Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Validation of an Artificial Intelligence–Assisted Meta-Analysis System Using 400 Published Hazard Ratio Forest Plots: Retrospective Benchmark Validation Study

Kyriakos Polymenakos
Synthesa, Inc
Nikolaos Korasidis
Synthesa, Inc
Lefteris Teperikidis
lefteris@synthesa.ai

Synthesa, Inc

Research Article

Keywords: artificial intelligence, large language models, meta-analysis, evidence synthesis, reproducibility, statistical computing, forest plots, hazard ratios

Posted Date: November 3rd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11230257/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: Competing interest reported. Yes. K.P. and L.T. are co-founders of Synthesa, Inc., which develops and commercializes the Synthesa Meta-Analysis tool evaluated in this study. N.K. provides services to Synthesa, Inc. as a part-time independent contractor. These relationships constitute competing financial interests related to the technology evaluated in this manuscript. The authors declare no other competing financial or non-financial interests.

Abstract

Background

Large language models are increasingly being applied to evidence synthesis tasks, but their use for generating executable statistical code for meta-analysis remains poorly validated. Reproducing published forest plots from study-level data provides a practical and objective benchmark for this capability.

Objective

This study aimed to evaluate the numerical reproducibility and methodological appropriateness of an LLM-assisted meta-analysis system using published hazard-ratio forest plots as benchmarks.

Methods

We validated an LLM-based tool that accepts study-level data extracted from published forest plots and generates R code to conduct the meta-analysis. Forest plot images were processed using commercial OCR services, and the extracted study-level data were passed to the tool with all method-specific instructions removed, ensuring the tool's choices were independent of the published analysis. Reproduced outputs were compared against published pooled estimates across 400 benchmark analyses derived from published forest plots. For cases with numerical disagreement, iterative reproduction was used to recover the most likely original specification. All 400 benchmark analyses derived from published forest plots were subsequently evaluated by an external methodological adjudicator, who assessed whether the statistical choices made by the tool were appropriate relative to those identified in the published analyses.

Results

The reproduced pooled point estimate was identical to the published estimate in 315 of 400 cases (78.8%), within 0.01 in 93.8%, within 0.02 in 96.5%, and within 0.05 in 98.0%. The median absolute error was 0.00 and the mean was 0.005. The direction of effect was reproduced correctly in 399 of 400 cases. I^2 agreement was high, with 96.7% of cases within 5 percentage points. The main sources of discrepancy were the use of REML rather than DerSimonian-Laird estimation and the application of Hartung-Knapp confidence intervals, absent from many published analyses. The external methodological adjudicator confirmed that REML was an appropriate estimator in all 400 cases, and confidence interval differences were in all identifiable instances attributable to Hartung-Knapp adjustment or related methodological choices rather than errors in the generated code.

Conclusions

An LLM-based meta-analysis tool reproduced published pooled hazard-ratio estimates with high numerical fidelity across 400 benchmark analyses derived from published forest plots. Most discrepancies reflected differences between methods increasingly recommended for random-effects meta-analysis, namely REML estimation and Hartung-Knapp confidence intervals, and methods commonly used in published analyses. Expert methodological review judged the tool's methodological choices to be statistically appropriate in all reviewed cases. This validation framework, combining numerical comparison, iterative specification recovery, and methodological review, offers a reusable approach for assessing AI-assisted meta-analysis tools at scale.

Introduction

Systematic reviews and meta-analyses occupy a central role in evidence-based medicine, informing clinical guidelines, health technology assessments, and regulatory decisions. By pooling results across studies, meta-analyses increase statistical precision and provide quantitative summaries of treatment effects. Their reliability, however, depends not only on which studies are included but also on how the statistical analysis is implemented. Choices such as the heterogeneity estimator, the distinction between common-effect and random-effects models, the use of the Hartung–Knapp adjustment, and software-specific defaults can all materially influence the pooled estimate and its confidence interval.

Large language models (LLMs) are increasingly being applied across the evidence synthesis workflow, with most attention directed at tasks such as title and abstract screening, full-text review, data extraction, risk-of-bias assessment, and manuscript drafting. Less explored is the use of LLMs to generate executable statistical code for meta-analysis, which is a qualitatively different application. Unlike screening or summarization, code generation must produce syntactically valid, numerically correct output that correctly implements specific statistical methods and can be audited for methodological and implementation choices. Recent work has begun to demonstrate this potential: Reason et al. showed that GPT-4 could generate R scripts for network meta-analysis and produce interpretive reports [1], and their subsequent "AI Statistician" work extended this to automated model selection and output evaluation [2]. Prior work involving members of the present author team has demonstrated ChatGPT-assisted meta-analysis in individual clinical applications [3, 4], and several adjacent reports document LLM-assisted workflows under human supervision [5, 6]. The evidence base, however, remains limited to small case series and individual examples.

We evaluated the Synthesa Meta-Analysis tool, an LLM-assisted system, that accepts study-level data extracted from published forest plots and generates R code to conduct the meta-analysis. The validation design follows a three-stage logic. First, the tool's outputs are compared against published pooled estimates, and cases with negligible disagreement require no further investigation. Second, for cases that diverge, we perform iterative reproduction by systematically varying model parameters to identify which configuration the original authors most likely used. This step is not intended to show that the tool

can match historical analyses on demand; in practice, published results would not be available during actual use. Rather, it serves as a forensic step: once both sets of methodological choices are known, an expert adjudicator can assess whether the tool's decisions were statistically appropriate. That expert review constitutes the primary quality judgment, evaluating whether the tool's statistical approach is defensible relative to the published one, regardless of numerical agreement. This design separates numerical reproducibility from methodological appropriateness, which is the more meaningful criterion for an AI-assisted tool intended for prospective use.

