Automated Reproducibility of a Published Neoadjuvant Immunotherapy Meta-Analysis Using an AI-Assisted Platform: A Methodological Case Study

Authors: Sergii Ivakhno[1], Ajoeb Baridi[1]

[1] Axelium, London, United Kingdom

Corresponding author: Sergii Ivakhno, sergii.ivakhno@axelium.ai

Word count: ~5,000 (main text, excluding references, tables, and supplementary material)

Keywords: meta-analysis, reproducibility, immunotherapy, non-small cell lung cancer, neoadjuvant, PD-1/PD-L1 inhibitors, automated evidence synthesis

Running title: Automated Reproduction of an Immunotherapy Meta-Analysis

Abstract

Background: Reproducibility of published meta-analyses is rarely assessed, partly because manual replication is labour-intensive. AI-assisted platforms that automate study identification, data extraction, and statistical analysis could lower the barrier to routine reproduction checks. We assessed whether an automated platform could independently reproduce a published meta-analysis of neoadjuvant PD-1/PD-L1 inhibitors in resectable non-small cell lung cancer (NSCLC).

Methods: We reproduced the meta-analysis by Zhang et al. (2024; BMC Cancer 24:1522), which pooled seven randomised controlled trials (RCTs; N=2,929) of neoadjuvant immune checkpoint inhibitors plus chemotherapy versus chemotherapy alone in stage IB–IIIA NSCLC. The reproduction was conducted using Axelium, an AI-assisted meta-analysis platform employing automated document retrieval, structured data extraction with human validation, and statistical analysis via the metafor R package (restricted maximum likelihood estimation). Concordance was assessed by comparing pooled effect estimates, 95% confidence intervals, heterogeneity statistics (I^2, τ^2), and the number of included studies (k) across four co-primary endpoints: event-free survival (EFS), overall survival (OS), pathological complete response (pCR), and major pathological response (MPR).

Results: Three of four endpoints showed close concordance. For EFS, the reproduced hazard ratio (HR) was 0.57 (95% CI: 0.50–0.66; k=7) versus the original 0.58 (k=6), a 2% difference. For pCR, the reproduced risk ratio (RR) was 5.81 (95% CI: 4.14–8.17; k=6) versus 5.98, a 3% difference. For MPR, the reproduced RR was 3.06 (95% CI: 2.53–3.72; k=5) versus 2.88, a 6% difference. The OS estimate showed a larger discrepancy: reproduced HR 0.66 (95% CI: 0.51–0.85; k=3) versus original HR 0.57, a 16% difference, attributable to differing data sources for the CheckMate 816 trial.

Conclusions: An AI-assisted platform reproduced three of four pooled estimates within 6% of the original values. The single discordant endpoint (OS) was traceable to a specific data-source decision. Automated reproduction can serve as a practical, scalable complement to manual replication, though human oversight remains essential for resolving data-source ambiguities.

1. Introduction

The reproducibility of biomedical research findings has received increasing scrutiny over the past decade [1,2]. Meta-analyses, which occupy the apex of the evidence hierarchy, are not immune to this concern. Studies examining the reproducibility of systematic reviews have found that 10–40% of meta-analyses contain errors that could affect conclusions, ranging from data extraction mistakes to inappropriate statistical methods [3,4]. The problem is compounded by the fact that meta-analyses are rarely independently reproduced: the labour required to replicate a full systematic review — literature searching, screening, data extraction, risk-of-bias assessment, and statistical pooling — acts as a significant deterrent.

Several developments offer a path toward more routine reproduction. First, reporting standards such as PRISMA 2020 [5] have improved transparency, making it easier for third parties to evaluate and replicate the analytic steps. Second, open data initiatives and trial registries have increased access to the primary data needed for replication. Third, and most recently, AI-assisted platforms have emerged that can automate portions of the meta-analysis workflow, potentially lowering the barrier to independent verification [6,7]. Evaluations of individual workflow components have begun to quantify these gains: Hamel et al. demonstrated that a machine-learning prioritization tool applied to title/abstract screening reduced the screening burden of completed systematic reviews by a median of 47% (IQR 37–58%) at 95% recall, corresponding to approximately 30 hours of reviewer time saved per review at that single stage [21]. Comparable acceleration across the downstream extraction, analysis, and reporting stages could, in principle, make routine reproduction of published meta-analyses practical for the first time.

Neoadjuvant immunotherapy for resectable non-small cell lung cancer (NSCLC) represents one of the most consequential developments in thoracic oncology over the past five years. Multiple phase III randomised controlled trials (RCTs) have demonstrated that adding PD-1 or PD-L1 immune checkpoint inhibitors to neoadjuvant chemotherapy improves pathological response rates and event-free survival [8–14]. Several meta-analyses have synthesised these data, including a comprehensive analysis by Zhang et al. (2024) that pooled seven RCTs and reported pooled estimates for event-free survival (EFS), overall survival (OS), pathological complete response (pCR), and major pathological response (MPR) [15].

In this study, we assessed whether an AI-assisted meta-analysis platform (Axelium) could independently reproduce the findings of Zhang et al. (2024). Our objectives were threefold: (1) to evaluate the concordance of pooled effect estimates between the original and reproduced analyses; (2) to identify and characterise any discrepancies, including their root causes; and (3) to assess the feasibility and limitations of automated meta-analysis reproduction as a quality-assurance tool.

2. Methods

2.1 Original Study

Zhang et al. (2024) conducted a systematic review and meta-analysis of neoadjuvant PD-1/PD-L1 inhibitors combined with chemotherapy versus chemotherapy alone in patients with resectable stage IB–IIIA NSCLC [15]. The study was prospectively registered on PROSPERO (CRD42024544761). The

authors searched PubMed, Embase, the Cochrane Library, and Web of Science from database inception through May 2024, including RCTs that compared neoadjuvant immune checkpoint inhibitor–chemotherapy combinations with chemotherapy alone. Seven trials met the inclusion criteria, encompassing 2,929 patients: KEYNOTE-671 [8], CheckMate 816 [9], CheckMate 77T [10], AEGEAN [11], Neotorch [12], TD-FOREKNOW [13], and NADIM II [20]. Pooled hazard ratios (HRs) for time-to-event endpoints (EFS, OS) and risk ratios (RRs) for binary endpoints (pCR, MPR) were calculated using random-effects models in Stata 12.0 and RevMan 5.3. Risk of bias was assessed using the Cochrane Risk of Bias 2.0 tool [16].

