DTA meta-analyses [5]. In this context, meta-regression represents a key analytical extension of hierarchical models, enabling the investigation of study-level covariates that may influence diagnostic performance.

Meta-regression evaluates whether variability in sensitivity, specificity, or overall test performance can be explained by factors such as study design (e.g., prospective vs. retrospective), patient population (e.g., pediatric vs. adult), risk of bias domains (e.g., QUADAS-2 assessment), disease prevalence, or applied diagnostic thresholds. This approach facilitates the identification of methodological or clinical characteristics associated with differences in diagnostic performance across studies.

While univariable meta-regression provides an initial assessment of the effect of individual covariates, more advanced multivariable models allow for the simultaneous adjustment of multiple factors. These models provide a more comprehensive exploration of heterogeneity, but they require larger datasets to ensure statistical stability and minimize the risk of overfitting. Methodological guidelines generally recommend including at least 10 studies per covariate to avoid inflated type I error rates and spurious associations—an important consideration in DTA meta-analyses, where the number of available studies is often limited [5].

It is also essential to emphasize that meta-regression in DTA meta-analyses is inherently exploratory and observational. Associations detected cannot be interpreted as causal and may be influenced by residual confounding, ecological bias, or imbalance in covariate distribution.

Despite these limitations, when appropriately applied, meta-regression provides valuable insight into how diagnostic test performance varies across different clinical settings, populations, and methodological designs. This information is critical for refining evidence synthesis, guiding the interpretation of summary estimates, and informing future research priorities.

Publication Bias in Diagnostic Test Accuracy Meta-Analyses

Publication bias represents a well-recognized threat to the validity of meta-analyses, including those evaluating DTA. In this setting, publication bias occurs when studies reporting favorable diagnostic performance—such as high sensitivity, specificity, or DOR—are more likely to be published, inflating pooled estimates and potentially generating misleading conclusions.

Detecting publication bias in DTA meta-analyses presents specific methodological challenges. Classical tools such as funnel plots, Begg's test, and the Egger regression test—commonly applied in intervention research—are based on evaluating the relationship between study size and effect magnitude to identify asymmetry suggestive of publication bias. Funnel plots provide a visual assessment, while Begg's test uses a rank correlation approach, and the Egger test applies a regression framework to detect asymmetry formally. However, these methods assume a single, continuous effect size and independence between outcomes. These assumptions do not hold in diagnostic test accuracy meta-analyses due to the paired nature of sensitivity and specificity [33,34]. Consequently, the use of conventional methods such as Begg's or Egger's tests is inappropriate in DTA meta-analyses and may produce misleading results [5,35,36]. To address this limitation, Deeks' regression test was specifically developed for DTA settings. This method evaluates the association between study size and diagnostic performance by regressing the inverse square root of the effective sample size against the log-transformed diagnostic odds ratio (DOR). In this plot, smaller studies with exaggerated performance estimates tend to cluster asymmetrically. A statistically significant slope ($p < 0.1$) suggests the presence of small-study effects, which may indicate publication bias. Major methodological guidelines, including the Cochrane Handbook for Systematic Reviews of Interventions currently endorse Deeks' test. It is important to note that Deeks' test may lack statistical power when fewer than 10 studies are included, limiting its reliability in small datasets. As a result, findings from this test must be interpreted with caution in underpowered meta-analyses.

 doi:10.20944/preprints202506.2461.v1

In Stata, the *midas* command incorporates a dedicated subcommand (*pubbias*) that automates the implementation of Deeks' test and generates a corresponding graphical output to facilitate the assessment of asymmetry. In contrast, *metandi* and *metadta* do not natively incorporate this functionality, making it necessary to rely on complementary tools—such as *midas* or manual coding—to evaluate publication bias comprehensively.

It is also essential to recognize that publication bias represents only one potential source of bias in DTA meta-analyses. Conceptually, numerous other biases may influence the accuracy and generalizability of diagnostic test performance estimates. These include spectrum bias (arising when the study population does not represent the full disease spectrum), selection bias, partial verification bias, misclassification bias, information bias, and disease progression bias, among others. Most of these biases tend to overestimate diagnostic performance, although in specific contexts, they may also lead to underestimation [2]. Systematic consideration of these factors, alongside formal assessments of publication bias, is critical to ensure a reliable, transparent, and clinically meaningful synthesis of diagnostic evidence.

Complementary Tools for Interpreting DTA Meta-Analyses: Fagan Nomograms, Scatterplots, and Beyond

Beyond global summary estimates and HSROC curves, several complementary graphical and analytical tools enhance the clinical interpretability of DTA meta-analyses. These resources bridge the gap between complex statistical outputs and practical decision-making, particularly in surgical and high-stakes clinical scenarios.

The Fagan (Bayesian) nomogram remains one of the most widely used tools for translating diagnostic performance into clinically relevant terms [37]. This visual aid illustrates how pre-test probability, likelihood ratios (LR+ and LR−), and post-test probability interact. By applying likelihood ratios derived from a meta-analysis to an estimated pre-test probability, clinicians can approximate the probability of disease after a positive or negative test result. This facilitates risk stratification and improves diagnostic reasoning in everyday practice.