Methods

Dataset and Ingestion

The validation dataset consisted of 400 published forest plots from meta-analyses reporting hazard ratios. Forest plots were collected as images, typically screenshots from PDF documents. Eligible forest plots were obtained from reports published after 2020 in journals with a Journal Impact Factor greater than 5, with priority given to journals with a Journal Impact Factor greater than 10 and were required to contain sufficient study-level information to reconstruct the meta-analysis computationally. The search focused primarily on clinical specialties in which time-to-event analyses and HRs are frequently reported, particularly oncology and gastroenterology/hepatology, while also including relevant studies from cardiology and cardiovascular medicine, infectious diseases and epidemiology, nephrology and transplantation, neurology, critical care, and respiratory medicine. These areas were prioritized because outcomes such as overall survival, progression-free survival, disease-free survival, recurrence-free survival, graft survival, hospitalization, disease progression, and mortality are commonly evaluated using time-to-event methods and consequently reported as HRs. Targeted high-impact journals included The BMJ, The New England Journal of Medicine (NEJM), JAMA, JAMA Internal Medicine, JAMA Oncology, The Lancet Oncology, The Lancet Gastroenterology & Hepatology, The Lancet Neurology, The Lancet Diabetes & Endocrinology, The Lancet Respiratory Medicine, The Lancet Infectious Diseases, Journal of Hepatology, Gut, Gastroenterology, Clinical Gastroenterology and Hepatology, Clinical and Molecular Hepatology, Journal of Hematology & Oncology, Molecular Cancer, and Cancer Communications. Potentially eligible meta-analyses were identified through targeted searches of these and other relevant journals, and their bibliographic records were saved and organized in a reference manager. Full-text articles were subsequently retrieved and manually reviewed to identify forest plots reporting pooled HRs. Eligible forest plots were saved directly from the online full-text article using the available “Save Image” option or, where necessary, extracted from the corresponding PDF. Each plot was assigned a standardized filename comprising the first author's surname, publication year, effect measure, and sequential plot number (e.g., “Author 2022 HR (1)”), with subsequent plots from the same article numbered consecutively. The corresponding full-text PDF was simultaneously saved using the same author-based naming convention, and all extracted forest plots and source PDFs were maintained within a single designated folder. This standardized collection procedure enabled each extracted forest plot to be linked to its original publication and subsequently used for direct comparison of the published meta-

analysis results with those reproduced using the Synthesa Meta-Analysis tool. Forest plots from the same source report were treated as independent entries when they corresponded to different comparisons, outcomes, or subgroups. Forest plots were excluded from the quantitative comparison when no numerically comparable pooled estimate was available, for example when only subgroup-specific estimates were reported without an overall summary. The 400 benchmark analyses derived from published forest plots were derived from 158 unique published source meta-analyses.

Sample Size and Power Considerations

The sample size was calculated for the primary validation endpoint, defined as the proportion of hazard-ratio forest plots for which the reproduced pooled point estimate differed from the published pooled estimate by no more than 0.05 on the ratio scale. The 0.05 threshold was specified a priori as a pragmatic numerical-reproducibility tolerance on the original hazard-ratio scale rather than as a clinical-equivalence margin. Because absolute differences on a ratio scale are scale-dependent, agreement was additionally summarized using relative-error measures, including the proportions of analyses with relative errors of ≤ 1%, ≤ 5%, and ≤ 10%. A one-sample binomial non-inferiority framework was used, with a null success proportion of 90%, an expected success proportion of 97.5%, one-sided alpha of 0.025, and 80% power. Under these assumptions, the minimum required sample size was estimated to be 93 forest plots. To account for exclusions, incomplete reporting, OCR failures, and plots without numerically comparable pooled estimates, the target minimum sample size was rounded upward to at least 100 hazard-ratio forest plots. The final validation dataset substantially exceeded this minimum, providing additional precision for secondary agreement analyses and expert methodological review.

Data extraction

Data extraction from each image was performed in two stages. First, the image was processed using optical character recognition (OCR) services from Amazon Web Services and Google Cloud to identify and digitize the text and numerical content of the forest plot. Second, the extracted content was interpreted by gpt-5, through the OpenAI API, to produce a structured representation of the data. All structured outputs were manually checked against the source publication by a human reviewer for accuracy (KP).

The structured data were organized into two distinct datasets, which were kept separate throughout the analysis. The first dataset, referred to here as the study-level input data, contained the per-study information needed to run the meta-analysis: study identifiers, hazard ratios and corresponding confidence intervals from which log-scale estimates and standard errors were derived. The second dataset, referred to as the extracted published results, contained the reported outputs of the published meta-analysis: the pooled point estimate, confidence intervals, I^2, and τ^2. These two datasets served different roles in the validation workflow. The study-level input data were provided to the LLM-based tool

as the starting point for its analysis. The extracted published results were set aside and only reintroduced at the comparison stage, after the tool had produced its own outputs.

Core LLM-assisted meta-analysis

The system evaluated in this study was the Synthesa Meta-Analysis tool, a proprietary LLM-based meta-analysis platform developed by Synthesa, Inc. It uses gpt-5 through the OpenAI API together with a structured set of instructions and a purpose-built knowledge base. The knowledge base contains hand-selected excerpts from relevant textbooks, methodological guidance documents, and peer-reviewed articles covering effect-measure selection, transformation of ratio measures, estimation of between-study heterogeneity, the distinction between common-effect and random-effects models, construction of confidence intervals, and interpretation of forest-plot outputs. It also includes a curated set of worked examples, each comprising a representative study-level dataset, a corresponding R script, and instructions describing the scenario for which that example is appropriate. These examples were designed to cover the predictable cases arising in meta-analysis of ratio measures, including datasets requiring log-scale transformations, reconstruction of standard errors from published confidence intervals, and alternative model specifications.

When a study-level dataset is submitted, the tool evaluates its structure, including the type of effect measure, the form in which uncertainty is reported (confidence intervals or standard errors), and the number of contributing studies. The tool combines retrieval-augmented generation (RAG) over the curated knowledge base with the task-specific instructions to generate an R script, which it then executes, producing a forest plot and the full set of standard meta-analytic outputs. All analyses were conducted using R version 4.5.1 and version 8.2.1 of the meta package [7]. For the hazard-ratio analyses evaluated in this validation dataset, the system selected restricted maximum likelihood (REML) estimation for between-study heterogeneity and the Hartung-Knapp method for confidence intervals in random-effects analyses. Both common-effect and random-effects pooled estimates are computed for every dataset, along with estimates of between-study heterogeneity (I^2 and τ^2). For each dataset, R code generation and execution proceeded through a fully automated retry loop. If execution failed, the R error output was returned to the LLM, which revised the script and re-executed it, for a maximum of 10 execution attempts comprising the initial execution and up to nine automated correction rounds. No human intervention was permitted during this code-generation and debugging process, and the published pooled results were not available to the tool at this stage.