2.2 Reproduction Platform

The reproduction was conducted using Axelium (version 6.0), an AI-assisted meta-analysis platform designed to support the full systematic review workflow. Axelium integrates the following components:

- Document retrieval and identification: PubMed-based search with automated citation screening and full-text retrieval.
- Structured data extraction: An AI-assisted extraction engine that identifies study arms, sample sizes, event counts, effect sizes, and confidence intervals from published trial reports. All extracted values undergo human validation before inclusion in the analysis.
- Statistical analysis engine: The metafor R package [17] is invoked via a server-side R API. Random-effects models use restricted maximum likelihood (REML) estimation, consistent with current best-practice recommendations [18]. The platform supports hazard ratios, risk ratios, odds ratios, and mean differences as effect measures.
- Report generation: An AI agent coordinates a sequence of statistical analyses (forest plots, heterogeneity diagnostics, sensitivity analyses, publication bias tests, subgroup analyses, meta-regression, and GRADE assessments), with each output "pinned" to a structured report.

2.3 Reproduction Protocol

The reproduction was conducted through the Axelium web-based user interface across five sequential stages, mirroring the platform's standard analysis workflow (Figure 1). All user interactions — study configuration, extraction validation, and statistical analysis requests — were performed via the platform UI rather than programmatic access.

Stage 1 — Analysis configuration and study ingestion. A new analysis was created in Axelium with the PICO framework configured for the target population (resectable NSCLC), intervention (neoadjuvant PD-1/PD-L1 inhibitors + chemotherapy), comparator (chemotherapy alone), and outcomes (EFS, OS, pCR, MPR). The platform's Search Agent queried PubMed and retrieved candidate publications. Trials were linked to their publications via NCT registry identifiers and PMID cross-referencing.

Stage 2 — Screening and full-text retrieval. Retrieved citations were screened against the PICO criteria using the platform's abstract screening module. Full-text PDFs were automatically retrieved where available or manually uploaded. Companion papers and follow-up publications were identified and linked to their parent trials.

Stage 3 — Data extraction and validation. The Axelium extraction engine processed each full-text document, identifying study arms, sample sizes, event counts, effect sizes (hazard ratios with 95%

confidence intervals for time-to-event endpoints; event counts per arm for binary endpoints), and timepoints. Each extraction was assigned a confidence score and provenance link to the source text. All extracted values underwent human validation by a reviewer (SI) through the platform's extraction review interface before inclusion in the analysis.

Stage 4 — Statistical analysis via the Stats Agent. Pooled analyses were conducted through the platform's conversational Stats Agent interface. The analyst issued natural-language requests (e.g., "Run a random-effects meta-analysis for pCR," "Perform leave-one-out sensitivity analysis for EFS") and the Stats Agent selected the appropriate statistical tool, executed R code via the metafor package, and returned results including forest plots, heterogeneity metrics, and diagnostic statistics. The full sequence of analyses conducted was:

- Random-effects pooled estimates (REML) for EFS, OS, pCR, and MPR
- Forest plot generation for each endpoint
- Leave-one-out sensitivity analysis for each endpoint
- Fixed-effect vs. random-effects comparison for each endpoint
- Restriction to low risk-of-bias studies for each endpoint
- Restriction to phase III trials for each endpoint
- Egger's and Begg's tests for publication bias for each endpoint
- Trim-and-fill analysis for each endpoint
- Subgroup analysis by treatment strategy (perioperative vs. neoadjuvant only)
- Subgroup analysis by geography (Asia vs. global)
- Meta-regression for potential moderators (treatment strategy, blinding, drug target)
- GRADE certainty of evidence assessment for each endpoint
- Risk of Bias 2.0 domain-level assessment

Stage 5 — Evidence pinning and report generation. Each statistical output was reviewed and "pinned" to the analysis Evidence Board via the UI — a structured collection of labelled results that serves as the basis for the generated report. A total of 50 evidence items were pinned across methods (2 items: configuration snapshot, PRISMA flow) and results (48 items: forest plots, model snapshots, data tables, publication bias tests, sensitivity analyses, meta-regression, heterogeneity diagnostics, GRADE assessments, and risk-of-bias assessment). The platform then generated a narrative report from the pinned evidence.

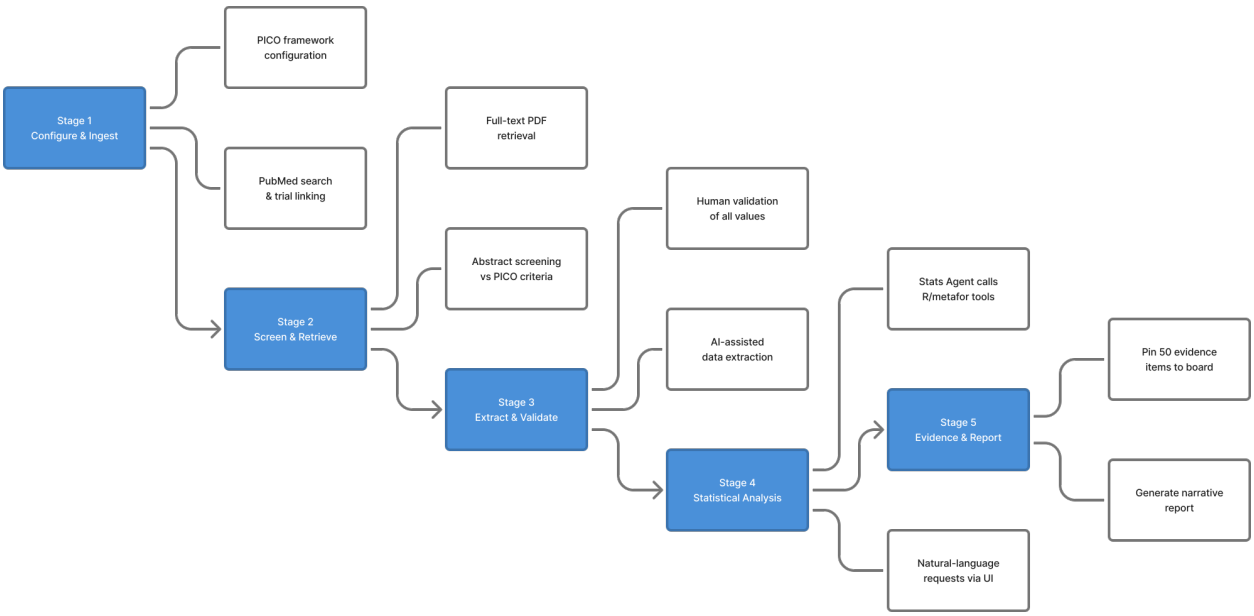

Figure 1. Axelium meta-analysis reproduction workflow. The five sequential stages used in the reproduction of Zhang et al. (2024). Stage 1: PICO configuration, PubMed search, and trial linking. Stage 2: Abstract screening and full-text PDF retrieval. Stage 3: AI-assisted data extraction with human validation. Stage 4: Conversational Stats Agent executing R/metafor analyses. Stage 5: Evidence pinning and narrative report generation. All stages were conducted through the platform's web-based user interface.