Additionally, likelihood ratio scatterplots offer an intuitive graphical representation of global diagnostic performance [27]. Unlike ROC scatterplots, which focus on sensitivity and specificity pairs, these plots visualize the distribution of LR+ and LR− across individual studies, highlighting patterns of variability, outliers, or subgroup effects. The likelihood ratio scattergram, as described by Stengel et al., defines four quadrants of informativeness based on evidence-based thresholds:

1. **Upper Left Quadrant**: LR+ < 10, LR− < 0.1 — diagnostic exclusion only
2. **Upper Right Quadrant**: LR+ > 10, LR− < 0.1 — both exclusion and confirmation
3. **Lower Right Quadrant**: LR+ > 10, LR− > 0.1 — diagnostic confirmation only
4. **Lower Left Quadrant**: LR+ < 10, LR− > 0.1 — neither exclusion nor confirmation

This approach allows clinicians to rapidly assess the test's confirmatory and exclusionary potential, complementing numerical estimates such as the AUC or DOR.

Model diagnostic tools further support the evaluation of heterogeneity and study influence. For example, Cook's distance quantifies the extent to which each study disproportionately affects the overall model estimates [27]. Elevated Cook's distance values identify influential studies that, if excluded, would substantially alter pooled sensitivity, specificity, or thresholds. Complementary to this, standardized residuals assess the degree of misfit between observed and model-predicted values for sensitivity and specificity, highlighting outliers or studies with methodological inconsistencies.

In *midas* [28], these analyses are integrated into a simplified workflow particularly suited for clinicians or researchers without advanced statistical expertise. Using a single command, *midas* generates four key diagnostic plots: (1) a goodness-of-fit plot based on residuals; (2) a probability plot assessing bivariate normality assumptions; (3) a Cook's distance plot for detecting influential studies; and (4) a standardized residual scatterplot to evaluate model fit at the study level. While practical, it is important to acknowledge that *midas* does not implement a formally parameterized HSROC model.

 doi:10.20944/preprints202506.2461.v1

Consequently, graphical outputs and numerical estimates from *midas* may differ from those derived with more rigorous tools such as *metandi* [27] or *metadta* [29,30].

Additionally, the prediction regions displayed by *midas* [28] in HSROC plots are often exaggerated, primarily due to reliance on conventional I^2 estimates that ignore the correlation between sensitivity and specificity [31]. In contrast, *metadta* [29,30] incorporates the bivariate I^2 proposed by Zhou et al. [32], offering a more accurate, methodologically robust representation of heterogeneity and prediction intervals.

Lastly, bivariate boxplots, available in *midas* [28], provide an additional visual tool to explore heterogeneity and detect outlier studies based on the joint distribution of sensitivity and specificity. Unlike univariate influence metrics or goodness-of-fit plots, the bivariate boxplot simultaneously considers the correlation structure inherent to diagnostic accuracy data by plotting logit-transformed sensitivity against specificity. Concentric regions represent the expected distribution under the assumed bivariate normal model, with studies falling outside the outer envelope flagged as potential outliers. These studies often exhibit atypical combinations of sensitivity and specificity, which are frequently attributed to threshold effects, differences in study populations, or methodological variability. In our analysis, the bivariate boxplot identified four studies as outliers, including those with unusually high sensitivity but poor specificity, as well as others with distorted diagnostic trade-offs associated with elevated thresholds or pediatric cohorts. When interpreted alongside Cook's distance and standardized residuals, this plot provides a robust yet intuitive approach to visually assessing data integrity and identifying sources of heterogeneity beyond numerical summaries.

Together, these complementary tools—when applied appropriately—enhance the transparency, diagnostic validity, and clinical utility of DTA meta-analyses, promoting reliable evidence synthesis that genuinely informs decision-making.

Table 1 provides a summary of the main differences between classical meta-analytical models and hierarchical models in their application to DTA reviews.

Application of Hierarchical Models in Surgical Diagnostic Research: A Practical Example

To illustrate the practical implications of model selection in DTA meta-analyses, I generated a simulated dataset of 30 studies evaluating a hypothetical biomarker for diagnosing a surgical condition (Supplementary File 1). The dataset was designed to reflect common features of real-world surgical research, including threshold variability, between-study heterogeneity, and a marked threshold effect.

The first essential step in conducting a DTA meta-analysis is assembling a properly structured dataset, typically in spreadsheet format (e.g., Excel). The dataset must include a study identifier (commonly labeled as *studyid*; *study_id*; or *id*), followed by 2×2 contingency data: true positives (*tp*), false positives (*fp*), false negatives (*fn*), and true negatives (*tn*), ideally in that order. Additional columns can incorporate study-level covariates (dichotomous or continuous) for subsequent meta-regression analyses if required, such as the risk of bias, the study design, the threshold or the population. The accompanying dataset exemplifies the recommended structure before initiating a DTA meta-analysis. It is important to note that this framework assumes a single threshold per study. When individual studies report multiple thresholds, one must be selected, or specialized methods—such as bivariate ROC curve meta-analysis—should be applied, which are beyond the scope of this guide.

Initially, I applied a conventional, non-bivariate meta-analysis using *Meta-DiSc* (version 1.4). Separate random-effects models (DerSimonian-Laird) were used to pool sensitivity and specificity independently, and a symmetric SROC curve was generated using the Moses-Littenberg method. As expected, this approach ignored threshold variability, between-study heterogeneity, and the intrinsic correlation between sensitivity and specificity. The results appeared falsely precise, with narrow confidence intervals: pooled sensitivity, 0.74 (95% CI: 0.73–0.76); specificity, 0.68 (95% CI: 0.66–0.69);

 doi:10.20944/preprints202506.2461.v1

and heterogeneity exceeding 94% (I^2) for both. The area under the SROC curve (AUC) was 0.7881; however, no confidence intervals were provided, which limits interpretability (Figure 1).