An important design feature of the validation setup is that the tool's methodological choices were made independently of the published analysis. Any method-specific instructions that may have been present in the structured notes accompanying the extracted data, such as references to DerSimonian-Laird estimation or specific confidence interval methods, were systematically removed before being passed to the tool. This ensured that the tool's decisions regarding heterogeneity estimator, model type, and confidence interval construction reflected its own defaults and knowledge base, rather than information recovered from the published forest plot.

Result Comparison and Iterative Reproduction

After the tool completed its analysis, the extracted published results were reintroduced for comparison. For each forest plot, the tool had produced both common-effect and random-effects pooled estimates; the choice of which to compare against the published estimate followed a prespecified rule based on what the published forest plot reported.

When the published forest plot explicitly reported both a common-effect and a random-effects estimate, the comparison used the estimate corresponding to the model deemed more appropriate given the observed heterogeneity: the common-effect estimate when I^2 was below 20%, and the random-effects estimate otherwise. When only one model type was reported in the published forest plot, that model's estimate was used directly. When the published forest plot did not explicitly indicate which model had been used, the reproduced common-effect and random-effects estimates were compared against the published estimate separately, and the reproduced estimate with the smaller absolute difference was selected. This was not intended to minimize the reported errors; rather, it was a principled attempt to recover which model the original authors had used, in the absence of that information. The model selected in each case was recorded. Exact numerical agreement was defined as equality between the published and reproduced pooled point estimates, rounded to the precision reported in the published forest plot. When exact agreement was achieved, no further steps were taken for that forest plot. When the reproduced estimate did not match, the forest plot entered an iterative reproduction stage.

The purpose of iterative reproduction was to identify which combination of statistical parameters would reproduce the published estimate, thereby recovering the methodological choices of the original analysis. This was not a claim that the tool could replicate historical analyses on demand; in practice no published result would be available. Rather, it was a forensic step enabling a meaningful methodological comparison. The iterative procedure used the original study-level data and the tool's initial R script as a starting point. The R script was adapted to expose the key meta-analysis function call as a parameter interface, and the code was split into the portion preceding and following that call. A comprehensive grid of parameter combinations was then assembled, and each was executed in turn, with all other aspects of the analysis held constant. The full parameter grid was designed to cover pooling approaches and random-effects options implemented in the meta R package [7]. Parameters varied were: pooling method (inverse-variance, Mantel–Haenszel, Peto, and generalized linear mixed models, where applicable) [7, 8]; between-study heterogeneity estimator (DerSimonian–Laird [9], REML [10, 11], Paule–Mandel, maximum likelihood, empirical Bayes, Hunter–Schmidt, Sidik–Jonkman, and Hedges) [10–12]; confidence interval method for random-effects analyses (Hartung–Knapp [13–16], classic Wald-type, and Kenward–Roger) [7, 8, 15]; and the formula used to calculate I^2 [8]. For Mantel–Haenszel combined with DerSimonian–Laird estimation, additional Cochrane-specific variants were included [7, 8].

The number of applicable combinations varies by data type. For binary outcomes with odds ratios, where all pooling methods were available, the full grid reached 238 combinations. For binary outcomes with risk ratios or risk differences, where Peto's method and certain other approaches did not apply, this

reduced to 88. For continuous outcomes analyzed with raw mean differences, the applicable τ estimators were more limited, giving 22 combinations. For pre-computed ratio measures on the log scale, including hazard ratios, the analysis used inverse-variance weighting via the metagen function of the meta R library [7], which does not take a pooling method parameter; the applicable grid therefore covered only heterogeneity estimator, confidence interval method, and I^2 formula, yielding 38 combinations. In the present validation, all forest plots reported hazard ratios, so 38 combinations were evaluated per forest plot. Count-based pooling methods such as Mantel-Haenszel and Peto, which require individual event counts and are designed for binary outcomes, are not appropriate for hazard ratio data and were excluded [8]. These methods are retained in the general grid because they have been useful in reproducing older binary-outcome meta-analyses; as expected, they did not contribute to reproducing any of the analyses in this dataset. The combination producing the closest match to the published pooled estimate, confidence intervals, and heterogeneity statistics was recorded as the most likely specification used in the original published analysis.

External Methodological Review

All 400 benchmark analyses derived from published forest plots and their corresponding reproduced analyses were evaluated by an external methodological adjudicator who was not a member of the author team and had no role in development of the Synthesa Meta-Analysis tool. For each case, the adjudicator was provided with the published meta-analysis results and reported methodology, where available, together with the corresponding Synthesa-generated analysis, including the generated R code, statistical specification, pooled estimates, confidence intervals, and observed numerical differences. The adjudicator was asked to assess the methodological appropriateness of the statistical choices rather than numerical proximity to the published result.

The review was not blinded to whether an analysis originated from the published report or from Synthesa. Full blinding was considered impractical for this retrospective methodological adjudication because substantially more implementation information, including generated code and explicit statistical parameters, was available for the reproduced analyses, whereas corresponding implementation details were frequently unavailable or incompletely reported in the source publications. Masking source labels would therefore not have removed this underlying information asymmetry and could have restricted access to information required for methodological assessment. Moreover, the consistent methodological approach and presentation of the Synthesa-generated forest plots could have allowed their origin to be inferred even if source labels had been removed. The absence of blinding was nevertheless considered a potential source of assessment bias and is acknowledged as a limitation.

For each forest plot, the adjudicator recorded the preferable heterogeneity estimator given the available data, assessed confidence-interval differences and their likely methodological explanation, and identified any instance in which the statistical approach used by the tool was considered less appropriate than that used in the corresponding published analysis. The primary methodological question was whether the

tool's default use of REML estimation and Hartung-Knapp confidence intervals was justifiable for the dataset relative to the alternative specification identified from the published analysis.