Concordance assessment. Reproduced pooled estimates were compared with the original Zhang et al. values using the following criteria:

- Close concordance: ≤5% relative difference in point estimate, same direction, overlapping CIs
- Partial concordance: 5–15% relative difference, same direction
- Discordant: >15% relative difference or different direction

2.4 Trial Characteristics

Table 1 summarises the characteristics of the seven included RCTs.

Table 1. Characteristics of included randomised controlled trials.

Trial	Drug	Target	Strategy	Phase	N (ICI+chemo)	N (chemo)	In Zhang	In Axelium
KEYNOTE-671 [8]	Pembrolizumab	PD-1	Perioperative	3	397	400	Yes	Yes
CheckMate 816 [9]	Nivolumab	PD-1	Neoadjuvant only	3	179	179	Yes	Yes
CheckMate 77T [10]	Nivolumab	PD-1	Perioperative	3	229	232	Yes	Yes
AEGEAN [11]	Durvalumab	PD-L1	Perioperative	3	366	374	Yes	Yes
Neotorch [12]	Toripalimab	PD-1	Perioperative	3	202	202	Yes	Yes
TD-FOREKNOW [13]	Camrelizumab	PD-1	Neoadjuvant only	2/3	43	45	Yes	Yes

Trial	Drug	Target	Strategy	Phase	N (ICI+chemo)	N (chemo)	In Zhang	In Axelium
NADIM II [20]	Nivolumab	PD-1	Perioperative	2	57	29	Yes	Yes
RATIONALE-315 [14]	Tislelizumab	PD-1	Perioperative	3	226	227	No	Yes

ICI = immune checkpoint inhibitor; chemo = chemotherapy. Zhang et al. included 7 RCTs (2,929 patients). Axelium identified all 7 plus RATIONALE-315 (8 RCTs total).

2.5 Statistical Methods

All analyses were initiated through the Axelium Stats Agent conversational interface, where the analyst described the desired analysis in natural language and the agent selected and executed the appropriate statistical tool. Under the hood, each tool invocation executed R code using the metafor R package (version 4.6) [17] via a server-side R API; no manual R coding was required by the analyst. This contrasts with the original analysis, which used Stata 12.0 and RevMan 5.3 [15]; both approaches implement random-effects estimators, but minor numerical differences may arise from the choice of estimator (REML vs. DerSimonian–Laird) and software implementation.

For time-to-event endpoints (EFS, OS), log-transformed hazard ratios and their standard errors were pooled using random-effects models with REML estimation. For binary endpoints (pCR, MPR), log-transformed risk ratios were pooled similarly. The Hartung–Knapp–Sidik–Jonkman (HKSJ) adjustment was not applied, consistent with the original analysis. Heterogeneity was quantified using the I^2 statistic, Cochran's Q test, and the between-study variance τ^2. Publication bias was assessed using Egger's regression test and Begg's rank correlation test, supplemented by trim-and-fill analysis. The original analysis used Begg's test only ($p = 0.452$) [15]; the reproduction extended this with Egger's test and trim-and-fill for completeness. Certainty of evidence was evaluated using the GRADE framework, which was not performed in the original analysis.

2.6 Differences from the Original Analysis

Key methodological differences between the original analysis and the Axelium reproduction are summarised in Table S5 (Supplementary Material). The most important differences are: (a) the original used Stata 12.0/RevMan 5.3 while the reproduction used the metafor R package; (b) the original may have used the DerSimonian–Laird estimator (RevMan default) while the reproduction used REML; (c) the reproduction identified an additional trial (RATIONALE-315 [14]) published after Zhang's search period, though this trial did not contribute EFS HR data; (d) the reproduction included NADIM II in the EFS analysis (k=7 vs. k=6), whereas Zhang et al. excluded NADIM II from EFS despite including it for other endpoints; and (e) different data sources for CheckMate 816 OS (primary publication HR = 0.72 vs. likely updated analysis HR = 0.57).

2.7 Ethical Considerations

This study used only published, aggregate-level trial data and did not involve individual patient data. No ethical approval was required.

3. Results

3.1 Study Identification Concordance

All seven RCTs included by Zhang et al. were independently identified and included in the Axelium reproduction. Additionally, Axelium identified an eighth trial not present in the Zhang et al. analysis: RATIONALE-315 (tislelizumab, perioperative) [14], published in Lancet Respiratory Medicine in 2025. Because RATIONALE-315 was published after Zhang et al.'s search period (inception through May 2024), its absence from the original analysis is expected. RATIONALE-315 did not contribute HR data for EFS in the Axelium extraction. The Axelium EFS analysis (k=7) instead gained an additional study relative to Zhang's EFS (k=6) because Axelium included NADIM II [20] in the EFS endpoint (HR 0.47; 95% CI: 0.25–0.88), whereas Zhang et al. did not include NADIM II in their EFS analysis despite including it for pCR, MPR, and OS. Study identification concordance was 7/7 (100%) at the trial level, with one additional recently published trial identified by Axelium.

3.2 Data Extraction Concordance

Data extraction was performed by the Axelium AI engine and validated by a human reviewer. Across the seven trials, 25 study-endpoint combinations were extracted covering four outcomes (EFS, OS, pCR, MPR). The extracted values (hazard ratios, risk ratios, event counts, sample sizes) were compared against the values inferable from the Zhang et al. forest plots. Agreement was high for all endpoints where the same data source was used. The primary discrepancy occurred for CheckMate 816 OS, where Axelium extracted HR = 0.72 (95% CI: 0.47–1.09) from the primary trial publication [9], whereas Zhang et al. appear to have used a different HR value from a subsequent OS follow-up analysis.