Figure 1. Non-bivariate meta-analytic model applied to the simulated dataset using Meta-Disc. Sensitivity and specificity were pooled separately using DerSimonian-Laird random-effects models (upper left and upper right panels). The pooled diagnostic odds ratio (DOR) is displayed in the lower left panel. A symmetrical summary receiver operating characteristic (SROC) curve was generated using the Moses-Littenberg method (lower right panel), which does not provide confidence regions and exhibits the typical symmetric morphology associated with this classical approach. The Q* value corresponds to the Q-index, which reflects the theoretical point on the SROC curve where sensitivity and specificity are equal, providing a single summary measure of diagnostic performance. However, as it is derived from the simplified Moses-Littenberg model—which fails to account for heterogeneity and the correlation between sensitivity and specificity—the Q-index often overestimates diagnostic precision and lacks the methodological robustness of modern hierarchical models.

In contrast, re-analysis using hierarchical models in Stata (*metandi*, *midas*, *metadta*) implemented the BRMA framework, accounting for threshold effects, heterogeneity, and the correlation between sensitivity and specificity. All three approaches generated both pooled summary points and HSROC curves. The estimated AUC was 0.79 (95% CI: 0.75–0.82), with summary sensitivity 0.74 (95% CI: 0.66–0.81) and specificity 0.71 (95% CI: 0.63–0.79). LR+ was 2.6 (95% CI: 2.1–3.2), and LR- was 0.36 (95% CI: 0.3–0.44). The DOR was 7 (95% CI: 6-9). Wider confidence intervals accurately reflected underlying uncertainty. Between-study variance (tau^2) remained substantial: 0.99 for sensitivity, 1.14 for specificity. A strong negative correlation ($\rho = -0.87$) confirmed a pronounced threshold effect—completely overlooked by the non-hierarchical model (Figure 2).

Figure 2. Hierarchical meta-analysis of the same simulated dataset was performed using STATA. The upper panel displays a forest plot of sensitivity and specificity, jointly modeled within the bivariate framework (generated with *metadta*). The lower panel shows summary receiver operating characteristic (SROC) curves generated using three different STATA commands: *metandi* (left), *metadta* (center), and *midas* (right). Despite minor differences in output format, the three approaches produce highly similar global estimates, underscoring their methodological equivalence in terms of summary sensitivity and specificity. Each dot represents an individual study's estimate of sensitivity and specificity; in the *metandi* plot, circle size is proportional to the study's sample size. Both *metandi* and *metadta* generate a true HSROC curve explicitly applying the HSROC parameterization (Rutter and Gatsonis) and a summary point based on the bivariate random-effects framework. In contrast, *midas* generates the ROC curve as a continuous extrapolation derived from the bivariate model output, rather than through formal HSROC modeling, which results in a complete inferential ROC curve extending across the entire 0–1 specificity range. It is noteworthy that the prediction region generated by *midas* appears substantially wider than those produced by *metandi* and *metadta*, consistent with the methodological reasons discussed in the main text (i.e., the use of the classical Higgins' I^2 statistic, which tends to inflate heterogeneity estimates). Lastly, the visual distribution of individual studies follows an ascending pattern along the ROC space, consistent with the expected inverse relationship between sensitivity and specificity due to threshold variability. This alignment suggests that differences in positivity thresholds across studies may partially explain the observed dispersion.

Visual inspection of the HSROC plots revealed that the prediction regions produced by *metandi* and *metadta* were moderate in size, consistent with residual heterogeneity. In contrast, *midas* generated a markedly exaggerated prediction region. This discrepancy arises from *midas* relying on conventional Higgins' I^2 applied separately to sensitivity and specificity, ignoring their correlation, thus overstating uncertainty. Conversely, *metadta* implements the bivariate I^2 of Zhou et al., providing a more accurate, correlation-adjusted estimate of heterogeneity and prediction intervals.

Despite similar global AUCs (0.79 vs. 0.7881), the conventional model created a false sense of precision, lacking confidence intervals and failing to model key elements such as heterogeneity and threshold effects. Only hierarchical bivariate models adequately captured the complexity of diagnostic performance, yielding statistically valid, clinically interpretable estimates.

 doi:10.20944/preprints202506.2461.v1

To explore heterogeneity sources, I conducted univariable meta-regressions in *midas*, assessing covariates including risk of bias (QUADAS-2 index test domain), sample size, study design, and population type (pediatric vs. adult). Studies with high risk of bias—typically due to retrospective (*post hoc*) threshold selection—demonstrated inflated diagnostic estimates (sensitivity 0.79; specificity 0.80) relative to low-risk studies (sensitivity 0.72; specificity 0.67), consistent with overfitting. Smaller studies tended toward higher sensitivity, although the results were non-significant. No substantial performance differences emerged by study design or population (Figure 3).

