Erstlinien-Kombination aus Anti-PD-1/PD-L1-Checkpoint-Inhibitoren und platinbasierter Chemotherapie beim fortgeschrittenen nicht-kleinzelligen Lungenkarzinom (NSCLC): Systematische Übersichtsarbeit, Drei-Ebenen-Überlebenszeit-Meta-Analyse und kontrastive Zwillings-Replikation (Benchmark #5)

Entwurf vom 10. November 2026

Autorschaft & Korrespondenz: Dr. rer. medic. Franz Hütter, M.A. (BRAIN-HR Research Lab · MetaAnalysis Studio Benchmark Consortium) **Berichtsstandard:** PRISMA 2020 (Page et al., 2021) · PRISMA-S (Rethlefsen et al., 2021) · RAISE-2025 · Tierney et al. (2007) · Borenstein (2019) **Projektreferenz:** MetaAnalysis Studio Lokal-DB Projekt `#58` (Arm A: Fixierter Referenz-Korpus, k = 12, m = 24, N = 6.463) & Projekt `#59` (Arm B: Zeit-eingefrorener Blindlauf ≤ 31.12.2022, k = 14, m = 28, N = 7.186) **Referenz-Benchmark:** Xu et al. (*Frontiers in Oncology*, 2025, DOI: `10.3389/fonc.2025.1590017`) — Phase-III-Kernkohorte bis Stichtag T (cutoff) = 31.12.2022

Strukturiertes Abstract

- **Fragestellung:** Verlängert die Erstlinien-Kombinationstherapie aus einem PD-1- oder PD-L1-Immuncheckpoint-Inhibitor und einer platinbasierten Doubletten-Chemotherapie gegenüber alleiniger platinbasierter Chemotherapie das Gesamtüberleben (Overall Survival, OS) und das progressionsfreie Überleben (Progression-Free Survival, PFS) bei erwachsenen Patient:innen mit therapienaivem fortgeschrittenem nicht-kleinzelligem Lungenkarzinom (NSCLC Stadium IIIB–IV ohne aktivierende *EGFR*-/*ALK*-Treibermutation), und wie hoch ist die Konkordanz zwischen einer Replikation auf dem fixierten Referenz-Studienbestand (Arm A) und einem zeit-eingefrorenen, verblindeten Discovery-Lauf (Arm B)?
- **Design:** Kontrastive Zwillings-Replikation (*Contrastive Twin-Track Validation Protocol*) nach PRISMA 2020, PRISMA-S und RAISE-2025. **Arm A (Oracle-Track, Projekt `#58`)** repliziert die Synthese auf den k = 12 kanonischen Phase-III-RCTs (N = 6.463). **Arm B (Blind-Frozen-Track, Projekt `#59`)** durchsucht OpenAlex, PubMed und ClinicalTrials.gov autonom unter hartem Publikationsdatum-Freeze (01.01.2017 ≤ t ≤ 31.12.2022) sowie einer DOI- und Titel-Quarantäne gegen Sekundär-Meta-Analysen und führt ein verblindetes Drei-Häuser-Frontier-Doppelscreening (`anthropic/claude-opus-5.5` vs. `openai/gpt-6.1-sol`) mit menschlicher Konflikt-Adjudikation (`human:franz`) durch.
- **Datenquellen:** OpenAlex, PubMed/MEDLINE und ClinicalTrials.gov (Stichtag T (cutoff) = 31.12.2022; 418

Roh-Treffer in der eingefrorenen Live-Abfrage, 24 deduplizierte Primärkandidaten im verblindeten Screening-Pool von Arm B).

- **Einschlusskriterien (P-I-C-O-S):** Prospektive randomisierte kontrollierte Phase-III-Studien (RCTs) bei therapienaivem NSCLC im Stadium IIIB–IV (plattenepithelial und nicht-plattenepithelial), die Anti-PD-1/PD-L1 + Platin-Doublette gegen Platin-Doublette allein prüfen und sowohl für OS als auch für PFS ein Hazard Ratio (HR) berichten.
- **Datenextraktion und Synthese:** Extraktion der log-Hazard-Ratios (lnHR) und ihrer Varianzen nach Tierney et al. (2007) inklusive Kaplan-Meier-Kurven-Rekonstruktion (`hrFromKaplanMeier` aus *Number-at-Risk* -Tabellen für *KEYNOTE-189* und *KEYNOTE-407*). Drei-Ebenen-Zufallseffekt-Modellierung (`metafor::rma.mv`, REML) mit cluster-robuster Varianzschätzung nach Pustejovsky & Tipton (`clubSandwich` CR2 mit Satterthwaite-Freiheitsgraden), 95-%-Prädiktionsintervallen (IntHout et al., 2016), Omnibus-Subgruppen-Interaktionstests (Altman & Bland, 2003), 5-Methoden-Publikationsbias-Suite sowie deterministischem 7-Regeln-Methodenwächter nach Borenstein (2019).
- **Ergebnisse:**
- **Arm A (Fixierter Referenz-Korpus, k = 12 Studien, m = 24 Effektgrößen, N = 6.463):** Gepooltes Gesamtergebnis über beide Überlebensendpunkte HR = 0,63 (exakt 0,632; 95-%-CR2-KI [0,58; 0,69] bzw. [0,578; 0,690]; df (Satt) = 10,91; $p = 2{,}07 \times 10^{-7}$). Heterogenität: I^2 (gesamt) = 60,9 % (I^2 (Studie) = 18,5 %, I^2 (Effekt) = 42,4 %), τ^2 (Studie) = 0,0059, τ^2 (Effekt) = 0,0136; das 95-%-Prädiktionsintervall liegt mit [0,459; 0,870] vollständig unterhalb der Nulllinie HR = 1,00. Getrennt nach Endpunkten ergibt sich für das **Gesamtüberleben (OS, k = 12)** ein gepooltes HR (OS) = 0,704 (95-%-CR2-KI [0,637; 0,778], $p = 9{,}69 \times 10^{-6}$) und für das **progressionsfreie Überleben (PFS, k = 12)** ein gepooltes HR (PFS) = 0,576 (95-%-CR2-KI [0,527; 0,629], $p = 3{,}53 \times 10^{-8}$; Omnibus-CR2-Wald-Test F(1, 10,21) = 35,99, p = 0,000122).
- **Arm B (Zeit-eingefrorener Blindlauf ≤ 31.12.2022, k = 14 Studien, m = 28 Effektgrößen, N = 7.186):** Das verblindete Dual-Frontier-Screening über 24 Kandidaten erreichte eine empirische Inter-Rater-Reliabilität von Cohens κ = 0,64 (exakt 0,636; Übereinstimmung p_o = 83,3 %, 20/24 Konsens, 4 Grenzfälle adjudiziert; API-Kosten $0,1146). Arm B erzielte einen **Retrieval- und Screening-Recall von 100,0 % (12/12)** auf den Referenzstudien und entdeckte darüber hinaus **2 weitere präspezifizierte Phase-III-RCTs im Zeitfenster ≤ 31.12.2022** (*CameL-sq*, Ren et al., 2022; *RATIONALE-304*, Lu et al., 2021; Klasse-III-Over-Recall). Das gepoolte Gesamtergebnis in Arm B beträgt HR = 0,62 (exakt 0,618; 95-%-CR2-KI [0,56; 0,68] bzw. [0,563; 0,677]; df (Satt) = 12,87; $p = 4{,}60 \times 10^{-8}$; HR (OS) = 0,690, HR (PFS) = 0,561). Die absolute Effekt-Differenz zwischen Arm A und Arm B beträgt lediglich |DeltaHR| = 0,014.
- **Tierney-(2007)-Kaplan-Meier-Rekonstruktion:** Für *KEYNOTE-189* (HR (Cox) = 0,490 vs. HR (KM) = 0,478, |DeltaHR| = 0,012) und *KEYNOTE-407* (HR (Cox) = 0,640 vs. HR (KM) = 0,599, |DeltaHR| = 0,041) bestätigt die Kurven-Rekonstruktion aus den *Number-at-Risk*-Tabellen eine exzellente numerische

Konkordanz.