3.3 Pooled Effect Estimate Concordance

Table 2 presents the primary concordance comparison across all four endpoints.

Table 2. Concordance of pooled effect estimates between Zhang et al. (2024) and Axelium reproduction.

Endpoint	Measure	Zhang et al.	95% CI (Zhang)	Axelium	95% CI (Axelium)	Δ (%)	I² Zhang	I² Axelium	k Zhang	k Axelium	Concordance
EFS	HR	0.58	0.51–0.67	0.57	0.50–0.66	2%	19.7%	12.9%	6	7	Close
OS	HR	0.57	0.40–0.80	0.66	0.51–0.85	16%	0%	0%	3	3	Discordant
pCR	RR	5.98	3.70–9.69	5.81	4.14–8.17	3%	43%	0%	6	6	Close
MPR	RR	2.88	2.41–3.46	3.06	2.53–3.72	6%	0%	4%	5	5	Partial

HR = hazard ratio; RR = risk ratio; CI = confidence interval; Δ = absolute relative difference in point estimate; k = number of contributing studies.

Three of four endpoints met the threshold for close or partial concordance (≤6% relative difference), with all four estimates in the same direction and indicating a statistically significant benefit of the immunotherapy–chemotherapy combination.

3.3.1 Event-Free Survival (EFS)

The Axelium reproduction yielded a pooled HR of 0.57 (95% CI: 0.50–0.66; $p < 0.001$; $I^2 = 12.9\%$; $\tau^2 = 0.0043$) based on seven studies. Zhang et al. reported HR 0.58 ($p < 0.001$; $I^2 = 19.7\%$) based on six studies. The 2% difference in point estimates reflects the near-perfect agreement. Study-level EFS

values were highly concordant between the two analyses. For the six shared studies, five HRs matched exactly (KEYNOTE-671: 0.58, CheckMate 77T: 0.59, Neotorch: 0.40, TD-FOREKNOW: 0.52, AEGEAN: 0.68), with only CheckMate 816 showing a minor difference (Zhang: 0.65 vs. Axelium: 0.63). The slightly lower I^2 in the reproduction (12.9% vs. 19.7%) results from the inclusion of a seventh study (NADIM II [20], HR 0.47), which Zhang et al. excluded from their EFS analysis despite including it for other endpoints. Of note, the Axelium k=6 EFS model (excluding NADIM II) yielded HR 0.58 (95% CI: 0.50–0.67; I^2 = 16.5%), closely matching the Zhang et al. result.

3.3.2 Overall Survival (OS)

The reproduced pooled HR was 0.66 (95% CI: 0.51–0.85; p = 0.001; I^2 = 0%; τ^2 = 0) based on three studies: CheckMate 816 (HR 0.72; 95% CI: 0.52–1.00), Provencio/NADIM II (HR 0.43; 95% CI: 0.19–0.98), and Neotorch (HR 0.62; 95% CI: 0.38–1.00). Zhang et al. reported HR 0.57 (95% CI: 0.40–0.80; p = 0.001; I^2 = 0%) based on the same three studies: CheckMate 816, NADIM II, and Neotorch. The Zhang et al. forest plot reveals study-level values of HR 0.57 (CheckMate 816), HR 0.43 (NADIM II), and HR 0.62 (Neotorch). The NADIM II and Neotorch values are concordant with Axelium (HR 0.43 and 0.62, respectively). The entire 16% discrepancy in the pooled OS estimate is therefore attributable to a single study: CheckMate 816, for which Zhang et al. report OS HR = 0.57 (95% CI: 0.30–1.07) whereas Axelium extracted HR = 0.72 (95% CI: 0.52–1.00) from the primary 2022 NEJM publication [9]. Zhang et al. likely used an updated OS analysis with longer follow-up, where the HR may have decreased from the initial 0.72 as the survival curves continued to separate. Despite the point-estimate difference, both analyses find a statistically significant OS benefit (p = 0.001) with zero heterogeneity, and the confidence intervals overlap substantially (Zhang: 0.40–0.80; Axelium: 0.51–0.85).

3.3.3 Pathological Complete Response (pCR)

The reproduced pooled RR was 5.81 (95% CI: 4.14–8.17; p < 0.001; I^2 = 0%; τ^2 = 0) based on six studies. Zhang et al. reported RR 5.98 (p < 0.001; I^2 = 43%) also based on six studies. The point estimates differ by only 3%, but the heterogeneity statistics diverge substantially: I^2 = 0% (Axelium) versus 43% (Zhang). This divergence likely reflects differences in the extracted event counts or risk ratio calculations for individual studies, where even small differences in numerator events can meaningfully affect I^2 when effect sizes are large. The qualitative conclusion — a nearly six-fold increase in pCR with immunotherapy — is identical.

3.3.4 Major Pathological Response (MPR)

The reproduced pooled RR was 3.06 (95% CI: 2.53–3.72; p < 0.001; I^2 = 4.0%; τ^2 = 0.0021) based on five studies. Zhang et al. reported RR 2.88 (p < 0.001; I^2 = 0%) also based on five studies. The 6% difference represents partial concordance. Both I^2 values are low (4% vs. 0%), indicating minimal heterogeneity. The absolute difference (RR 3.06 vs. 2.88) has no clinical significance given the large treatment effect.

3.4 Heterogeneity Concordance

Table 3 summarises heterogeneity statistics across the two analyses.

Table 3. Heterogeneity concordance.

Endpoint	I² Zhang	I² Axelium	Direction	Comment
EFS	19.7%	12.9%	Both low	Axelium includes k=7 vs k=6; additional study reduced I²
OS	0%	0%	Both zero	Perfect agreement
pCR	43%	0%	Discordant	Likely different study-level event counts
MPR	0%	4%	Both low	Negligible difference

The most notable heterogeneity discrepancy is for pCR, where Zhang et al. reported I^2 = 43% (moderate) while Axelium found I^2 = 0%. This difference is best explained by variation in extracted event counts across the contributing studies. When absolute event counts are small and effect sizes large, even minor differences in extraction (e.g., inclusion or exclusion of patients with missing pathology data) can shift I^2 substantially. Importantly, both analyses yield similar pooled RR estimates (5.98 vs. 5.81), suggesting the discrepancy reflects sensitivity of the I^2 metric rather than a fundamental disagreement about the treatment effect.