Results

Meta-analysis tool results

The validation dataset comprised 400 benchmark analyses derived from published forest plots of hazard ratios. The published model type was not reported in 377 cases (94.25%); 18 forest plots explicitly reported a random-effects analysis and 5 reported a common-effect analysis. All 400 benchmark analyses (100%) successfully completed the automated code-generation and execution workflow and produced executable R code and a forest plot without human code correction. Attempt-level execution counts were not retained, and the proportion succeeding on the first execution attempt therefore could not be determined.

Point estimate agreement

The reproduced pooled point estimate was identical to the published estimate in 315 of 400 cases (78.8%). Agreement within an absolute difference of 0.01 was achieved in 375 cases (93.8%), within 0.02 in 386 cases (96.5%), and within 0.05 in 392 cases (98.0%). The median absolute error was 0.00 and the mean was 0.005 (RMSE 0.023). The median relative error was 0.0% and the mean was 0.41%, with 86.8% of cases showing a relative error of 1% or less, 98.8% within 5%, and 99.3% within 10%. The maximum absolute difference observed was 0.27, occurring in a single case. Pearson and Spearman correlations between published and reproduced point estimates were 0.999 and 1.000, respectively. The direction of effect was reproduced correctly in 399 of 400 cases. Figure 1 shows the cumulative distribution of absolute errors in pooled point estimates, and Fig. 2 shows published versus reproduced values on the log scale.

Heterogeneity

Agreement in I^2 was high. The median absolute difference in I^2 was 0.1 percentage points and the mean was 0.94 percentage points, with 96.7% of cases within 5 percentage points. Figure 3 shows cumulative results for absolute errors in I^2 and tau². Figure 4 shows the distribution of I^2 category transitions between published and reproduced analyses. Of 399 forest plots with available I^2 data (one published forest plot did not report I^2), 387 (97.0%) remained in the same heterogeneity category (Low: $I^2 < 25\%$, Moderate: 25–74%, High: ≥75%), and category changes were observed in 12 cases (3.0%), in all instances involving transitions between adjacent categories. Agreement in between-study variance (τ^2) was also high, with a median absolute difference of 0.0005 (0.00 when rounded to two decimal places) and a mean absolute difference of 0.02

Confidence intervals

Confidence interval agreement was assessed qualitatively and semi-quantitatively. For the published interval $[L_{pub}\ ,\ U_{pub}]$ and the reproduced interval $[L_{rep},\ U_{rep}]$, overlap was defined as the length of their intersection divided by the length of their union (Jaccard similarity for intervals on the real line) [17]:

$$overlap\ (\%) = \ 100\frac{\max\left(0, \min\left(U_{pub}, U_{rep}\right) - \max\left(L_{rep}, L_{pub}\right)\right)}{\max\left(U_{pub}, U_{rep}\right) - \min\left(L_{rep}, L_{pub}\right)}$$

Values of 0% and 100% corresponded to non-overlapping and identical intervals, respectively. The median CI overlap between published and reproduced confidence intervals was 97.7% and the mean was 85.0%. Cases with lower overlap were predominantly attributable to methodological differences in confidence interval construction rather than errors in the pooled point estimate itself, as discussed further in the Expert Review Results section. Nominal statistical significance differed between the published and reproduced analyses in 30 of 400 cases (7.5%), attributed mostly to wider confidence intervals estimated by the tool crossing over the null hypothesis threshold.

External Methodological Review Results

The external methodological adjudicator evaluated all 400 benchmark analyses derived from published forest plots. In every case, the adjudicator judged REML estimation to be the appropriate or preferable choice for between-study heterogeneity; DerSimonian-Laird estimation was not preferred in any instance. The reconstructed specification of the published analyses was consistent with DerSimonian-Laird estimation in 321 of 400 cases. Confidence interval discrepancies were noted in 207 of 400 benchmark analyses derived from published forest plots (51.75%). Of these, 176 (85.0% of discrepant cases) were attributable to the Hartung-Knapp adjustment applied in the reproduced analyses but absent from the published forest plots. In 22 additional cases, confidence interval inconsistencies were observed without a specific method change being identified. In 4 cases, changing from REML to DerSimonian-Laird estimation improved confidence interval agreement, and in 1 case the reverse was true. The remaining 4 cases involved a mismatch between the common-effect and random-effects model reported in the published analysis versus that produced by the tool.

In summary, the adjudicator found no case where the tool's methodological approach was considered inferior to that of the published analysis. Differences in confidence intervals were in all identifiable cases explained by differences in the Hartung-Knapp adjustment or related methodological choices, and the adjudicator consistently endorsed REML with Hartung-Knapp as the appropriate contemporary default.

Discussion

Quality of Automated Meta-analysis

The AI-assisted system reproduced published pooled hazard-ratio estimates with high numerical fidelity across 400 benchmark analyses derived from published forest plots. The majority of discrepancies were small, and the direction of effect was preserved in all but one case. Crucially, the external review found

no instance in which the tool's methodological approach was considered inferior to that of the published analysis, providing a basis for evaluating quality that goes beyond numerical agreement alone.

The main source of difference between the tool's output and the published forest plots was the choice of heterogeneity estimator. Published analyses predominantly used DerSimonian-Laird (DL) estimation, which has been the default in most meta-analysis software since its introduction in 1986 [9]. The tool consistently applied restricted maximum likelihood (REML) estimation, which is now widely recommended as a more statistically appropriate default for random-effects meta-analysis [11, 12]. Simulation studies have shown that REML provides less biased estimates of between-study variance in most practical settings, particularly when the number of studies is small to moderate [10, 12, 18]. The use of DL by published authors reflects established convention rather than principled method selection and should not be interpreted as a better approach. The external methodological adjudicator likewise judged REML to be appropriate or preferable across the included benchmark cases.

A second source of discrepancy was the Hartung-Knapp (HK) adjustment for confidence intervals in random-effects analyses. The tool applied HK adjustment in all such cases, whereas most published forest plots did not. The HK method uses a t-distribution reference for the pooled estimate rather than the standard normal, usually resulting in wider confidence intervals that better reflect uncertainty about the between-study variance, particularly in analyses with few studies [13, 14]. Omitting the adjustment in analyses with a small number of contributing studies can produce confidence intervals that are too narrow, giving an overstated impression of precision. The widening of confidence intervals observed in the tool's reproduced analyses is therefore a feature of appropriate practice rather than a discrepancy requiring correction.