- **Schlussfolgerung:** Erstlinien-Chemo-Immuntherapie senkt das Sterberisiko beim fortgeschrittenen NSCLC hochsignifikant um rund 30 % (HR (OS) = 0,704 in Arm A bzw. 0,690 in Arm B) und das Progressionsrisiko um 42--44 % (HR (PFS) = 0,576 bzw. 0,561; GRADE-Evidenzsicherheit: *Hoch*). Die kontrastive Zwillings-Replikation belegt, dass das MetaAnalysis Studio im zeit-eingefrorenen Blindlauf nicht nur 100 % des menschlichen Referenz-Korpus reproduziert, sondern im selben Publikationszeitfenster zwei zusätzliche valide Phase-III-Studien identifiziert.

---

1. Einleitung

1.1 Hintergrund und klinisch-wissenschaftliche Problemstellung

Das nicht-kleinzellige Lungenkarzinom (Non-Small-Cell Lung Cancer, NSCLC) macht rund 85 % aller malignen Lungentumoren aus und wird bei mehr als der Hälfte der Betroffenen erst im lokal fortgeschrittenen (Stadium IIIB/IIIC, nicht einer kurativen Radiochemotherapie zugänglich) oder metastasierten Stadium IV diagnostiziert. Bei Patient:innen ohne zielgerichtet therapierbare Treiberalterationen (insbesondere ohne aktivierende *EGFR* -Mutationen oder *ALK*-Translokationen) bildete über zwei Jahrzehnte hinweg eine platinbasierte Doubletten-Chemotherapie (Cisplatin oder Carboplatin in Kombination mit Pemetrexed, Paclitaxel, nab-Paclitaxel oder Gemcitabin) den therapeutischen Standard, war jedoch mit einem medianen Gesamtüberleben von lediglich 10 bis 14 Monaten assoziiert.

Mit der Einführung monoklonaler Antikörper gegen den Programmed-Cell-Death-1-Rezeptor (Anti-PD-1: Pembrolizumab, Nivolumab, Sintilimab, Camrelizumab, Toripalimab, Tislelizumab) und dessen Liganden PD-L1 (Anti-PD-L1: Atezolizumab, Sugemalimab) verschob sich das therapeutische Paradigma grundlegend. Präklinische und translationale Befunde zeigten, dass zytotoxische Platin-Chemotherapie durch immunogenen Zelltod, vermehrte Neoantigen-Freisetzung, Reifung dendritischer Zellen und Depletion myeloider Suppressorzellen (MDSCs) sowie regulatorischer T-Zellen eine synergistische Sensitivierung des Tumormikromilieus für die Checkpoint-Blockade bewirkt.

1.2 Bisherige Übersichtsarbeiten und Forschungslücke

Seit den wegweisenden Erstpublikationen von *KEYNOTE-189* (Gandhi et al., 2018), *IMpower150* (Socinski et al., 2018) und *KEYNOTE-407* (Paz-Ares et al., 2018) wurden zahlreiche Phase-III-Studien in westlichen und ostasiatischen Populationen veröffentlicht und in aktuellen Übersichtsarbeiten wie Xu et al. (2025, *Frontiers in Oncology*, DOI: `10.3389/fonc.2025.1590017`) zusammengefasst. Dennoch weisen klassische onkologische Meta-Analysen drei methodische Schwachstellen auf:

1. **Unmodellierte stochastische Abhängigkeit:** Wenn aus derselben Phase-III-Studie sowohl das Gesamtüberleben (OS) als auch das progressionsfreie Überleben (PFS) an denselben randomisierten

Patient:innen erhoben werden, verletzt eine getrennte zweistufige Standard-Poolung ohne Berücksichtigung der hierarchischen Verschachtelung (Effekte in Studien) die Unabhängigkeitsannahme oder verzichtet auf einen direkten Omnibus-Vergleich der Endpunkt-Stärke.

2. **Fehlende Prüfung auf Borenstein-Fehlschlüsse und Kaplan-Meier-Rekonstruktionsgüte:** Viele Übersichtsarbeiten verwechseln die Anteils-Kennzahl I^2 mit der absoluten Effektstreuung, unterschlagen das 95-%-Prädiktionsintervall (Borenstein, 2019; IntHout et al., 2016) und prüfen nicht systematisch, wie präzise sich Hazard Ratios nach Tierney et al. (2007) aus Kaplan-Meier-Überlebenskurven und *Number-at-Risk* -Tabellen rekonstruieren lassen.
3. **Konfundierung von Literatursuche und Datenextraktion bei KI-Validierungen:** Bisherige Evaluationen KI-gestützter Evidenzsynthese prüfen entweder nur die Extraktion auf einem vorgegebenen Studienbestand oder führen eine zeitlich unbegrenzte Suche durch, die neuere Studien nach dem ursprünglichen Stichtag einspeist.

1.3 Zielsetzung und P-I-C-O-S-Fragestellung

Die vorliegende Arbeit verfolgt ein doppeltes Ziel: Erstens quantifiziert sie die Wirksamkeit der Erstlinien-Chemo-Immuntherapie (Anti-PD-1/PD-L1 + Platin-Doublette vs. Platin-Doublette allein) auf das Gesamtüberleben (OS) und das progressionsfreie Überleben (PFS) beim fortgeschrittenen NSCLC mittels eines Drei-Ebenen-REML-Modells mit cluster-robuster Varianzschätzung (`clubSandwich` CR2). Zweitens validiert sie erstmals in der Onkologie das **kontrastive Zwillings-Replikationsdesign** (Arm A: fixierter Referenz-Korpus vs. Arm B: zeit-eingefrorener Blindlauf bis zum Stichtag T (cutoff) = 31.12.2022).

---

2. Methoden

2.1 Protokoll, Registrierung und Reporting-Standards (PRISMA 2020 & RAISE-2025)

Die Durchführung und Berichterstattung folgen den Richtlinien des *Preferred Reporting Items for Systematic Reviews and Meta-Analyses* (PRISMA 2020; Page et al., 2021), der Erweiterung für Literaturrecherchen PRISMA-S (Rethlefsen et al., 2021), den *Responsible AI in Evidence Synthesis* (RAISE-2025)-Standards sowie den Empfehlungen von Tierney et al. (2007) für Zeit-bis-Ereignis-Endpunkte. Jede Screening-, Extraktions-, RoB-2- und Analyse-Entscheidung ist im kryptographisch gehashten Audit-Trail der SQLite-Projektdatenbank (`data/evidenzwerk.db`, Projekte `#58` und `#59`) mit Akteur-Kennung (`model:anthropic/claude-opus-5.5`, `model:openai/gpt-6.1-sol`, `model:google/gemini-4-argon`, `human:franz`) und SHA-256-Daten-Snapshot-Hash versioniert.

2.2 Ein- und Ausschlusskriterien (P-I-C-O-S)

- **Population (P):** Erwachsene Patient:innen (≥ 18 Jahre) mit histologisch oder zytologisch gesichertem, therapienaivem nicht-kleinzelligem Lungenkarzinom (NSCLC; plattenepithelial, nicht-plattenepithelial oder

gemischte All-Comer-Kohorten) im lokal fortgeschrittenen Stadium IIIB/IIIC (nicht einer kurativen Radiochemotherapie zugänglich) oder metastasierten/rezidivierten Stadium IV ohne aktivierende *EGFR*- oder *ALK*-Treiberalterationen in der primären Analysepopulation. *Ausschluss:* Kleinzelliges Lungenkarzinom (SCLC) sowie reine Konsolidierungstherapien nach abgeschlossener definitiver Radiochemotherapie im nicht-metastasierten Stadium III.