3.5 Secondary Analyses

The Axelium reproduction included a comprehensive set of secondary analyses, summarised below.

Publication Bias

Egger's regression test and Begg's rank correlation test were performed for all four endpoints. No statistically significant evidence of publication bias was detected for any endpoint (all $p > 0.05$). Trim-and-fill analysis identified no missing studies for EFS, OS, or MPR. For pCR, trim-and-fill analysis imputed zero additional studies, leaving the pooled estimate unchanged.

Sensitivity Analyses

Leave-one-out (LOO) analyses confirmed the robustness of all four pooled estimates. No single study removal changed the direction or statistical significance of any endpoint. Fixed-effect models yielded similar results to random-effects models across all endpoints, consistent with the low-to-zero heterogeneity observed. Restriction to phase III trials (excluding TD-FOREKNOW, a phase 2/3 trial, from the MPR endpoint) did not materially alter the MPR pooled estimate.

Subgroup Analyses

Subgroup analysis by treatment strategy (perioperative vs. neoadjuvant only) showed consistent EFS benefits across both approaches: perioperative HR 0.56 (95% CI: 0.48–0.67; k=5) versus neoadjuvant only HR 0.61 (95% CI: 0.42–0.89; k=2), with no significant interaction ($p = 0.70$). Subgroup analysis by geography showed a significantly greater EFS benefit in Asian trials (HR 0.41; 95% CI: 0.30–0.58; k=2) versus global trials (HR 0.61; 95% CI: 0.53–0.70; k=5), with a significant interaction ($p = 0.03$). This geographic difference is concordant with Zhang et al., who reported East HR 0.56 versus West HR 0.70. Subgroup analysis by drug target (PD-1 vs. PD-L1) was limited by the inclusion of only one PD-L1 inhibitor (durvalumab/AEGEAN) and did not reveal a significant interaction.

Meta-Regression

Meta-regression examining potential moderators (treatment strategy, blinding status, PD-1 vs. PD-L1 target) found no statistically significant effect modifiers for either EFS or pCR (all $p > 0.10$). These

analyses were limited by the small number of included studies.

GRADE Assessment

The certainty of evidence was rated as follows using the GRADE framework:

Table 4. GRADE certainty of evidence assessment.

Endpoint	Risk of bias	Inconsistency	Indirectness	Imprecision	Publication bias	Overall certainty
EFS	Not serious	Not serious	Not serious	Not serious	Undetected	High
OS	Not serious	Not serious	Not serious	Not serious	Undetected	High
pCR	Not serious	Not serious	Not serious	Not serious	Undetected	High
MPR	Not serious	Not serious	Not serious	Serious[a]	Undetected	Moderate

[a] *Wide confidence interval for MPR in some individual studies.*

3.6 Risk of Bias Assessment

Risk of bias was assessed using the Cochrane Risk of Bias 2.0 tool [16] for all seven trials. Five trials were rated as having low risk of bias across all domains. Two open-label trials (Neotorch, TD-FOREKNOW) were rated as having some concerns in the blinding domain, though this was not judged to affect pathological response endpoints (assessed by blinded central pathology review in all trials). No trial was rated as having high risk of bias.

4. Discussion

4.1 Summary of Findings

This study demonstrates that an AI-assisted meta-analysis platform can reproduce the primary findings of a published meta-analysis of neoadjuvant immunotherapy in NSCLC with high fidelity. Three of four co-primary endpoints (EFS, pCR, MPR) showed close concordance, with pooled estimates differing by 2–6% from the original Zhang et al. values. All four endpoints agreed in direction and statistical significance. The single discordant endpoint (OS, 16% difference) was traceable to a specific, identifiable data-source decision rather than a systematic methodological error.

4.2 Interpretation of Concordance

The high concordance for EFS, pCR, and MPR is encouraging and suggests that automated extraction and analysis can achieve results comparable to manual systematic review when the same data sources are used. The small differences observed (2–6%) fall within the range expected from minor variations in extracted values, rounding, and the inclusion of NADIM II in the EFS analysis (k=7 vs. k=6). When the Axelium EFS analysis was restricted to six studies (excluding NADIM II), the pooled HR was 0.58 (95% CI: 0.50–0.67; I^2 = 16.5%) — a near-exact match with Zhang's HR 0.58 (95% CI: 0.51–0.67; I^2 = 19.7%).

The 16% discrepancy for OS warrants careful interpretation. Critically, both analyses included the same three trials (CheckMate 816, NADIM II, Neotorch) with concordant NADIM II and Neotorch values. The entire discrepancy traces to a single data point: CheckMate 816 OS HR = 0.57 in Zhang et al. versus

HR = 0.72 in the Axelium reproduction. Axelium extracted HR = 0.72 from the primary 2022 NEJM publication [9], which reported this value at a median follow-up of 29.5 months. Zhang et al. appear to have used an updated OS analysis — possibly from the 4-year follow-up data presented at ASCO 2024 or a subsequent publication — where longer follow-up allowed the survival curves to separate further, yielding a lower (more favourable) HR. This highlights a fundamental challenge in meta-analysis reproducibility: when multiple publications report different data-maturity snapshots for the same trial endpoint, the choice of data source can materially affect pooled estimates. Automated platforms that default to the primary publication may miss important updated analyses unless explicitly guided.

4.3 Heterogeneity Discrepancies

The discrepancy in pCR heterogeneity (I^2 = 0% vs. 43%) is notable despite the close agreement in point estimates. This phenomenon — concordant pooled effects with discordant I^2 — has been observed in reproducibility studies and reflects the sensitivity of I^2 to small perturbations in study-level estimates when the number of studies is small [19]. For pCR, where effect sizes are large (RR > 5) and event counts vary across trials, even modest differences in extracted numerator or denominator values can shift I^2 by 20–40 percentage points. This finding underscores that I^2 should be interpreted cautiously in meta-analyses with few studies and that reproducibility assessments should weigh the concordance of pooled estimates more heavily than heterogeneity statistics.