A secondary consequence of HK-widened confidence intervals was a change in nominal statistical significance in 30 of 400 cases (7.5%). In all of these cases the pooled point estimate itself was unchanged or nearly identical; the significance change was an artifact of the wider intervals, not a difference in the underlying analysis. This is an expected and well-understood consequence of applying a more conservative inference procedure and should not be interpreted as a disagreement in effect size. One known limitation of the HK adjustment is that it can produce degenerate confidence intervals in cases where the estimated between-study variance is zero, effectively collapsing the interval to the point estimate [15]. This issue has been documented in the literature, and several modifications have been proposed to guard against it [15, 16]. In the present study, we used the standard (unadjusted) Hartung-Knapp implementation, and no such collapse was observed across the 400 analyses, suggesting that this pathological behavior did not arise in any of the datasets included.

A Practical Tool for Reproducing Meta-Analysis at Scale

Reproducibility of published meta-analyses matters for science, yet it remains non-trivial in practice. The statistical methods used are often incompletely reported, and even when they are described, the combination of software defaults, heterogeneity estimator, confidence interval method, and model

choice creates a large enough parameter space that exact reconstruction without prior knowledge of each decision is uncertain.

The tool evaluated here, combined with the validation framework developed alongside it, provides a scalable and low-cost approach to this problem. The pipeline, from OCR-based data extraction to LLM-generated R code to systematic comparison with published results, requires minimal manual effort per forest plot and can operate across large collections.

A key design principle of the framework is that iterative reproduction and expert review are complementary and both necessary. Numerical comparison alone cannot distinguish a methodological difference from a genuine error. Iterative reproduction resolves this by identifying, in most cases, which specification the original authors most likely used, thereby making both sets of methodological choices explicit. Without that step, an expert adjudicator would be presented with two sets of numerical results without knowing the methodology behind either, making it impossible to comment on which approach is more appropriate. With both specifications known, the adjudicator can go directly to the methodological question: are the tool's choices statistically defensible relative to the published ones?

This combination of capabilities could be of value beyond retrospective validation. During peer review, the same workflow could be used to verify that a submitted meta-analysis is reproducible and that its methodological choices are appropriate, with the potential to reduce the manual burden of statistical replication. As the volume of published meta-analyses continues to grow, scalable tools for assessing their statistical integrity become increasingly relevant for both editorial processes and evidence synthesis practice.

Limitations

This study has several limitations. First, the validation was restricted to forest plots reporting hazard ratios, and the findings may not generalize directly to binary, continuous, diagnostic accuracy, network, or individual participant data meta-analyses. Second, published meta-analyses are an imperfect reference standard because model type, heterogeneity estimator, confidence interval method, and software defaults are often incompletely reported. Third, methodological adjudication was performed by a single external, non-author adjudicator. The review was not blinded to whether the statistical specifications originated from the Synthesa-generated or published analyses, and no second adjudicator was used. The absence of blinding and, more importantly, formal inter-rater assessment therefore remains a potential source of assessment bias. Multiple eligible forest plots could originate from the same source meta-analysis and, particularly for related outcomes or subgroup analyses, could include overlapping sets of primary studies. The 400 benchmark analyses derived from 158 published meta-analyses therefore represent 400 benchmark analysis instances rather than 400 fully independent source populations. Future work should evaluate additional effect measures and designs, incorporate multiple blinded external methodological adjudicators, and assess fully end-to-end workflows from full-text ingestion to final meta-analytic output.

Conclusions

An LLM-based meta-analysis tool was evaluated against 400 published forest plots reporting hazard ratios. The AI-assisted system reproduced the published pooled point estimate exactly in 78.8% of cases and within 0.05 in 98.0%, with the direction of effect preserved in all but one case. The main sources of discrepancy were differences in the heterogeneity estimator and the use of Hartung-Knapp confidence intervals, both of which reflect deliberate methodological choices rather than errors. External methodological review confirmed that the tool's approach, applying REML estimation and Hartung-Knapp adjustment as defaults, was not inferior to the published analysis in any of the 400 cases reviewed.

The validation framework developed alongside the tool, combining systematic numerical comparison with iterative recovery of the original authors' specification and expert methodological review, provides a reusable approach for assessing AI-assisted meta-analysis at scale. Together, the tool and this framework offer a low-cost and scalable basis for reproducing or auditing published meta-analyses, with potential applications in evidence synthesis workflows and editorial quality control. Further validation across other effect measures and meta-analytic designs is warranted.

Abbreviations

AI
artificial intelligence
API
application programming interface
CI
confidence interval
DL
DerSimonian-Laird
HK
Hartung-Knapp
HR
hazard ratio
I^2
I-squared statistic
LLM
large language model
OCR
optical character recognition
PDF
Portable Document Format
RAG
retrieval-augmented generation

REML
restricted maximum likelihood
RMSE
root mean square error
τ^2
tau-squared; between-study variance

Declarations

Conflicts of Interest

K.P. and L.T. are co-founders of Synthesa, Inc., which develops and commercializes the Synthesa Meta-Analysis tool evaluated in this study. N.K. provides services to Synthesa, Inc. as a part-time independent contractor. These relationships constitute competing commercial interests related to the technology evaluated in this manuscript.

Clinical trial number

Not applicable

Ethics approval and consent to participate

Not applicable. No human participants were recruited, and no identifiable patient-level data were accessed.

Funding

This study was conducted using internal resources of Synthesa, Inc. and received no external funding.

Data Availability

We are providing a directory with all outcomes of the analysis. The 400 published forest plots are included, along with the 400 benchmark analyses derived from published forest plots generated by the tool. R Code for every meta-analysis performed by the tool is also provided. Please note that the production Synthesa implementation, full prompt orchestration, and retrieval configuration remain proprietary.

Acknowledgments

We want to thank Paschalis Karakasis, a cardiologist with extensive experience in conducting and interpreting meta-analyses, for performing the external review.

The authors used ChatGPT (GPT-5.6 Sol; OpenAI, 2026) for editorial review and language refinement of the manuscript. The authors reviewed all suggested changes and take full responsibility for the final

content.