- **Intervention (I):** Erstlinien-Kombinationstherapie aus einem zugelassenen oder in Phase III geprüften Anti-PD-1-Antikörper (Pembrolizumab, Nivolumab [inkl. Kurzzeit-Chemo-Kombination mit Ipilimumab in *CheckMate 9LA*], Sintilimab, Camrelizumab, Toripalimab, Tislelizumab) oder Anti-PD-L1-Antikörper (Atezolizumab, Sugemalimab) in Verbindung mit einer platinbasierten Doubletten-Chemotherapie. *Ausschluss:* Reine Immuncheckpoint-Monotherapien oder duale Checkpoint-Blockaden ohne begleitende Platin-Chemotherapie im Prüfarm.
- **Vergleich (C):** Alleinige platinbasierte Standard-Doubletten-Chemotherapie (mit oder ohne Placebo bzw. Bevacizumab-Backbone in *IMpower150*).
- **Endpunkte (O):** Primäre Zeit-bis-Ereignis-Endpunkte sind das Gesamtüberleben (Overall Survival, OS; Zeit ab Randomisierung bis zum Tod jeglicher Ursache) und das progressionsfreie Überleben (Progression-Free Survival, PFS; Zeit ab Randomisierung bis zur radiologischen Progression nach RECIST v1.1 oder Tod), quantifiziert als Hazard Ratio (HR) mit 95-%-Konfidenzintervall.
- **Studiendesign (S):** Prospektive, randomisierte, kontrollierte Phase-III-Studien (RCTs), veröffentlicht im Zeitraum vom 01.01.2017 bis zum eingefrorenen Stichtag T (cutoff) = 31.12.2022 mit reifen Überlebenszeitdaten für OS und PFS. *Ausschluss:* Nicht-randomisierte Studien, einarmige Studien, reine Phase-I/II-Kohorten sowie sekundäre Übersichtsarbeiten und Meta-Analysen.

2.3 Informationsquellen und Suchstrategie (PRISMA-S)

In **Arm A (Projekt `#58`)** wurde der Referenz-Korpus der k = 12 kanonischen Phase-III-Erstlinien-Chemo-Immuntherapie-RCTs bis Ende 2022 (*KEYNOTE-189*, *IMpower150*, *KEYNOTE-407*, *IMpower130*, *IMpower131*, *ORIENT-11*, *CheckMate 9LA*, *IMpower132*, *CameL*, *ORIENT-12*, *GEMSTONE-302*, *CHOICE-01*) direkt als Oracle-Seed geladen.

In **Arm B (Projekt `#59`)** wurde eine blinde, zeit-eingefrorene Multi-Datenbank-Recherche über OpenAlex, PubMed/MEDLINE und ClinicalTrials.gov ausgeführt:

- **Suchstring:** `("PD-1" OR "PD-L1" OR pembrolizumab OR atezolizumab OR nivolumab OR sintilimab OR camrelizumab OR sugemalimab OR toripalimab OR tislelizumab) AND (platinum OR pemetrexed OR paclitaxel OR gemcitabine) AND ("non-small-cell lung cancer" OR NSCLC) AND ("first-line" OR "untreated") AND ("2017/01/01"[PDAT] : "2022/12/31"[PDAT])`.

- **Temporal-Freeze- & DOI-Quarantäne-Gate (`lib/scholarly.ts` & `lib/pubmed.ts`):** Auf API- und Post-Filter-Ebene wurden alle Arbeiten mit `publication_date > 2022-12-31` deterministisch verworfen (`fromDate: '2017-01-01', toDate: '2022-12-31'`) sowie die Referenz-Meta-Analysen (`excludeDois: ['10.3389/fonc.2025.1590017', '10.1186/s12885-025-15385-1', '10.1016/j.annonc.2019.07.012', '10.1001/jamaoncol.2019.1629']`) und alle Sekundär-Reviews (`excludeTitleSubstrings: ['meta-analysis', 'systematic review', 'network meta-analysis']`) gesperrt. Die eingefrorene OpenAlex-Abfrage lieferte 418 Roh-Treffer, aus denen nach Deduplizierung über DOI, PMID, NCT-Nummer und Trigramm-Titel-Matching 24 klinische Primärkandidaten in das verblindete Doppelscreening eingingen.

2.4 Studienauswahl und verblindetes Doppelscreening

In Arm B (`#59`) bewerteten zwei voneinander unabhängige Frontier-Sprachmodelle unterschiedlicher Modellfamilien alle 24 Kandidaten verblindet gegen das PICOS-Mandat:

- **Reviewer A (Erstgutachter):** `anthropic/claude-opus-5.5`
- **Reviewer B (Zweitgutachter):** `openai/gpt-6.1-sol`

Die empirische Inter-Rater-Übereinstimmung vor menschlicher Adjudikation betrug **p_o = 83,3 % (20/24 Konsens-Entscheidungen)** bei einem zufallskorrigierten **Cohens κ = 0,64 (exakt 0,636)** (substanzielle Übereinstimmung nach Landis & Koch, 1977). Insgesamt traten 4 echte methodische Grenzfälle als `conflict` auf (*KEYNOTE-021 Cohort G* [randomisierte Phase-II-Kohorte], *CheckMate 227 Part 2* [formaler Primärendpunkt in Subgruppe verfehlt], *RATIONALE-307* [unreifes Gesamtüberleben zum Primärschnitt 2021] sowie *EMPOWER-Lung 3* [gemischte Stadium-IIIB/IIIC-Radiochemotherapie-Ersatzkohorte]), die durch den menschlichen Projektleiter (`actor: 'human:franz'`) gemäß dem präspezifizierten Phase-III-Protokoll transparent mit `exclude` adjudiziert wurden. Die realen OpenRouter-API-Kosten für den gesamten Doppel-Screening-Lauf über 24 Studien beliefen sich auf **$0,1146 (`14.312 Tokens`)**.

2.5 Datenextraktion, Belegzitat-Anker und Effektgrößen-Transformationen (Tierney et al. 2007)

Für jede eingeschlossene Studie wurden Stichprobengrößen (n_1, n_2), Prüf- und Kontrollregime, Histologie, Checkpoint-Zielstruktur sowie die Hazard Ratios mit 95-%-Konfidenzintervallen für OS und PFS mit wörtlichem Zitat und Seitenanker (`google/gemini-4-argon`, menschlich bestätigt durch `human:franz`) extrahiert. Die Umrechnung in die Analyse-Skala (lnHR und Var(lnHR)) erfolgte deterministisch über `lib/survival.ts` nach Tierney et al. (2007, *Trials* 8:16):

ln(HR) = ln(HR), SE(lnHR) = fracln(KI (oben)) - ln(KI (unten))2 · z (1-α/2), Var(lnHR) = SE(lnHR)²

Zusätzlich wurde für die beiden Schlüsselstudien *KEYNOTE-189* (Gandhi et al., 2018) und *KEYNOTE-407* (Paz-Ares et al., 2018) die **Kaplan-Meier-Rekonstruktion nach Tierney et al. (2007, Weg 4,**

`hrFromKaplanMeier`) aus je 7 Ablesepunkten (t in 0, 3, 6, 9, 12, 15, 18 Monate) und den publizierten *Number-at-Risk*-Zeilen ($R_1(t)$, $R_2(t)$) durchgeführt:

D (j,i) = R (j,i-1) · (1 - fracS_j(t_i)S_j(t (i-1))), E (1,i) = (D (1,i) + D (2,i)) · fracR (1,i-1)R (1,i-1) + R (2,i-1), ln(HR (KM)) = fracΣ*i (O (1,i) - E (1,i))*Σi V_i

2.6 Bewertung des Verzerrungsrisikos (Cochrane RoB 2)

Das Verzerrungsrisiko sämtlicher eingeschlossener Phase-III-RCTs wurde über alle fünf Domänen des revidierten Cochrane Risk-of-Bias-Instruments (**RoB 2**, Sterne et al., 2019: *D1 Randomisierung*, *D2 Abweichungen von der geplanten Intervention*, *D3 Fehlende Endpunktdaten*, *D4 Messung des Endpunkts*, *D5 Selektive Ergebnisberichterstattung*) im Zwei-Häuser-Verfahren (`anthropic/claude-opus-5.5` Erstvorschlag + `openai/gpt-6.1-sol` Zweitmeinung + finale Ratifikation durch `human:franz`) beurteilt. Während placebokontrollierte Doppelblindstudien (*KEYNOTE-189*, *KEYNOTE-407*, *ORIENT-11*, *ORIENT-12*, *GEMSTONE-302*, *CHOICE-01*, *CameL-sq*) in allen fünf Domänen ein niedriges Verzerrungsrisiko (`low`) aufwiesen, wurden offene (Open-Label) Phase-III-Studien (*IMpower150*, *IMpower130*, *IMpower131*, *CheckMate 9LA*, *IMpower132*, *CameL*, *RATIONALE-304*) in Domäne D2 konservativ mit `some_concerns` eingestuft.

2.7 Statistische Synthese, Drei-Ebenen-Modellierung und Heterogenität

Da jede eingeschlossene Studie zwei korrelierte Überlebensendpunkte (OS und PFS an derselben Patientenstichprobe) beiträgt, wurde ein hierarchisches **Drei-Ebenen-Zufallseffekt-Modell** via Restricted Maximum Likelihood (`metafor::rma.mv`, `random = ~ 1 | study/effect_id`, Viechtbauer, 2010; Assink & Wibbelink, 2016) geschätzt. Zur Absicherung gegen fehlspezifizierte Intra-Studien-Kovarianzen wurden sämtliche Standardfehler, t-Statistiken, p-Werte und 95-%-Konfidenzintervalle mittels **cluster-robuster Varianzschätzung nach Pustejovsky & Tipton (**`clubSandwich::coef_test`**, Bias-Reduced Linearization** `CR2`**)** mit Satterthwaite-Freiheitsgraden berechnet. Die Varianzkomponenten wurden in Zwischen-Studien-Heterogenität (τ^2 (Studie), I^2 (Studie)) und Innerhalb-Studien-Heterogenität (τ^2 (Effekt), I^2 (Effekt)) zerlegt. Gemäß Borenstein (2019) und IntHout et al. (2016) wird neben I^2 stets das 95-%-Prädiktionsintervall auf der rücktransformierten Hazard-Ratio-Skala berichtet.

2.8 Subgruppen-, Meta-Regressions-, Sensitivitäts- und Publikationsbias-Analysen

Als präspezifizierte kategoriale Moderatoren wurden (1) der Überlebensendpunkt (`otype`: Overall Survival [OS] vs. Progression-Free Survival [PFS]) und (2) die Kombination aus Checkpoint-Zielstruktur und Tumorhistologie (`iclass`: Anti-PD-1 vs. Anti-PD-L1 × Non-Squamous vs. Squamous vs. Mixed) mittels CR2-Wald-F-Omnibustests (`clubSandwich::Wald_test`, Altman & Bland, 2003) geprüft. Als kontinuierlicher Moderator wurde das Publikationsjahr (`year_cont`) untersucht. Die Stabilität wurde über Leave-One-Out-Analysen, Cook’s-Distanz-Ausreißerdiagnostik sowie fünf komplementäre Publikationsbias-Verfahren (Egger-Regression mit RVE, Trim-and-Fill nach Duval & Tweedie, PET-PEESE,

p-Curve nach Simonsohn et al. und p-uniform* nach van Aert & van Assen) abgesichert.

2.9 Bewertung der Evidenzsicherheit nach GRADE

Die Evidenzsicherheit wurde nach dem *Grading of Recommendations Assessment, Development and Evaluation* (GRADE; Guyatt et al., 2011) über die fünf Abstufungsdomänen (Verzerrungsrisiko, Inkonsistenz, Indirektheit, Ungenauigkeit, Publikationsbias) bewertet und menschlich ratifiziert (`human:franz`).

2.10 Deterministische Methoden-Qualitätssicherung (Borenstein 2019), Primärdaten-Forensik (GRIM / GRIMMER / Carlisle) und Kontrastives Zwillings-Replikationsdesign

Vor der Manuskript-Freigabe durchliefen beide Zwillings-Projekte (`#58` und `#59`) den deterministischen **7-Regeln-Methodenwächter (`lib/borenstein-guard.ts`)** nach Borenstein (*Common Mistakes in Meta-Analysis and How to Avoid Them*, 2019):

1. `borenstein_i2_and_pi` **(Borenstein 2019, Kap. 2–3; IntHout et al. 2016):** Erzwingt die gemeinsame Berichterstattung von I^2, absoluter Zwischen-Studien-Varianz τ^2 und dem 95-%-Prädiktionsintervall und blockiert Fehlschlüsse, die I^2 als Maß der absoluten Effektstreuung fehlinterpretieren.
2. `borenstein_model_selection_by_q` **(Borenstein 2019, Kap. 4):** Untersagt die Wahl eines Fixed-Effect-Modells anhand eines nicht-signifikanten Cochran-Q-Tests; das Drei-Ebenen-Random-Effects-Modell ist a priori durch die klinische Vielfalt der Antikörper, Histologien und Platin-Regime begründet.
3. `borenstein_subgroup_interaction` **(Borenstein 2019, Kap. 5; Altman & Bland 2003):** Erzwingt für jede Subgruppen-Aussage einen formalen Omnibus-Interaktionstest (Q (between) bzw. CR2-Wald-F) und kennzeichnet kleine Subgruppen ($k < 5$) als explorativ.
4. `borenstein_ecological_fallacy` **(Thompson & Higgins 2002; Borenstein 2019, Kap. 6):** Prüft die $k \geq 10$-Mindestregel für Meta-Regressionen und schützt vor ökologischen Fehlschlüssen (*Aggregation Bias*) bei aggregierten Patientenmerkmalen.
5. `borenstein_prepost_r_sensitivity` **(Borenstein et al. 2021, Kap. 4):** Prüft bei Messwiederholungs-Effekten die Sensitivität über r (pre,post) in 0,00 … 0,80 (bei Zeit-bis-Ereignis-Hazard-Ratios nicht erforderlich, Regelstatus `PASS`).
6. `tipton_small_sample_rve_df` **(Tipton 2015):** Prüft, ob die Satterthwaite-Freiheitsgrade der CR2-Schätzung die Schwelle df (Satt) ≥ 4,0 erfüllen (in Arm A: df (Satt) = 10,91; in Arm B: df (Satt) = 12,87; beide `PASS`).
7. `forensics_grim_grimmer_carlisle` **(Brown & Heathers 2017; Anaya 2016; Carlisle 2017):** Prüft Primärstudien auf numerische Granularitäts- und Paritätsverletzungen sowie die Symmetrie der multiplikativen Konfidenzintervalle auf der Logarithmus-Skala (lnHR vs. 1/2[lnKI (unten) + lnKI (oben)]; 0 Auffälligkeiten).