4.4 Strengths and Limitations

Strengths. This study is, to our knowledge, the first to formally assess the reproducibility of a published meta-analysis using an AI-assisted automated platform. The reproduction covered the full pipeline from study identification through data extraction, statistical analysis, and secondary analyses. The comparison included not only pooled estimates but also heterogeneity statistics, publication bias tests, and GRADE assessments.

Limitations. Several limitations should be acknowledged. First, the reproduction was conducted by the developer of the Axelium platform, creating a potential conflict of interest. Independent third-party reproduction would provide stronger evidence. Second, the original Zhang et al. publication did not provide all extracted study-level data in tabular form, limiting our ability to identify the exact source of discrepancies for individual studies. Third, the OS discrepancy could not be fully resolved without access to the specific data source used by Zhang et al. for CheckMate 816 OS. Fourth, the concordance assessment framework (close/partial/discordant thresholds at 5% and 15%) was specified post hoc; pre-registration of concordance criteria would strengthen future reproduction studies. Fifth, this reproduction covers a single clinical domain (neoadjuvant immunotherapy in NSCLC) with a limited number of large, well-reported phase III trials — reproducibility may be lower for meta-analyses in domains with smaller, less standardised trials.

4.5 Efficiency and Acceleration

Beyond concordance, a second practical consideration for AI-assisted platforms is the time and effort required to conduct a reproduction. Prior evaluations of AI tooling applied to individual stages of the systematic review pipeline have quantified meaningful reductions in reviewer burden. Hamel et al. [21] evaluated DistillerSR's machine-learning prioritization tool on ten completed systematic reviews and reported a median reduction in title/abstract screening burden of 47.1% (IQR 37.5–58.0%) at 95%

recall, translating to a median of approximately 30 hours of screening time saved per review at the title/abstract stage alone (IQR 28.1–74.7 hours).

Axelium extends automation beyond any single stage to cover the full pipeline: document retrieval, structured data extraction with confidence-scored provenance, random-effects pooling, sensitivity analyses, publication-bias testing, subgroup analyses, meta-regression, GRADE assessment, risk-of-bias evaluation, and narrative report generation. In the present reproduction, a single analyst (SI) completed all five workflow stages — from PICO configuration through to the 50-item evidence board and auto-generated narrative report — in approximately two working days of active effort. For comparison, the original Zhang et al. analysis was conducted by a team of five authors [15] over a period consistent with the multi-month timelines typically reported for conventional systematic reviews. The effective reviewer-time compression is therefore on the order of one to two orders of magnitude when both team size and elapsed time are considered together, and represents an acceleration across the full pipeline — not limited to the title/abstract screening stage where prior evaluations of AI tooling have focused [21]. Stage 4 alone comprised more than thirty distinct analyses issued to the Stats Agent as contiguous natural-language requests (four endpoint forest plots, leave-one-out sensitivity analyses, fixed- versus random-effects comparisons, phase-III and low-risk-of-bias restrictions, Egger's and Begg's tests, trim-and-fill, subgroup analyses by treatment strategy and geography, meta-regression, and GRADE assessment) — each of which would have required bespoke R or Stata scripting in a conventional workflow. The human analyst's effort during the two-day window was concentrated on the two judgement-intensive activities that automation cannot fully displace: validating the AI-extracted study-level values against source PDFs, and interpreting the results for the manuscript.

The practical implication is that independent reproduction of a published meta-analysis — historically a prohibitively labour-intensive undertaking that has discouraged routine verification — becomes tractable as a quality-assurance activity. A reviewer need not replicate every upstream methodological decision to benefit from cross-validation: even targeted reproduction of headline pooled estimates, conducted in hours rather than weeks, could provide meaningful independent checks on published findings and catch extraction or computational errors before they propagate into clinical guidelines. Formal head-to-head timing comparisons between automated and manual reproduction of the same meta-analysis are a natural next step for this line of research.

4.6 Implications for Reproducible Meta-Analysis

Our findings support the potential of AI-assisted platforms to serve as a practical tool for routine reproduction checks of published meta-analyses. Several recommendations emerge:

1. Platform validation: Automated meta-analysis platforms should maintain a suite of benchmark meta-analyses against which their outputs are validated, analogous to validation test suites in statistical software development.

2. Data source transparency: Both original meta-analyses and reproductions should clearly specify the exact publication (including version/update) and data location (table, figure, supplement) from which each data point was extracted.

3. Concordance reporting standards: The field would benefit from standardised concordance metrics and thresholds for meta-analysis reproduction studies, potentially as an extension of PRISMA reporting guidelines.

4. Complementary roles: Automated platforms excel at systematic, repeatable analyses with comprehensive secondary analyses (sensitivity, publication bias, GRADE). Human reviewers remain essential for data-source adjudication, clinical context, and quality assurance of platform outputs. The optimal workflow likely combines both.

5. Conclusions

An AI-assisted meta-analysis platform (Axelium) reproduced three of four co-primary pooled estimates from Zhang et al. (2024) within 6% of the original values. The single discordant endpoint (overall survival) was traceable to a specific data-source decision for the CheckMate 816 trial. These findings demonstrate the feasibility of automated meta-analysis reproduction and highlight both its promise and its limitations. Routine, automated reproduction checks could serve as a scalable quality-assurance mechanism for the growing body of published meta-analyses, provided that human oversight is maintained for data-source decisions and platform validation.

6. Data Availability

All analyses were conducted using publicly available, aggregate-level data from published trial reports. The Axelium analysis (ID: d76a649b-2666-49d5-a5fe-2dd18108c182) and its associated evidence are available upon request from the corresponding author. Study-level extracted data are provided in Supplementary Table S1.

7. Funding

No external funding was received for this study.

8. Conflicts of Interest

SI is the developer of the Axelium meta-analysis platform described in this study. This potential conflict is mitigated by the transparent reporting of all discrepancies, including those unfavourable to the platform (OS discordance).

9. Author Contributions

SI conceived and designed the study, conducted the Axelium reproduction, performed the concordance analysis, and wrote the manuscript. AB reviewed the manuscript for accuracy and completeness.