Author Contribution

K.P. and L.T. conceived and designed the study. K.P. collected the published forest plots, performed data extraction and manual verification of the extracted study-level data, and contributed to the validation analyses. N.K. contributed to the development and implementation of the statistical validation workflow, including the iterative reproduction analyses and comparison of published and reproduced meta-analytic results. L.T. supervised the study, contributed to the methodological design, interpretation of the findings, and overall study coordination. K.P. drafted the manuscript. N.K. and L.T. critically reviewed and revised the manuscript for important intellectual content. All authors reviewed and approved the final manuscript.

References

1. Reason T, et al. Artificial Intelligence to Automate Network Meta-Analyses: Four Case Studies to Evaluate the Potential Application of Large Language Models. Pharmacoecon Open. 2024;8(2):205–20.
2. Reason T, et al. The Artificial Intelligence Statistician: Utilizing Generative Artificial Intelligence to Select an Appropriate Model and Execute Network Meta-Analyses. Value Health. 2025;28(11):1646–54.
3. Teperikidis L, et al. Colchicine for Major Adverse Cardiovascular Events: An Updated ChatGPT-Assisted Systematic Review and Meta-Analysis. J Cardiovasc Pharmacol. 2026;87(6):420–31.
4. Teperikidis L, et al. Validating ChatGPT's role in systematic reviews and meta-analyses: A case study on GLP-1 receptor agonists and all-cause mortality. Eur J Intern Med. 2025;134:145–7.
5. Jiang C, et al. ChatGPT-based meta-analysis for evaluating the temporal and spatial characteristics of deoxynivalenol contamination in Chinese wheat. J Hazard Mater. 2024;480:135888.
6. Xing X, et al. Leveraging AI for Meta-Analysis: Evaluating LLMs in Detecting Publication Bias for Next-Generation Evidence Synthesis. Cochrane Evid Synth Methods. 2025;3(5):e70047.
7. Schwarzer G. *meta: General Package for Meta-Analysis*. 2025, Comprehensive R Archive Network (CRAN).
8. Deeks JJ, Higgins JPT, Altman DG et al. *Analysing data and undertaking meta-analyses*, in *Cochrane Handbook for Systematic Reviews of Interventions*, J.P.T. Higgins, Editors. 2019, John Wiley & Sons: Chichester (UK). pp. 241–284.
9. DerSimonian R, Laird N. Meta-analysis in clinical trials. Control Clin Trials. 1986;7(3):177–88.
10. Viechtbauer W. Bias and Efficiency of Meta-Analytic Variance Estimators in the Random-Effects Model. J Educational Behav Stat. 2005;30(3):261–93.

11. Veroniki AA, et al. Methods to estimate the between-study variance and its uncertainty in meta-analysis. Res Synth Methods. 2016;7(1):55–79.
12. Langan D, et al. A comparison of heterogeneity variance estimators in simulated random-effects meta-analyses. Res Synth Methods. 2019;10(1):83–98.
13. Hartung J, Knapp G. A refined method for the meta-analysis of controlled clinical trials with binary outcome. Stat Med. 2001;20(24):3875–89.
14. Knapp G, Hartung J. Improved tests for a random effects meta-regression with a single covariate. Stat Med. 2003;22(17):2693–710.
15. Röver C, Knapp G, Friede T. Hartung-Knapp-Sidik-Jonkman approach and its modification for random-effects meta-analysis with few studies. BMC Med Res Methodol. 2015;15:99.
16. Jackson D, et al. The Hartung-Knapp modification for random-effects meta-analysis: A useful refinement but are there any residual concerns? Stat Med. 2017;36(25):3923–34.
17. Jaccard P. *Etude de la distribution florale dans une portion des Alpes et du Jura.* Bulletin de la Societe Vaudoise des Sciences Naturelles, 1901. 37: pp. 547–579.
18. Tanriver-Ayder E, et al. Comparison of commonly used methods in random effects meta-analysis: application to preclinical data in drug discovery research. BMJ Open Sci. 2021;5(1):e100074.

Figures

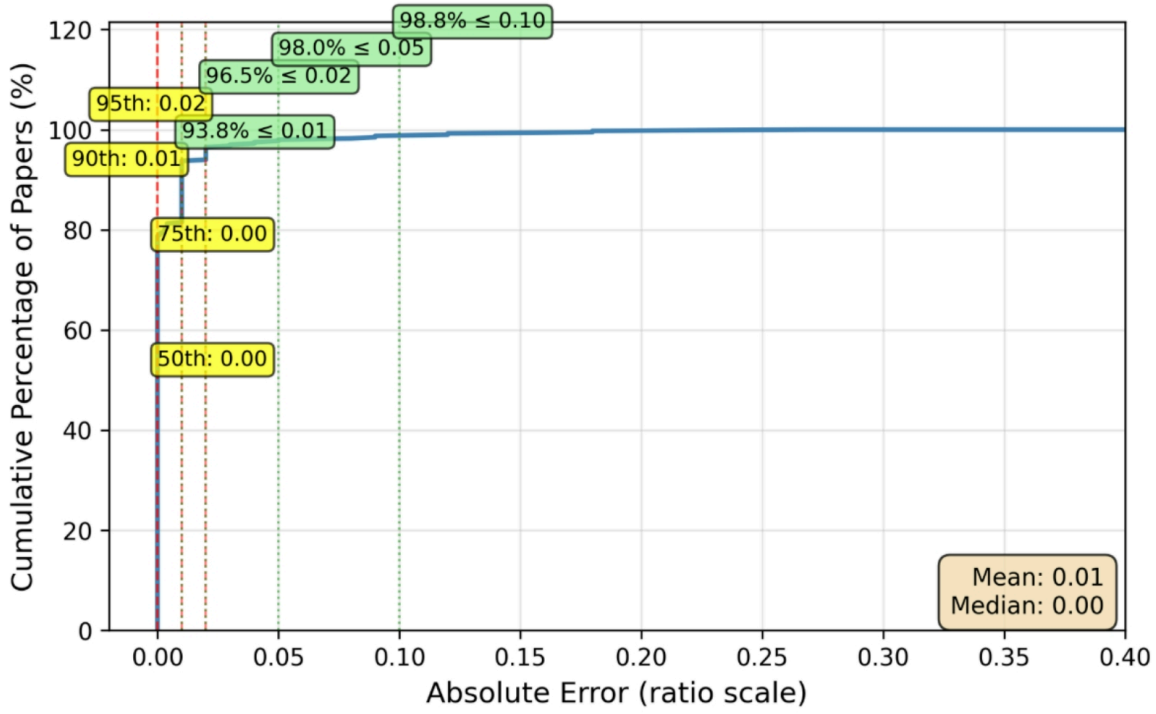


Figure 1

Distribution of absolute differences between published benchmark estimates and AI-assisted reproduced pooled hazard-ratio estimates.