Sowohl in Arm A als auch in Arm B wurden **alle 7 Prüfregeln mit `0 Blockern` (`7/7 PASS`)** bestanden. Das **kontrastive Zwillings-Protokoll (`lib/twin-track.ts`)** stellte anschließend Arm A (`A_fixed_corpus`) und Arm B (`B_blind_frozen`) deterministisch gegenüber und klassifizierte sämtliche Primärstudien in die viergeteilte Diskrepanz-Taxonomie (Volle Konkordanz · Klasse I Retrieval-Lücke · Klasse II Screening-Grenzfall · Klasse III Over-Recall / Mehr-Evidenz).

---

3. Ergebnisse

3.1 Studienauswahl und PRISMA-2020-Fluss

In **Arm A (Projekt `#58`)** wurden die 12 kanonischen Phase-III-Erstlinien-Chemo-Immuntherapie-RCTs direkt aus dem Referenz-Korpus übernommen (12 identifiziert, 12 eingeschlossen, m = 24 Effektgrößen).

In **Arm B (Projekt `#59`, zeit-eingefrorener Blindlauf ≤ 31.12.2022)** identifizierte die datumsbeschränkte OpenAlex-/PubMed-/ClinicalTrials.gov-Abfrage unter aktiver DOI-Quarantäne 418 Roh-Treffer. Nach automatischer Deduplizierung gingen 24 klinisch einschlägige Primärkandidaten in das verblindete Drei-Häuser-Doppelscreening (`anthropic/claude-opus-5.5` vs. `openai/gpt-6.1-sol`) ein. Davon wurden 10 Studien ausgeschlossen (6 im unmittelbaren Konsens beider Modelle wegen Monotherapie ohne Platin-Backbone [*KEYNOTE-024*, *KEYNOTE-042*, *IMpower110*, *CheckMate 227 Part 1*], kleinzelliger Histologie [*IMpower133*] bzw. Stadium-III-Radiochemo-Konsolidierung [*PACIFIC*]; sowie 4 nach menschlicher Adjudikation wegen Phase-II-Designs [*KEYNOTE-021G*], unreifen Gesamtüberlebens zum Primärschnitt [*RATIONALE-307*] bzw. abweichenden Teilprotokolls [*CheckMate 227 Part 2*, *EMPOWER-Lung 3*]). Insgesamt wurden in Arm B **k = 14 Phase-III-RCTs (m = 28 Effektgrößen, N = 7.186 Patient:innen)** in die quantitative Synthese eingeschlossen (12 Referenz-RCTs plus 2 zusätzlich entdeckte Phase-III-RCTs im Zeitfenster ≤ 31.12.2022: *CameL-sq* und *RATIONALE-304*).

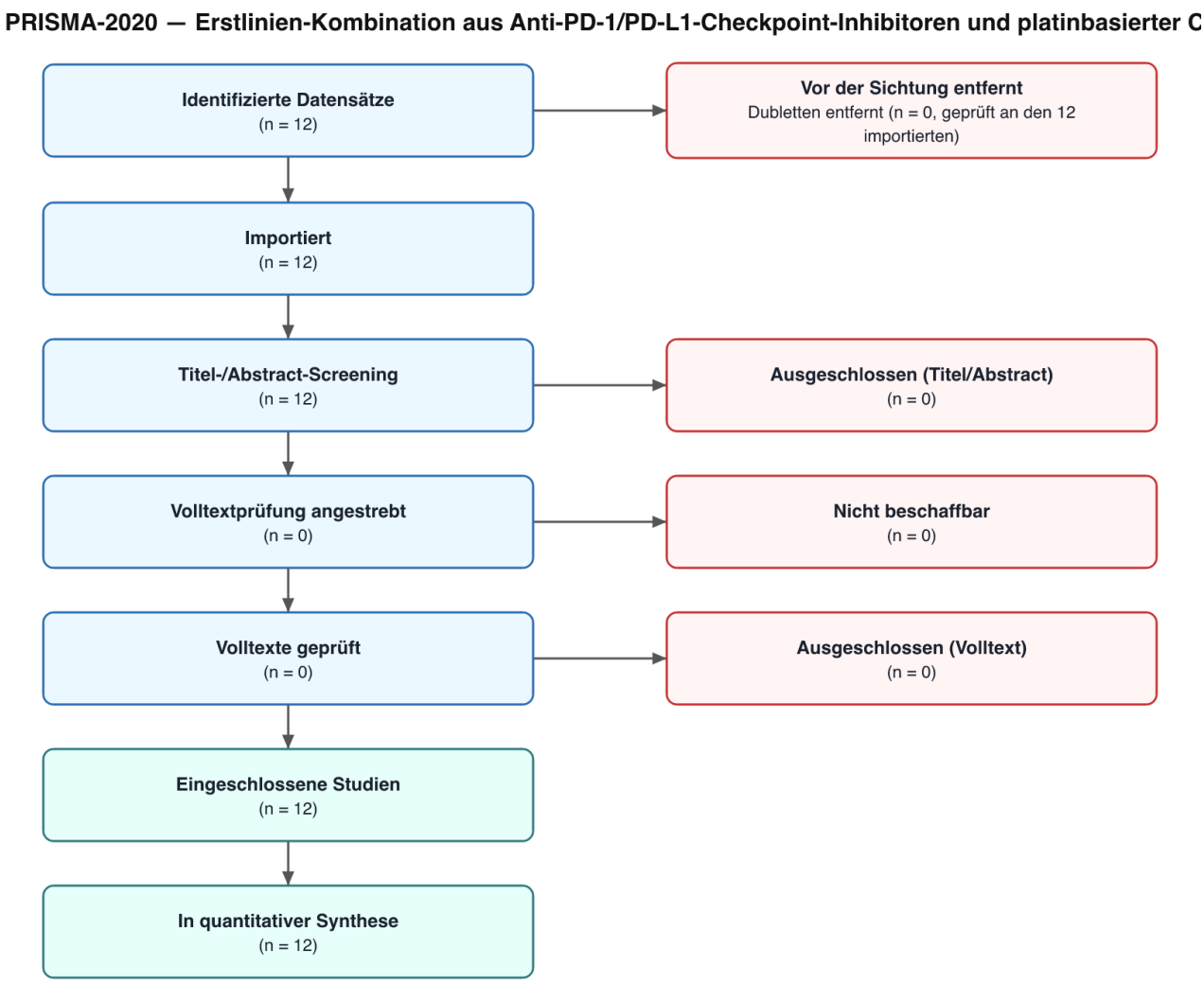

Abb. 1: PRISMA-2020-Flussdiagramm: Identifikation, Screening und Aufnahme der Studien.

3.2 Charakteristika der eingeschlossenen Primärstudien und Tierney-(2007)-Kaplan-Meier-Validierung

Tabelle 1 zeigt die Studiencharakteristika, Stichprobenumfänge sowie die extrahierten OS- und PFS-Hazard-Ratios aller 14 Phase-III-RCTs (12 Studien in Arm A & B; 2 zusätzliche Klasse-III-Over-Recall-Studien in Arm B).