Figure 3. Univariable meta-regression model applied to the simulated dataset in STATA. The covariates included: high risk of bias in the index test domain of the QUADAS-2 tool (Yes vs. No), large sample size (defined as >150 participants, Yes vs. No), prospective versus retrospective study design, and pediatric versus non-pediatric study population. For each potential effect modifier, stratified estimates of sensitivity and specificity are presented, allowing evaluation of their impact on diagnostic performance.

Further heterogeneity exploration included influence diagnostics using Cook's distance and standardized residuals derived from *metandi*. This dual approach quantified the influence of individual studies and their deviation from model expectations. Most studies aligned well with predicted values; however, Study 26 exhibited high influence (Cook's distance > 0.8) and a marked standardized residual for sensitivity near –1, indicating disproportionately low sensitivity compared to model expectations, significantly affecting global estimates (Figure 4). To assess the impact of this outlier, I repeated the hierarchical meta-analysis excluding study 26. The results remained broadly consistent, with only minor variations. The summary AUC was unchanged at 0.79 (95% CI: 0.75–0.82). Sensitivity increased slightly to 0.75 (95% CI: 0.67–0.81), which is consistent with the low sensitivity reported in the excluded study. Specificity decreased marginally to 0.70 (95% CI: 0.61–0.78). The positive likelihood ratio decreased slightly to 2.5 (95% CI: 2.0–3.1), and the negative likelihood ratio remained stable at 0.36 (95% CI: 0.30–0.44). The diagnostic odds ratio decreased marginally to 7 (95% CI: 5–9). Notably, the strong negative correlation between sensitivity and specificity ($\rho = -0.87$) persisted, consistent with a pronounced threshold effect accounting for 75% of the heterogeneity. Between-study variance decreased moderately following exclusion: τ^2 for sensitivity was reduced from 0.99 to 0.72, and τ^2 for specificity from 1.14 to 0.73. These findings confirm that, although study 26 exerted disproportionate influence on model estimates, its removal does not materially alter the overall interpretation: hierarchical models consistently reveal modest diagnostic accuracy, substantial heterogeneity, and a dominant threshold effect.

Figure 4. Influence analysis combining Cook's distance (upper panel), standardized residuals (middle panel), and bivariate boxplot (lower panel). The top plot shows Cook's distance for each study, quantifying its influence on global model estimates. Study 26 exhibits a markedly elevated Cook's distance, indicating disproportionate impact. The middle plot displays standardized residuals for sensitivity (y-axis) and specificity (x-axis); Study 26 demonstrates notably lower-than-expected sensitivity, while its specificity aligns with model predictions. The lower panel presents a bivariate boxplot of logit-transformed sensitivity and specificity, identifying outliers based on their joint distribution. Studies 1, 11, 12, and 21 fall outside the outer envelope, indicating atypical combinations of diagnostic performance, likely influenced by varying thresholds, study design, or specific populations such as pediatrics. Specifically, Study 1 shows unusually high sensitivity coupled with markedly low specificity, while Studies 11 and 12, both conducted in pediatric populations with elevated diagnostic thresholds (5.37 and 7.83 respectively), demonstrate imbalanced diagnostic performance likely linked to threshold effects and age-related variability. Study 21, despite being prospective, also presents an atypical sensitivity–specificity trade-off driven by a high false-positive rate and moderate cut-off value. These findings underscore the influence of population characteristics and threshold heterogeneity on diagnostic accuracy estimates.

To complement these analyses, a bivariate boxplot of logit-transformed sensitivity and specificity was generated, providing a graphical assessment of joint distribution patterns and outlier detection. This plot identified four studies (Studies 1, 11, 12, and 21) falling outside the expected

 doi:10.20944/preprints202506.2461.v1

envelope, consistent with atypical diagnostic trade-offs. Specifically, Study 1 demonstrated unusually high sensitivity with markedly low specificity, while Studies 11 and 12, both conducted in pediatric populations with elevated diagnostic thresholds, showed imbalanced diagnostic performance likely driven by threshold and age-related effects. Study 21, despite being prospective, also exhibited a distorted sensitivity–specificity relationship due to a high false-positive rate. Diagnostic performance estimates showed moderate variation compared to the full model. I repeated the model excluding these four studies. Summary AUC decreased slightly to 0.77 (95% CI: 0.73–0.81). Sensitivity declined to 0.69 (95% CI: 0.61–0.75), while specificity increased to 0.74 (95% CI: 0.66–0.82). The positive likelihood ratio improved to 2.7 (95% CI: 2.1–3.4), and the negative likelihood ratio increased to 0.42 (95% CI: 0.36–0.49), with the diagnostic odds ratio stable at 6 (95% CI: 5–8). The strong negative correlation between sensitivity and specificity ($\rho = -0.89$) persisted, consistent with a dominant threshold effect, which now explained 79% of the heterogeneity. Importantly, residual heterogeneity decreased notably: between-study variance (τ^2) dropped from 0.99 to 0.68 for sensitivity and from 1.14 to 0.73 for specificity. The ICC for sensitivity decreased to 0.16 (95% CI: 0.08–0.24), indicating a reduction in unexplained variance in sensitivity estimates. Despite the removal of influential studies, substantial heterogeneity and threshold effects remained, confirming that variability is intrinsic to the dataset and not solely driven by outliers

Together, these findings highlight specific contributors to between-study heterogeneity, reinforcing the need for cautious interpretation of global estimates.