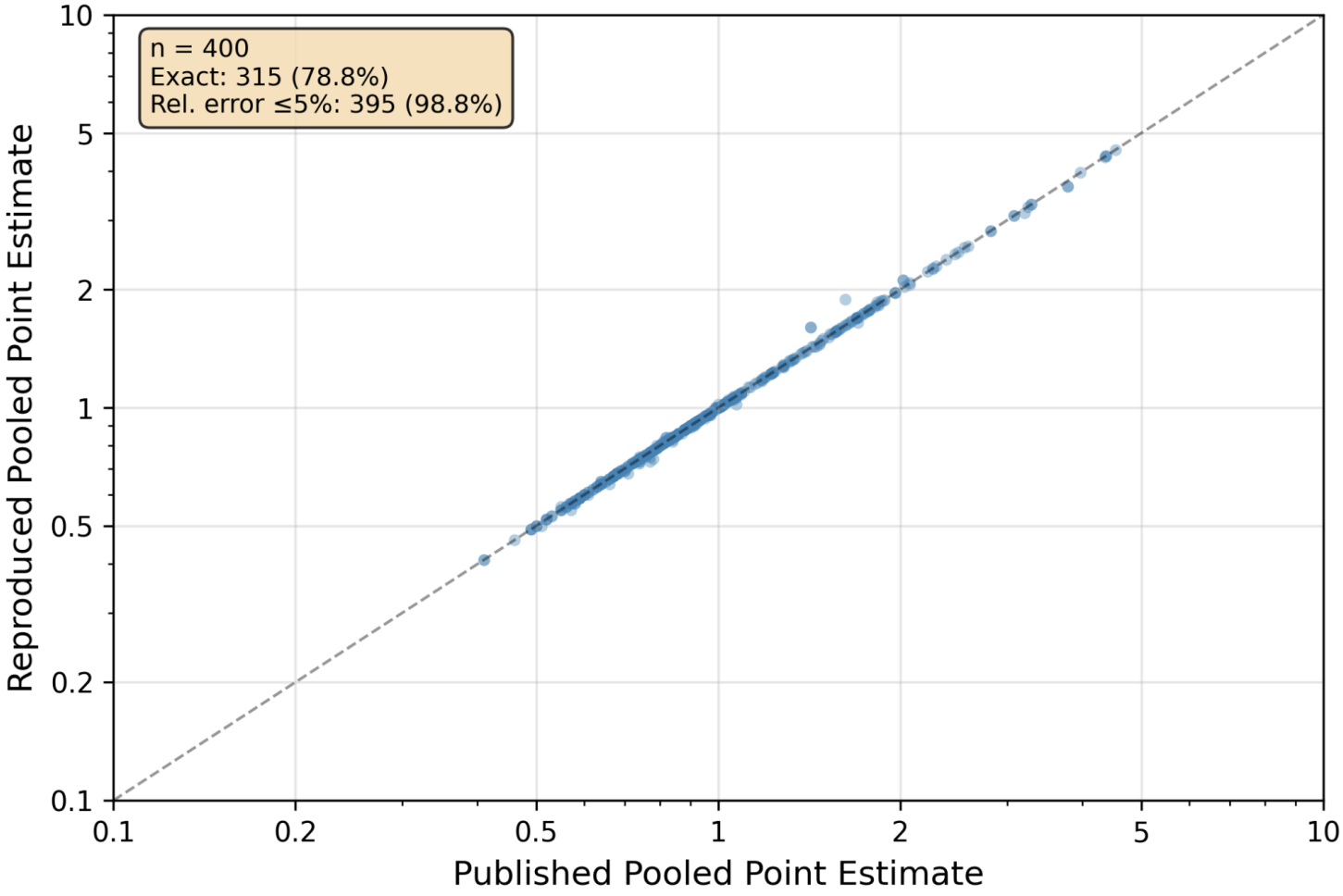


Figure 2

Agreement between published benchmark estimates and AI-assisted reproduced pooled estimates.