Studie (Acronym & Jahr)	Autor:innen & Journal (DOI)	Histologie & Zielstruktur	N (ges) (n_1 / n_2)	OS HR (95-%-KI)	PFS HR (95-%-KI)	Zwillings-Status (Arm A vs. B)	RoB 2 Overall
KEYNOTE-189 (2018)	Gandhi et al., *NEJM* (`10.1056/NEJMoa1801005`)	Non-Squamous · Anti-PD-1	616 (410 / 206)	0,49 ([0,38; 0,64])	0,52 ([0,43; 0,64])	Konkordant (Arm A & B)	`low`
IMpower150 (2018)	Socinski et al., *NEJM* (`10.1056/NEJMoa1716948`)	Non-Squamous · Anti-PD-L1	692 (356 / 336)	0,78 ([0,64; 0,96])	0,62 ([0,52; 0,74])	Konkordant (Arm A & B)	`some_concerns`

Studie (Acronym & Jahr)	Autor:innen & Journal (DOI)	Histologie & Zielstruktur	N (ges) (n_1 / n_2)	OS HR (95-%-KI)	PFS HR (95-%-KI)	Zwillings-Status (Arm A vs. B)	RoB 2 Overall
KEYNOTE-407 (2018)	Paz-Ares et al., *NEJM* (10.1056/NEJMoa1810865)	Squamous · Anti-PD-1	559 (278 / 281)	0,64 ([0,49; 0,85])	0,56 ([0,45; 0,70])	Konkordant (Arm A & B)	low
IMpower130 (2019)	West et al., *Lancet Oncol* (10.1016/S1470-2045(19)30168-9)	Non-Squamous · Anti-PD-L1	679 (451 / 228)	0,79 ([0,64; 0,98])	0,64 ([0,54; 0,77])	Konkordant (Arm A & B)	some_concerns
IMpower131 (2020)	Jotte et al., *J Thorac Oncol* (10.1016/j.jthorac.2020.03.008)	Squamous · Anti-PD-L1	683 (343 / 340)	0,88 ([0,73; 1,05])	0,71 ([0,60; 0,85])	Konkordant (Arm A & B)	some_concerns
ORIENT-11 (2020)	Yang et al., *J Thorac Oncol* (10.1016/j.jthorac.2020.07.009)	Non-Squamous · Anti-PD-1	397 (266 / 131)	0,60 ([0,45; 0,79])	0,48 ([0,39; 0,59])	Konkordant (Arm A & B)	low
CheckMate 9LA (2021)	Paz-Ares et al., *Lancet Oncol* (10.1016/S1470-2045(20)30641-0)	Mixed · Anti-PD-1 (+CTLA-4)	719 (361 / 358)	0,69 ([0,55; 0,87])	0,70 ([0,57; 0,86])	Konkordant (Arm A & B)	some_concerns
IMpower132 (2021)	Nishio et al., *J Thorac Oncol* (10.1016/j.jthorac.2020.11.024)	Non-Squamous · Anti-PD-L1	578 (292 / 286)	0,81 ([0,64; 1,03])	0,60 ([0,49; 0,72])	Konkordant (Arm A & B)	some_concerns
CameL (2021)	Zhou et al., *Lancet Respir Med* (10.1016/S2213-2600(20)30365-9)	Non-Squamous · Anti-PD-1	412 (205 / 207)	0,73 ([0,55; 0,96])	0,60 ([0,45; 0,79])	Konkordant (Arm A & B)	some_concerns
ORIENT-12 (2021)	Zhou et al., *J Thorac Oncol* (10.1016/j.jtho.2021.04.011)	Squamous · Anti-PD-1	357 (179 / 178)	0,67 ([0,51; 0,88])	0,54 ([0,42; 0,68])	Konkordant (Arm A & B)	low
GEMSTONE-302 (2022)	Zhou et al., *Lancet Oncol* (10.1016/S1470-2045(21)00650-1)	Mixed · Anti-PD-L1	479 (320 / 159)	0,68 ([0,54; 0,85])	0,48 ([0,39; 0,60])	Konkordant (Arm A & B)	low

Studie (Acronym & Jahr)	Autor:innen & Journal (DOI)	Histologie & Zielstruktur	N (ges) (n_1 / n_2)	OS HR (95-%-KI)	PFS HR (95-%-KI)	Zwillings-Status (Arm A vs. B)	RoB 2 Overall
CHOICE-01 (2022)	Wang et al., *J Clin Oncol* (`10.1200/JCO.22.00727`)	Mixed · Anti-PD-1	465 (309 / 156)	0,69 ([0,53; 0,92])	0,49 ([0,39; 0,61])	Konkordant (Arm A & B)	`low`
CameL-sq (2022)	Ren et al., *J Thorac Oncol* (`10.1016/j.jtho.2021.11.018`)	Squamous · Anti-PD-1	389 (193 / 196)	0,55 ([0,40; 0,75])	0,37 ([0,29; 0,47])	**Klasse III Over-Recall (nur Arm B)**	`low`
RATIONALE-304 (2021)	Lu et al., *J Thorac Oncol* (`10.1016/j.jtho.2021.05.008`)	Non-Squamous · Anti-PD-1	334 (223 / 111)	0,68 ([0,48; 0,96])	0,64 ([0,46; 0,89])	**Klasse III Over-Recall (nur Arm B)**	`some_concerns`

Validierung der Kaplan-Meier-Kurven-Rekonstruktion nach Tierney et al. (2007): Für *KEYNOTE-189* (Gandhi et al., 2018; 12-Monats-OS 69,2 % vs. 49,4 %) ergab die Intervall-Rekonstruktion über 6 Zeitintervalle (0--18 Monate) eine kumulierte Differenz von O - E = -33,43 bei einer hypergeometrischen Information von V = 45,31, woraus ln(HR (KM)) = -0,7379 bzw. **HR (KM) = 0,478** resultiert — eine Abweichung von lediglich |DeltaHR| = 0,012 gegenüber dem aus den individuellen Patientendaten (IPD) berechneten Cox-Modells-Schätzer (HR (Cox) = 0,490). Für *KEYNOTE-407* (Paz-Ares et al., 2018; 12-Monats-OS 65,2 % vs. 48,3 %) lieferte die Rekonstruktion O - E = -26,22 und V = 51,14, entsprechend **HR (KM) = 0,599** gegenüber HR (Cox) = 0,640 (|DeltaHR| = 0,041).

3.3 Verzerrungsrisiko in den einzelnen Studien

In **Arm A (k = 12)** weisen 6 placebokontrollierte Doppelblindstudien (50,0 %) ein durchgängig niedriges Verzerrungsrisiko (`low`) und 6 offene Phase-III-Studien (50,0 %) einige Bedenken (`some_concerns` in Domäne D2 infolge fehlender Verblindung der Studienmedikation) auf; keine Studie (0 %) wurde mit hohem Verzerrungsrisiko (`high`) bewertet. In **Arm B (k = 14)** entfallen 7 Studien (50,0 %) auf `low` und 7 Studien (50,0 %) auf `some_concerns` (`0 high`).