Lastly, Fagan nomograms translated these findings into clinical terms for two scenarios: low pre-test probability (10%, e.g., general screening) and high pre-test probability (50%, e.g., symptomatic surgical patients). With LR+ ≈ 3 and LR− ≈ 0.36, the test modestly altered diagnostic probability: in low prevalence settings, a positive result increased post-test probability to 22%, while a negative result reduced it to 4%; in high-risk populations, post-test probabilities were 72% and 27%, respectively (Figure 5).

Figure 5. Fagan nomograms illustrating the impact of diagnostic test results on post-test probability in two distinct clinical scenarios. The left panel represents a low-prevalence setting (pre-test probability = 10%), where a positive test result increases the post-test probability to 22%, and a negative result reduces it to 4%. This demonstrates the test's limited ability to confirm or exclude disease in this context. The right panel illustrates a high-prevalence scenario (pre-test probability = 50%), where a positive test result increases the post-test probability to 72%, while a negative result decreases it to 27%. These findings emphasize that the clinical utility of the test is highly dependent on the baseline probability of disease and remains moderate, even under favorable conditions.

The likelihood ratio scattergram further summarized diagnostic performance, with pooled LR+ ≈ 3 and LR− ≈ 0.3, positioned in the lower right quadrant—indicating limited confirmatory and exclusionary value. The wide distribution of individual studies reinforced the presence of substantial heterogeneity (Figure 6).

Figure 6. Likelihood ratio scattergram summarizing the diagnostic performance of the index test in the simulated dataset. The pooled estimate, displayed as a central diamond with 95% confidence intervals, reveals moderate positive (LR+ ≈ between 2 and 3) and negative (LR− ≈ between 0.3 and 0.4) likelihood ratios, suggesting that the test offers only limited diagnostic utility for both ruling in and ruling out the disease. Consequently, it should not be considered suitable for definitive exclusion or confirmation. The wide dispersion of individual study estimates underscores the substantial heterogeneity in diagnostic performance and reinforces the need for cautious interpretation in clinical settings. Notably, due to the narrow distribution of negative likelihood ratio values (ranging from 0.1 to 1), the software did not generate the typical four visual quadrants within the scatter plot. This emphasizes that scattergram interpretation must always be contextualized alongside the quantitative results from the original hierarchical model to avoid misinterpretation.

 doi:10.20944/preprints202506.2461.v1

To assess publication bias, I applied Deeks' test via *midas*. The regression analysis revealed no significant small-study effect (P = 0.6), and the plot showed no visual asymmetry, indicating a low risk of publication bias (Figure 7).

Figure 7. Assessment of publication bias using Deeks' regression test implemented via the *pubbias* subcommand in *midas* (STATA).

The plot displays the inverse square root of the effective sample size on the x-axis and the log of the diagnostic odds ratio (DOR) on the y-axis for each included study. The regression line estimates the relationship between study size and diagnostic performance. A statistically significant slope indicates the presence of small-study effects, which are suggestive of publication bias. In this example, the slope is negative but non-significant (P = 0.6), suggesting no evidence of publication bias. The intercept reflects the expected diagnostic performance for a hypothetical infinitely large study. This graphical output provides both visual and inferential assessment of potential bias, facilitating transparent interpretation of meta-analytic results.

This example illustrates that outdated approaches, such as the Moses-Littenberg method, overlook crucial methodological factors, leading to misleading precision. In contrast, when applied properly, hierarchical models and complementary tools yield robust and clinically meaningful insights into diagnostic performance.

Common Pitfalls, Expertise Requirements, and Practical Recommendations

Despite the clear methodological advantages of hierarchical and bivariate models, their correct application in DTA meta-analyses requires both statistical expertise and clinical judgment. Numerous methodological errors continue to undermine the validity and clinical relevance of published diagnostic reviews.

A frequent pitfall is the inappropriate use of outdated methods, such as univariate pooling of sensitivity and specificity, which ignores their inverse relationship and fails to account for threshold effects or heterogeneity. Similarly, the use of platforms that do not support hierarchical modeling, such as *Meta-DiSc* or *RevMan*, often results in artificially narrow confidence intervals, underestimation of heterogeneity, and inflated performance estimates.

Even when hierarchical models are applied, an incorrect interpretation of outputs—such as a misunderstanding of heterogeneity metrics (e.g., tau^2, I^2), an overemphasis on global indicators like the AUC or DOR, or neglecting threshold variability—can produce misleading conclusions. In particular, reporting the AUC without clarifying its limitations or relying solely on the DOR as a surrogate of diagnostic performance oversimplifies the interpretation and obscures clinically relevant nuances.

To ensure robust, clinically meaningful evidence synthesis, several recommendations should be followed:

1. **Select appropriate, validated software** for hierarchical modeling, prioritizing tools such as *metandi*, *midas*, or *metadta* in Stata, or equivalent R packages (*mada*, *diagmeta*), depending on available expertise and project complexity.
2. **Report comprehensive diagnostic metrics**, including pooled sensitivity and specificity with confidence intervals, HSROC curves, and likelihood ratios. The DOR may be included as a secondary indicator, but should not be the sole measure of test performance. If reporting the AUC, clarify whether it derives from a rigorously parameterized HSROC model or a simplified approximation, as interpretations differ substantially.
3. **Use prediction intervals and heterogeneity estimates**, such as tau^2 or the bivariate I^2 by Zhou, to convey uncertainty and between-study variability transparently. Avoid overreliance on conventional I^2, which exaggerates heterogeneity by ignoring the correlation between sensitivity and specificity.