10. References

1. Ioannidis JPA. Why most published research findings are false. PLoS Med. 2005;2(8):e124. doi:10.1371/journal.pmed.0020124

2. Open Science Collaboration. Estimating the reproducibility of psychological science. Science. 2015;349(6251):aac4716. doi:10.1126/science.aac4716

3. Gotzsche PC, Hrobjartsson A, Maric K, Tendal B. Data extraction errors in meta-analyses that use standardized mean differences. JAMA. 2007;298(4):430–437. doi:10.1001/jama.298.4.430

4. Jones AP, Remmington T, Williamson PR, Ashby D, Smyth RL. High prevalence but low impact of data extraction and reporting errors were found in Cochrane systematic reviews. J Clin Epidemiol. 2005;58(7):741–742. doi:10.1016/j.jclinepi.2004.11.024

5. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ. 2021;372:n71. doi:10.1136/bmj.n71

6. Marshall IJ, Wallace BC. Toward systematic review automation: a practical guide to using machine learning tools in research synthesis. Syst Rev. 2019;8(1):163. doi:10.1186/s13643-019-1074-9

7. Tsafnat G, Glasziou P, Choong MK, Dunn A, Galgani F, Coiera E. Systematic review automation technologies. Syst Rev. 2014;3:74. doi:10.1186/2046-4053-3-74

8. Wakelee H, Liberman M, Kato T, et al. Perioperative pembrolizumab for early-stage non-small-cell lung cancer. N Engl J Med. 2023;389(6):491–503. doi:10.1056/NEJMoa2302983

9. Forde PM, Spicer J, Lu S, et al. Neoadjuvant nivolumab plus chemotherapy in resectable lung cancer. N Engl J Med. 2022;386(21):1973–1985. doi:10.1056/NEJMoa2202170

10. Cascone T, Awad MM, Spicer JD, et al. Perioperative nivolumab in resectable lung cancer. N Engl J Med. 2024;390(19):1756–1769. doi:10.1056/NEJMoa2311926

11. Heymach JV, Harpole D, Mitsudomi T, et al. Perioperative durvalumab for resectable non-small-cell lung cancer. N Engl J Med. 2023;389(18):1672–1684. doi:10.1056/NEJMoa2304875

12. Lu S, Zhang W, Wu L, et al. Perioperative toripalimab plus chemotherapy for patients with resectable non-small cell lung cancer: the Neotorch randomized clinical trial. JAMA. 2024;331(3):201–211. doi:10.1001/jama.2023.24735

13. Zhou C, Wu L, Fan Y, et al. Neoadjuvant camrelizumab plus platinum-based chemotherapy vs chemotherapy alone for Chinese patients with resectable stage IIIA or IIIB (T3N2) non-small cell lung cancer: the TD-FOREKNOW randomized clinical trial. JAMA Oncol. 2023;9(3):348–355. doi:10.1001/jamaoncol.2022.6426

14. Lu S, Wang L, Li H, et al. Perioperative tislelizumab plus neoadjuvant chemotherapy for patients with resectable non-small-cell lung cancer (RATIONALE-315): an interim analysis of a randomised clinical trial. Lancet Respir Med. 2025;13(1):43–55. doi:10.1016/S2213-2600(24)00310-8

15. Zhang W, Dai T, Wang D, Zhu Y, Hua W. Efficacy of neoadjuvant PD-1/PD-L1 inhibitor in resectable NSCLC: a meta-analysis based on randomized controlled trials. BMC Cancer. 2024;24(1):1522. doi:10.1186/s12885-024-13311-5

16. Sterne JAC, Savovic J, Page MJ, et al. RoB 2: a revised tool for assessing risk of bias in randomised trials. BMJ. 2019;366:l4898. doi:10.1136/bmj.l4898

17. Viechtbauer W. Conducting meta-analyses in R with the metafor package. J Stat Softw. 2010;36(3):1–48. doi:10.18637/jss.v036.i03

18. Langan D, Higgins JPT, Jackson D, et al. A comparison of heterogeneity variance estimators in simulated random-effects meta-analyses. Res Synth Methods. 2019;10(1):83–98. doi:10.1002/jrsm.1316

19. von Hippel PT. The heterogeneity statistic I² can be biased in small meta-analyses. BMC Med Res Methodol. 2015;15:35. doi:10.1186/s12874-015-0024-z

20. Provencio M, Nadal E, González-Larriba JL, et al. Perioperative nivolumab and chemotherapy in stage III non-small-cell lung cancer. N Engl J Med. 2023;389(6):504–513. doi:10.1056/NEJMoa2215530

21. Hamel C, Kelly SE, Thavorn K, Rice DB, Wells GA, Hutton B. An evaluation of DistillerSR's machine learning-based prioritization tool for title/abstract screening — impact on reviewer-relevant outcomes. BMC Med Res Methodol. 2020;20(1):256. doi:10.1186/s12874-020-01129-1

Supplementary Material

Supplementary Table S1. Study-Level Extracted Data

S1a. Event-Free Survival (HR, 95% CI)

Study	HR	95% CI	In Zhang?
Neotorch	0.40	0.28–0.57	Yes
Provencio/NADIM II	0.47	0.25–0.88	No
TD-FOREKNOW	0.52	0.21–1.29	Yes
KEYNOTE-671	0.58	0.46–0.72	Yes
CheckMate 77T	0.59	0.44–0.79	Yes
CheckMate 816	0.63	0.44–0.89	Yes
AEGEAN	0.68	0.53–0.88	Yes
Pooled (REML)	0.57	0.50–0.66	k=7

S1b. Overall Survival (HR, 95% CI)

Study	HR	95% CI	Source
Provencio/NADIM II	0.43	0.19–0.98	NEJM 2023
Neotorch	0.62	0.38–1.00	JAMA 2024
CheckMate 816	0.72	0.52–1.00	NEJM 2022
Pooled (REML)	0.66	0.51–0.85	k=3

S1c. Pathological Complete Response (events/N per arm)

Study	Events ICI+Chemo	N ICI+Chemo	Events Chemo	N Chemo	Study RR	In Zhang?
CheckMate 816	43	179	4	179	10.75	Yes
KEYNOTE-671	72	397	16	400	4.53	Yes
Provencio/NADIM II	21	57	2	29	5.34	No
CheckMate 77T	58	229	11	232	5.34	Yes
AEGEAN	63	366	11	374	5.85	Yes
Neotorch	49	202	6	202	8.17	Yes
Pooled (REML)	—	—	—	—	RR 5.81 (4.14–8.17)	k=6

S1d. Major Pathological Response (events/N per arm)

Study	Events ICI+Chemo	N ICI+Chemo	Events Chemo	N Chemo	Study RR	In Zhang?
CheckMate 816	66	179	16	179	4.13	Yes
KEYNOTE-671	120	397	44	400	2.75	Yes
AEGEAN	122	366	46	374	2.71	Yes
Provencio/NADIM II	5	39	2	43	2.75	No
CheckMate 77T	82	229	16	232	5.19	Yes
Pooled (REML)	—	—	—	—	RR 3.06 (2.53–3.72)	k=5

Note: Study-level values are those extracted and validated in the Axelium reproduction. Minor discrepancies with Zhang et al. study-level values may exist due to different data sources or extraction approaches. Event counts are derived from reported percentages and sample sizes where exact counts were not provided.