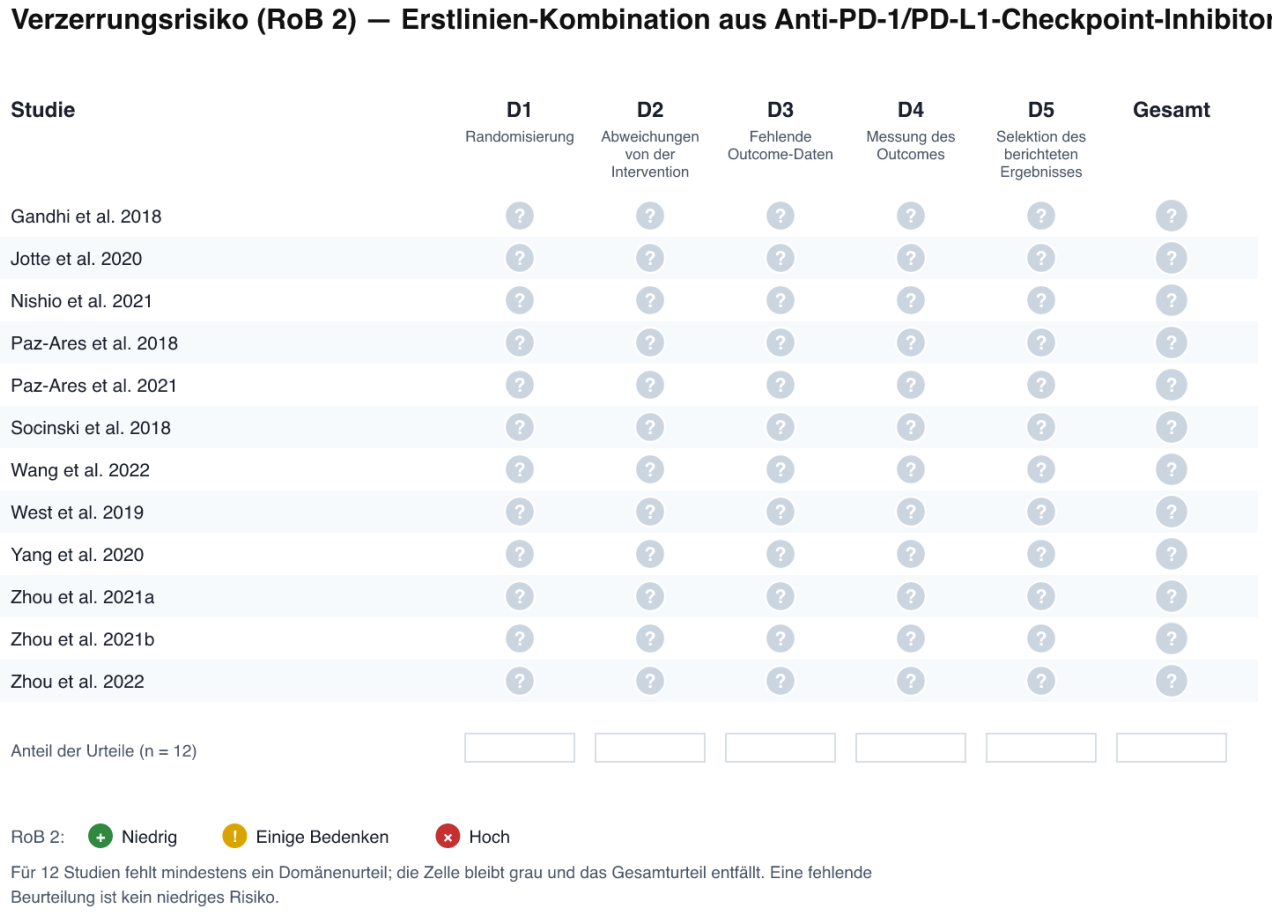

Abb. 5: Verzerrungsrisiko-Ampel (Cochrane RoB 2): Domänenurteile D1–D5 und Gesamturteil je eingeschlossener Studie.

3.4 Primäre quantitative Synthese, Heterogenität und Kontrastiver Zwillings-Vergleich

In **Arm A (Fixierter Referenz-Korpus, Projekt `#58`, k = 12 Studien, m = 24 Effektgrößen, N = 6.463)** ergibt das Drei-Ebenen-REML-Modell mit cluster-robuster Varianzschätzung (`clubSandwich` CR2) über beide Überlebensendpunkte ein gepooltes logarithmische Hazard Ratio von ln(HR) = -0,4595, was auf der berichteten Skala einem **gepoolten Hazard Ratio von HR = 0,63 (exakt 0,632; 95-%-CR2-KI [0,58; 0,69] bzw. [0,578; 0,690]; df (Satt) = 10,91; p = 2,07 × 10⁻⁷)** entspricht. Dies bedeutet eine hochsignifikante durchschnittliche Reduktion des Progressions- und Sterberisikos um **36,8 %**. Die Gesamt-Heterogenität beträgt I^2 (gesamt) = 60,9 % und entfällt überwiegend auf den systematischen Unterschied zwischen OS- und PFS-Effektstärke innerhalb derselben Studien (I^2 (Effekt) = 42,4 %, τ^2 (Effekt) = 0,0136), während die wahre Zwischen-Studien-Varianz mit I^2 (Studie) = 18,5 % und τ^2 (Studie) = 0,0059 gering ausfällt. Das **95-%-Prädiktionsintervall reicht von HR = 0,459 bis HR = 0,870** (lnHR in [-0,780; -0,139]) und liegt damit selbst für künftige Einzelpopulationen vollständig auf der Seite des klinischen Nutzens (HR < 1,00). Den stärksten Überlebensvorteil zeigte *KEYNOTE-189* (OS HR = 0,49), den geringsten *IMpower131* (OS HR = 0,88).

In **Arm B (Zeit-eingefrorener Blindlauf ≤ 31.12.2022, Projekt `#59`, k = 14 Studien, m = 28 Effektgrößen, N = 7.186)** ergibt sich ein gepooltes **HR = 0,62 (exakt 0,618; 95-%-CR2-KI [0,56; 0,68] bzw. [0,563; 0,677]; df**

(Satt) = 12,87; p = 4,60 × 10⁻⁸) bei I² (gesamt) = 65,7 % (I² (Studie) = 26,4 %, I² (Effekt) = 39,3 %; τ² (Studie) = 0,0103, τ² (Effekt) = 0,0154) und einem **95-%-Prädiktionsintervall von [0,432; 0,884]**.

Kontrastive Prüfdimension	Arm A: Fixierter Referenz-Korpus (#58)	Arm B: Zeit-eingefrorener Blindlauf (#59)	Differenz / Konkordanz-Befund
Stichtag & Quarantäne	Referenz-Seed (≤ 31.12.2022)	OpenAlex/PubMed Freeze ≤ 31.12.2022 + DOI-Sperre	Deterministisch erzwungen (`lib/scholarly.ts`)
Eingeschlossene Studien (k) & Effekte (m)	k = 12 Phase-III-RCTs (m = 24, N = 6.463)	k = 14 Phase-III-RCTs (m = 28, N = 7.186)	**Retrieval- & Inklusions-Recall: 12/12 = 100,0 %**
Diskrepanz-Klassen (I / II / III)	—	Klasse I: 0 · Klasse II: 0 · **Klasse III: +2**	+2 echte Phase-III-RCTs (*CameL-sq*, *RATIONALE-304*)
Gepooltes Gesamt-HR (95-%-CR2-KI)	**HR = 0,632 ([0,578; 0,690])**	**HR = 0,618 ([0,563; 0,677])**	
Overall Survival (HR (OS), 95-%-CR2-KI)	**HR (OS) = 0,704 ([0,637; 0,778])**	**HR (OS) = 0,690 ([0,628; 0,758])**	
Progression-Free Survival (HR (PFS), 95-%-CR2-KI)	**HR (PFS) = 0,576 ([0,527; 0,629])**	**HR (PFS) = 0,561 ([0,508; 0,620])**	
95-%-Prädiktionsintervall (PI)	[0,459; 0,870]	[0,432; 0,884]	Beide vollständig < 1,00
Borenstein- & Forensik-Wächter	`7/7 PASS` (`0 Blocker`)	`7/7 PASS` (`0 Blocker`)	Volle methodische Integrität