 doi:10.20944/preprints202506.2461.v1

4. **Incorporate complementary graphical and analytical tools**, including Fagan nomograms, likelihood ratio scatterplots, Cook's distance, standardized residuals, and meta-regression. These enhance interpretability, identify influential studies, and explore sources of heterogeneity.
5. **Interpret meta-regression findings cautiously**, particularly in meta-analyses with a limited number of studies. Introducing multiple covariates increases the risk of type I error and spurious associations due to overfitting. A commonly recommended rule of thumb is to include at least 10 studies per covariate to ensure model stability.
6. **Recognize and mitigate additional biases beyond publication bias**, such as spectrum bias (non-representative patient populations), selection bias, partial verification bias, misclassification, information bias, or disease progression bias. These sources of error frequently lead to overestimation of diagnostic performance, underscoring the need for rigorous methodological safeguards.
7. **Collaborate with experienced biostatisticians or methodologists** throughout all phases of the meta-analysis, from dataset preparation to statistical modeling and interpretation, to ensure adherence to best practices and maximize methodological rigor.
8. **Explicitly verify the statistical assumptions underpinning hierarchical models**. Many users apply BRMA or HSROC frameworks without assessing residual distribution, bivariate normality, or the influence of individual studies. Tools such as *midas modchk* or the influence diagnostics in *metandi* facilitate this evaluation and should be part of routine analysis. If bivariate normality is seriously violated (e.g., highly skewed distributions), consider data transformations, alternative modeling strategies such as more flexible Bayesian approaches, or, at the very least, interpret the results with great caution.
9. **Interpret results with caution in small meta-analyses**. When fewer than 10 studies are included, hierarchical models remain the preferred approach, but confidence intervals widen, prediction regions expand, and estimates of heterogeneity or threshold effects become less stable. In such scenarios, emphasis should shift towards transparency, identification of evidence gaps, and cautious interpretation rather than overconfident conclusions.

By applying these principles, researchers can avoid common pitfalls, reduce bias, and produce DTA meta-analyses that yield reliable, clinically applicable conclusions, ultimately improving diagnostic decision-making in both surgical and general medical practice.

Conclusions

While isolated diagnostic metrics—such as sensitivity, specificity, or the AUC—provide an initial overview of test performance, they offer an incomplete and potentially misleading representation of a test's true clinical value. Reporting these parameters without accounting for threshold effects, between-study heterogeneity, or pre-test probability disregards critical aspects of real-world diagnostic decision-making.

A valid evaluation of diagnostic accuracy requires a comprehensive, methodologically sound approach that integrates multiple parameters, explores sources of variability, and considers the specific clinical context in which the test is applied. In this regard, hierarchical models—specifically the HSROC and BRMA frameworks—represent the methodological gold standard for synthesizing DTA, particularly in heterogeneous fields such as surgical research.

These models overcome the limitations of traditional approaches by jointly modeling sensitivity and specificity, explicitly accounting for threshold variability, and incorporating both within- and between-study heterogeneity. Their use provides more realistic, clinically interpretable, and generalizable estimates of diagnostic performance.

However, hierarchical models alone are insufficient to guarantee valid conclusions. Complementary tools are essential to enhance both transparency and clinical applicability. Fagan nomograms translate diagnostic accuracy estimates into post-test probabilities, enabling clinicians to understand how test results modify disease likelihood in varying prevalence scenarios. Likelihood

 doi:10.20944/preprints202506.2461.v1

ratio scatterplots graphically summarize global test performance, clarifying whether the test reliably confirms (rule-in) or excludes (rule-out) disease [38].

Additionally, influence diagnostics—such as Cook's distance and standardized residual plots—help detect studies that disproportionately affect model parameters or exhibit poor fit, thereby enhancing model robustness. Prediction intervals derived from hierarchical models provide realistic estimates of between-study variability, while visual exploration of HSROC curves and scatterplots facilitates intuitive interpretation.

Despite these advances, methodological limitations persist, particularly in meta-analyses with few studies, limited covariate variability, or poorly reported primary data. In such settings, cautious interpretation, methodological transparency, and acknowledgment of uncertainty are essential.

In summary, appropriate application of hierarchical models, complemented by robust graphical tools and sensitivity analyses, enhances the validity, precision, and clinical relevance of DTA meta-analyses. Their correct use is essential for generating reliable, actionable evidence that genuinely informs diagnostic decision-making, ultimately improving patient care in both surgical and general medical contexts.

Supplementary Materials: The following supporting information can be downloaded at the website of this paper posted on Preprints.org.

CRediT authorship contribution statement: JAM: Conceptualization and study design; literature search and selection; investigation; methodology; project administration; resources; validation; visualization; writing – original draft; writing – review and editing.

CONFLICTS OF INTEREST: The author declares that he has no conflict of interest.

FINANCIAL STATEMENT/FUNDING: This review did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors, and the author has no external funding to declare.

ETHICAL APPROVAL: This study did not involve the participation of human or animal subjects, and therefore, IRB approval was not sought.

STATEMENT OF AVAILABILITY OF THE DATA USED: The dataset used for the simulated meta-analysis is provided as a supplementary file accompanying this manuscript.