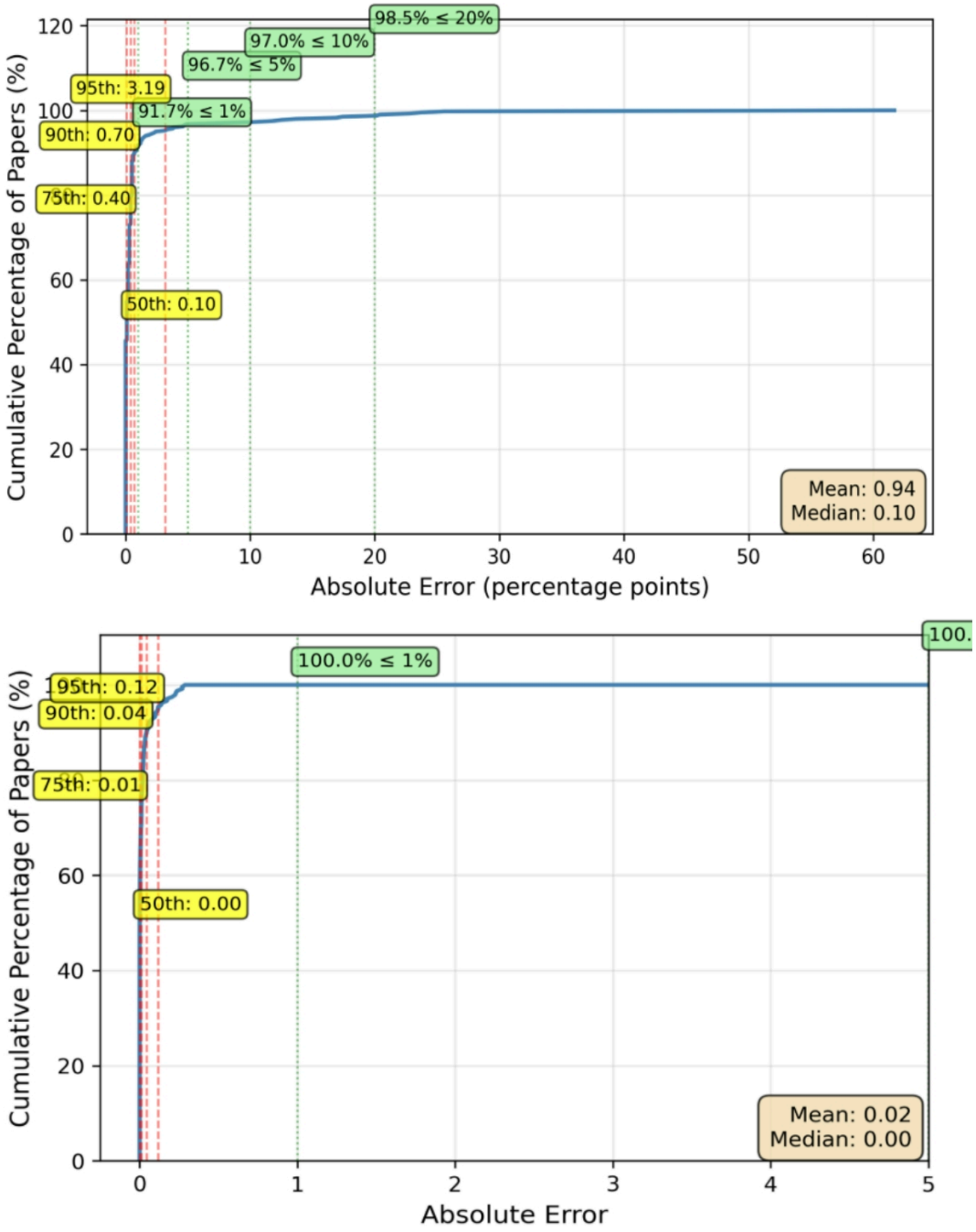


Figure 3

Agreement in heterogeneity estimates between published benchmark analyses and AI-assisted analyses.

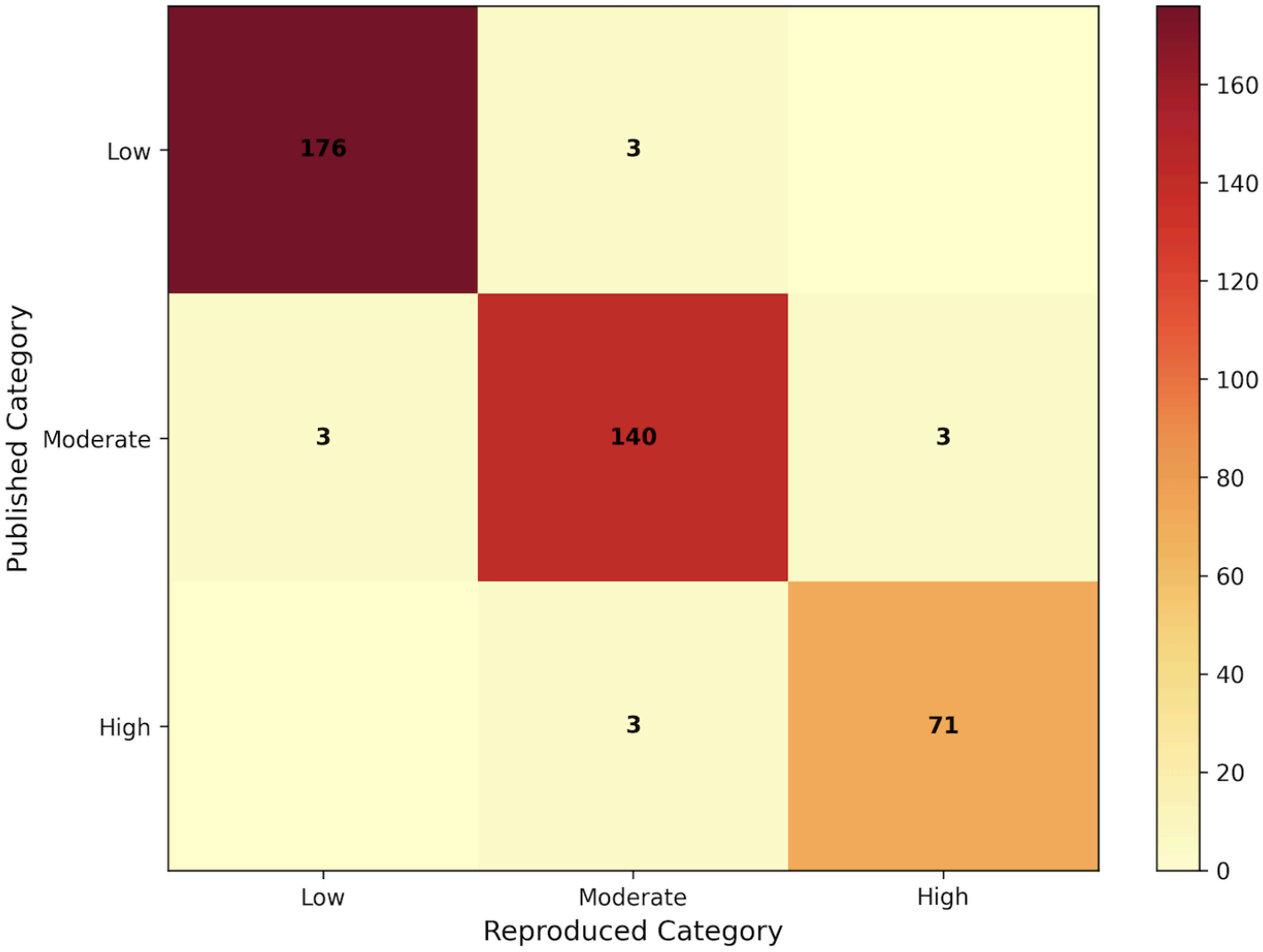


Figure 4

Changes in heterogeneity classification between published benchmark analyses and AI-assisted analyses.