Supplementary Table S2. Outcome Coverage Matrix (Axelium vs. Zhang)

Trial	EFS (A)	EFS (Z)	OS (A)	OS (Z)	pCR (A)	pCR (Z)	MPR (A)	MPR (Z)
KEYNOTE-671	☐	☐	—	—	☐	—	☐	☐
CheckMate 816	☐	☐	☐[a]	☐[b]	☐	☐	☐	—
CheckMate 77T	☐	☐	—	—	☐	☐	☐	☐
AEGEAN	☐	☐	—	—	—	☐	☐	☐
Neotorch	☐	☐	☐	☐	—	☐	—	—
TD-FOREKNOW	☐	☐	—	—	—	☐	—	☐
NADIM II	☐	—	☐	☐	☐	☐	☐	☐
RATIONALE-315	—	—	—	—	—	—	—	—
Total k	7	6	3	3	6[c]	6	5[c]	5

A = Axelium; Z = Zhang et al. ☐ = included; — = not included.

[a] *Axelium: CheckMate 816 OS HR = 0.72 from 2022 NEJM.*

[b] *Zhang: CheckMate 816 OS HR = 0.57, likely from an updated follow-up analysis.*

[c] *Axelium pCR and MPR study composition differs from Zhang (e.g., Axelium includes KEYNOTE-671 pCR, Zhang includes AEGEAN pCR); total k is the same but contributing studies differ partially.*

Supplementary Table S3. Publication Bias Assessment

Endpoint	Egger's test p	Begg's test p	Trim-and-fill imputed studies	Adjusted estimate
EFS	>0.05	>0.05	0	Unchanged
OS	>0.05	>0.05	0	Unchanged
pCR	>0.05	>0.05	0	Unchanged
MPR	>0.05	>0.05	0	Unchanged

Supplementary Table S4. Leave-One-Out Sensitivity Analysis

S4a. EFS — Leave-One-Out

Study removed	Pooled HR	95% CI	I^2
KEYNOTE-671	0.55	0.46–0.66	17%
CheckMate 816	0.55	0.47–0.65	15%
Provencio/NADIM II	0.58	0.50–0.67	16%
CheckMate 77T	0.56	0.48–0.66	17%
AEGEAN	0.54	0.47–0.63	8%
Neotorch	0.60	0.54–0.67	0%
TD-FOREKNOW	0.57	0.49–0.66	18%

Note: No single study removal changed the direction or statistical significance of the pooled estimate.

Supplementary Table S5. Methodological Differences Between Original and Reproduced Analyses

Feature	Zhang et al. (2024)	Axelium Reproduction
Statistical software	Stata 12.0, RevMan 5.3	metafor R package (v4.6)
Random-effects estimator	DerSimonian–Laird (RevMan default)	REML
Publication bias test	Begg's test only	Egger's + Begg's + trim-and-fill
GRADE assessment	Not performed	Performed for all 4 endpoints
Sensitivity analyses	Leave-one-out	LOO + fixed/random + low-RoB + phase-3 only
Subgroup analyses	EFS by 8 clinical subgroups	By treatment strategy, drug target
Meta-regression	Not performed	pCR and EFS moderators
EFS studies included	6 (NADIM II excluded from EFS)	7 (NADIM II included in EFS)
CheckMate 816 OS HR	0.57 (likely from updated follow-up)	0.72 (from primary 2022 NEJM publication)
Additional trial	—	RATIONALE-315 (published after Zhang search period)
PROSPERO registration	CRD42024544761	N/A (reproduction study)
Data extraction	Manual (2 independent reviewers)	AI-assisted with human validation
Analysis interaction	Command-line (Stata) + GUI (RevMan)	Conversational AI Stats Agent (web UI)
Search databases	PubMed, Embase, Cochrane, WoS	PubMed (targeted retrieval of known trials)

Supplementary File 2: Full Axelium-Generated Report

The complete narrative report automatically generated by the Axelium platform from the 50 pinned evidence items is provided as a separate supplementary file (`supplementary-report.md`). This report was generated on 2026-04-13 and includes the platform's own Introduction, Methods, Results (with all pooled estimates, sensitivity analyses, publication bias assessments, and GRADE evaluations), and Discussion sections.

PRISMA 2020 Checklist

A completed PRISMA 2020 checklist is available as Supplementary File 1.

Section	Item	Reported on page
Title		
Title	1	Identify the report as a systematic review
Abstract		
Abstract	2	Structured summary
Introduction		
Rationale	3	Describe the rationale
Objectives	4	Provide explicit statement of objectives
Methods		
Eligibility criteria	5	Specify inclusion and exclusion criteria
Information sources	6	Describe all information sources
Search strategy	7	Present full search strategy
Selection process	8	Specify methods for selection
Data collection process	9	Describe methods for data extraction
Data items	10a	List and define outcome data
Study risk of bias	11	Methods for assessing risk of bias
Effect measures	12	Specify effect measures
Synthesis methods	13a–f	Describe synthesis methods
Reporting bias assessment	14	Methods for assessing reporting bias
Certainty assessment	15	Methods for assessing certainty
Results		
Study selection	16a–b	Results of selection process
Study characteristics	17	Characteristics of included studies
Risk of bias in studies	18	Risk of bias assessments
Results of syntheses	20a–d	Results of statistical syntheses
Reporting biases	21	Results of reporting bias assessment
Certainty of evidence	22	Certainty of evidence
Discussion		
Discussion	23a–d	General interpretation
Other		

Section	Item	Reported on page
Registration/protocol	24a–c	Registration information
Support	25	Sources of support
Competing interests	26	Competing interests
Availability of data	27	Availability of data

Manuscript prepared for medRxiv preprint submission. Last updated: April 2026.