DECLARATION OF GENERATIVE AI AND AI-ASSISTED TECHNOLOGIES IN THE WRITING PROCESS: During the preparation of this work, the author used ChatGPT (OpenAI) exclusively for language polishing. All scientific content, methodological interpretation, data synthesis, and critical analysis were entirely developed by the author.

References

1. Dinnes J, Deeks J, Kirby J, Roderick P. A methodological review of how heterogeneity has been examined in systematic reviews of diagnostic test accuracy. Health Technol Assess. 2005 Mar;9(12):1-113, iii. doi: 10.3310/hta9120. PMID: 15774235.

 doi:10.20944/preprints202506.2461.v1

2. Leeflang MM, Deeks JJ, Gatsonis C, Bossuyt PM; Cochrane Diagnostic Test Accuracy Working Group. Systematic reviews of diagnostic test accuracy. Ann Intern Med. 2008 Dec 16;149(12):889-97. doi: 10.7326/0003-4819-149-12-200812160-00008. PMID: 19075208; PMCID: PMC2956514.
3. Rutter CM, Gatsonis CA. A hierarchical regression approach to meta-analysis of diagnostic test accuracy evaluations. Stat Med. 2001 Oct 15;20(19):2865-84. doi: 10.1002/sim.942. PMID: 11568945.
4. Reitsma JB, Glas AS, Rutjes AW, Scholten RJ, Bossuyt PM, Zwinderman AH. Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. J Clin Epidemiol. 2005 Oct;58(10):982-90. doi: 10.1016/j.jclinepi.2005.02.022. PMID: 16168343.
5. Deeks JJ, Bossuyt PM, Leeflang MM, Takwoingi Y (editors). Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy. Version 2.0 (updated July 2023). Cochrane, 2023. Available from https://training.cochrane.org/handbook-diagnostic-test-accuracy/current.
6. Altman DG, Bland JM. Diagnostic tests. 1: Sensitivity and specificity. BMJ. 1994 Jun 11;308(6943):1552. doi: 10.1136/bmj.308.6943.1552. PMID: 8019315; PMCID: PMC2540489.
7. Akobeng AK. Understanding diagnostic tests 1: sensitivity, specificity and predictive values. Acta Paediatr. 2007 Mar;96(3):338-41. doi: 10.1111/j.1651-2227.2006.00180.x. PMID: 17407452.

 doi:10.20944/preprints202506.2461.v1

8. Singh PP, Zeng IS, Srinivasa S, Lemanu DP, Connolly AB, Hill AG. Systematic review and meta-analysis of use of serum C-reactive protein levels to predict anastomotic leak after colorectal surgery. Br J Surg. 2014 Mar;101(4):339-46. doi: 10.1002/bjs.9354. Epub 2013 Dec 5. PMID: 24311257.
9. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology. 1982 Apr;143(1):29-36. doi: 10.1148/radiology.143.1.7063747. PMID: 7063747.
10. Mandrekar JN. Receiver operating characteristic curve in diagnostic test assessment. J Thorac Oncol. 2010 Sep;5(9):1315-6. doi: 10.1097/JTO.0b013e3181ec173d. PMID: 20736804.
11. Arredondo Montero J, Martín-Calvo N. Diagnostic performance studies: interpretation of ROC analysis and cut-offs. Cir Esp (Engl Ed). 2023 Dec;101(12):865-867. doi: 10.1016/j.cireng.2022.11.011. Epub 2022 Nov 24. PMID: 36436801.
12. Altman DG, Bland JM. Diagnostic tests 2: Predictive values. BMJ. 1994 Jul 9;309(6947):102. doi: 10.1136/bmj.309.6947.102. PMID: 8038641; PMCID: PMC2540558.
13. Deeks JJ, Altman DG. Diagnostic tests 4: likelihood ratios. BMJ. 2004 Jul 17;329(7458):168-9. doi: 10.1136/bmj.329.7458.168. PMID: 15258077; PMCID: PMC478236.
14. Akobeng AK. Understanding diagnostic tests 2: likelihood ratios, pre- and post-test probabilities and their use in clinical practice. Acta Paediatr. 2007 Apr;96(4):487-91. doi: 10.1111/j.1651-2227.2006.00179.x. Epub 2007 Feb 14. PMID: 17306009.

 doi:10.20944/preprints202506.2461.v1

15. Bai S, Hu S, Zhang Y, Guo S, Zhu R, Zeng J. The Value of the Alvarado Score for the Diagnosis of Acute Appendicitis in Children: A Systematic Review and Meta-Analysis. J Pediatr Surg. 2023 Oct;58(10):1886-1892. doi: 10.1016/j.jpedsurg.2023.02.060. Epub 2023 Mar 6. PMID: 36966018.

16. Glas AS, Lijmer JG, Prins MH, Bonsel GJ, Bossuyt PM. The diagnostic odds ratio: a single indicator of test performance. J Clin Epidemiol. 2003 Nov;56(11):1129-35. doi: 10.1016/s0895-4356(03)00177-x. PMID: 14615004.

17. Pepe MS, Janes H, Longton G, Leisenring W, Newcomb P. Limitations of the odds ratio in gauging the performance of a diagnostic, prognostic, or screening marker. Am J Epidemiol. 2004 May 1;159(9):882-90. doi: 10.1093/aje/kwh101. PMID: 15105181.

18. Moses LE, Shapiro D, Littenberg B. Combining independent studies of a diagnostic test into a summary ROC curve: data-analytic approaches and some additional considerations. Stat Med. 1993 Jul 30;12(14):1293-316. doi: 10.1002/sim.4780121403. PMID: 8210827.

19. Dinnes J, Mallett S, Hopewell S, Roderick PJ, Deeks JJ. The Moses-Littenberg meta-analytical method generates systematic differences in test accuracy compared to hierarchical meta-analytical models. J Clin Epidemiol. 2016 Dec;80:77-87. doi: 10.1016/j.jclinepi.2016.07.011. Epub 2016 Jul 30. PMID: 27485293; PMCID: PMC5176007.

20. Harbord RM, Whiting P, Sterne JA, Egger M, Deeks JJ, Shang A, Bachmann LM. An empirical comparison of methods for meta-analysis of diagnostic accuracy showed

 doi:10.20944/preprints202506.2461.v1

hierarchical models are necessary. J Clin Epidemiol. 2008 Nov;61(11):1095-103. doi: 10.1016/j.jclinepi.2007.09.013. PMID: 19208372.

21. Wang J, Leeflang M. Recommendedsoftware/packages for meta-analysis of diagnostic accuracy. J Lab Precis Med 2019;4:22.
22. Zamora J, Abraira V, Muriel A, Khan K, Coomarasamy A. Meta-DiSc: a software for meta-analysis of test accuracy data. BMC Med Res Methodol. 2006 Jul 12;6:31. doi: 10.1186/1471-2288-6-31. PMID: 16836745; PMCID: PMC1552081.
23. Review Manager (RevMan) [Computer program]. Version 5.4. The Cochrane Collaboration, 2020.
24. R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing; 2024
25. SAS Institute Inc. SAS Software, Version 9.4. Cary, NC: SAS Institute Inc.; 2024.
26. StataCorp. Stata Statistical Software: Release 19. College Station, TX: StataCorp LLC; 2025.
27. Harbord, R. M., & Whiting, P. (2009). Metandi: Meta-analysis of Diagnostic Accuracy Using Hierarchical Logistic Regression. The Stata Journal, 9(2), 211-229. https://doi.org/10.1177/1536867X0900900203.
28. Ben Dwamena, 2007. "MIDAS: Stata module for meta-analytical integration of diagnostic test accuracy studies," Statistical Software Components S456880, Boston College Department of Economics, revised 05 Feb 2009
29. Nyaga VN, Arbyn M. Metadta: a Stata command for meta-analysis and meta-regression of diagnostic test accuracy data - a tutorial. Arch Public Health. 2022 Mar 29;80(1):95.

 doi:10.20944/preprints202506.2461.v1

doi: 10.1186/s13690-021-00747-5. Erratum in: Arch Public Health. 2022 Sep 27;80(1):216. doi: 10.1186/s13690-022-00953-9. PMID: 35351195; PMCID: PMC8962039.

30. Nyaga VN, Arbyn M. Comparison and validation of metadta for meta-analysis of diagnostic test accuracy studies. Res Synth Methods. 2023 May;14(3):544-562. doi: 10.1002/jrsm.1634. Epub 2023 Apr 18. PMID: 36999350.
31. Higgins JP, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. BMJ. 2003 Sep 6;327(7414):557-60. doi: 10.1136/bmj.327.7414.557. PMID: 12958120; PMCID: PMC192859.
32. Zhou Y, Dendukuri N. Statistics for quantifying heterogeneity in univariate and bivariate meta-analyses of binary data: the case of meta-analyses of diagnostic accuracy. Stat Med. 2014 Jul 20;33(16):2701-17. doi: 10.1002/sim.6115. Epub 2014 Feb 19. PMID: 24903142.
33. Begg CB, Mazumdar M. Operating characteristics of a rank correlation test for publication bias. Biometrics. 1994 Dec;50(4):1088-101. PMID: 7786990.
34. Egger M, Davey Smith G, Schneider M, Minder C. Bias in meta-analysis detected by a simple, graphical test. BMJ. 1997 Sep 13;315(7109):629-34. doi: 10.1136/bmj.315.7109.629. PMID: 9310563; PMCID: PMC2127453.
35. Deeks JJ, Macaskill P, Irwig L. The performance of tests of publication bias and other sample size effects in systematic reviews of diagnostic test accuracy was assessed. J Clin Epidemiol. 2005 Sep;58(9):882-93. doi: 10.1016/j.jclinepi.2005.01.016. PMID: 16085191.

 doi:10.20944/preprints202506.2461.v1

36. van Enst WA, Ochodo E, Scholten RJ, Hooft L, Leeflang MM. Investigation of publication bias in meta-analyses of diagnostic test accuracy: a meta-epidemiological study. BMC Med Res Methodol. 2014 May 23;14:70. doi: 10.1186/1471-2288-14-70. PMID: 24884381; PMCID: PMC4035673.

37. Fagan TJ. Letter: Nomogram for Bayes's theorem. N Engl J Med. 1975 Jul 31;293(5):257. doi: 10.1056/NEJM197507312930513. PMID: 1143310.

38. Rubinstein ML, Kraft CS, Parrott JS. Determining qualitative effect size ratings using a likelihood ratio scatter matrix in diagnostic test accuracy systematic reviews. Diagnosis (Berl). 2018 Nov 27;5(4):205-214. doi: 10.1515/dx-2018-0061. PMID: 30243015.