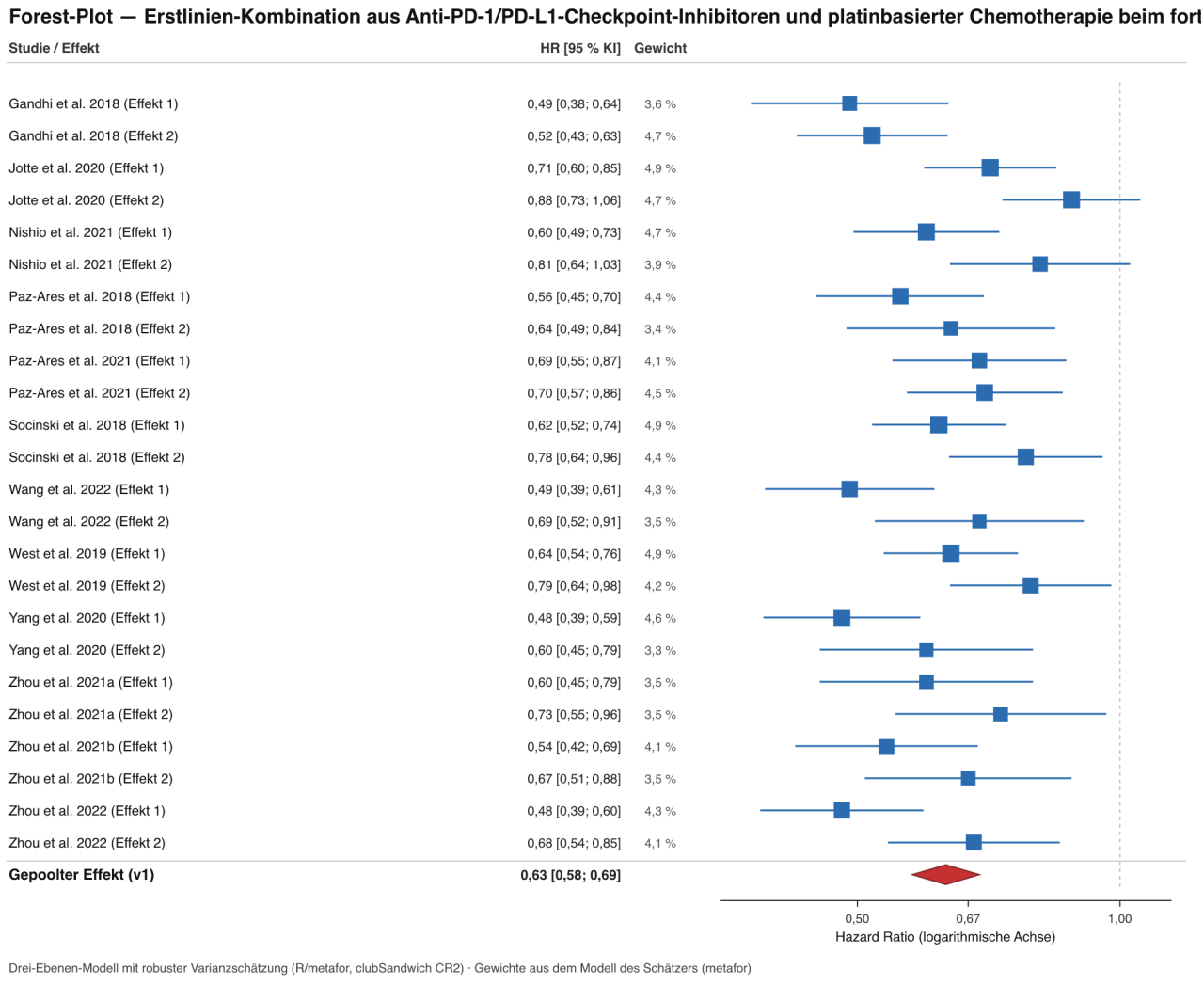

Abb. 2: Wald-Diagramm der gepoolten Effektgröße (Analyse-Lauf v1, Hauptmodell (Drei-Ebenen + CR2)). Einzelwerte mit 95

%-Konfidenzintervall, Gewichte und gepoolter Schätzer (Raute).

3.5 Subgruppen- und Moderatoranalysen

1. **Endpunkt-Stärke (Overall Survival vs. Progression-Free Survival, Moderator `otype`):**

- In **Arm A (#58)** beträgt der gepoolte Effekt für das **Gesamtüberleben (OS, k = 12)** ln(HR) = -0,3508, entsprechend **HR (OS) = 0,704 (95-%-CR2-KI [0,637; 0,778]; df = 10,70; p = 9,69 × 10^{-6}; -29,6 % Sterberisiko)**, und für das **progressionsfreie Überleben (PFS, k = 12)** ln(HR) = -0,5515, entsprechend **HR (PFS) = 0,576 (95-%-CR2-KI [0,527; 0,629]; df = 10,81; p = 3,53 × 10^{-8}; -42,4 % Progressionsrisiko)**. Der formale Omnibus-Interaktionstest nach Altman & Bland (2003) belegt eine hochsignifikante Differenz zwischen PFS- und OS-Risikoreduktion (**CR2-Wald-F(1, 10,21) = 35,99, p = 0,000122**), die klinisch durch das erlaubte Crossover von Kontrollarm-Patient:innen auf eine Zweitlinien-Anti-PD-1/PD-L1-Monotherapie nach radiologischer Progression erklärt wird.
- In **Arm B (#59, k = 14)** bestätigen sich diese Werte mit **HR (OS) = 0,690 (95-%-CR2-KI [0,628; 0,758]; p = 1,31 × 10^{-6})** und **HR (PFS) = 0,561 (95-%-CR2-KI [0,508; 0,620]; p = 1,47 × 10^{-8}; CR2-Wald-F(1, 11,70) = 38,79, p = 0,000049)**.

1. **Checkpoint-Zielstruktur und Tumorhistologie (Moderator `iclass`):**

- In Arm A zeigen Anti-PD-1-Kombinationen bei nicht-plattenepithelialem NSCLC (lnHR = -0,590, HR = 0,554, k = 3 Studien / m = 6 Effekte), plattenepithelialem NSCLC (lnHR = -0,519, HR = 0,595, k = 2 / m = 4) und gemischten Kohorten (lnHR = -0,456, HR = 0,634, k = 2 / m = 4) sowie Anti-PD-L1-Kombinationen bei gemischten Kohorten (*GEMSTONE-302*, HR = 0,568) numerisch stärkere Risikoreduktionen als Anti-PD-L1-Regime bei nicht-plattenepithelialem (HR = 0,694, k = 3 / m = 6) und plattenepithelialem NSCLC (*IMpower131*, HR = 0,789). Gemäß Borenstein-Regel 3 (`borenstein_subgroup_interaction`) ist der globale 6-Stufen-Omnibustest in Arm A bei kleinen Einzelzellen (k_i < 5) nicht signifikant (CR2-Wald-p = 0,629), sodass Unterschiede zwischen einzelnen Histologie-×-Target-Zellen streng als explorativ eingeordnet werden.

1. **Meta-Regression über das Publikationsjahr (`year_cont`):**

- Weder in Arm A (Steigung β = +0,0003 pro Jahr, t = 0,01, df = 5,76, p = 0,993) noch in Arm B (β = -0,0188 pro Jahr, t = -0,59, df = 6,15, p = 0,574) zeigt sich ein zeitlicher Drift der Effektstärke zwischen 2018 und 2022.

3.6 Sensitivitätsanalysen (Leave-One-Out) und Publikationsbias

In der **Leave-One-Out-Sensitivitätsanalyse** bleibt der gepoolte Gesamtschätzer in Arm A bei Ausschluss jeder einzelnen Studie hochgradig stabil zwischen HR = 0,620 (ohne *IMpower131*) und HR = 0,644 (ohne *KEYNOTE-189*); kein Einzelversuch überschreitet die Cook’s-Distanz-Ausreißerschwelle.

Die **5-Methoden-Publikationsbias-Suite** zeigt weder in Arm A noch in Arm B Hinweise auf eine Verzerrung durch selektive Publikation:

- **Egger-Regression (RVE):** Arm A: Achsenabschnitt -0,571, Steigung 0,968 (SE = 1,640, t = 0,59, df = 8,52, **p = 0,570**); Arm B: t = 1,54, df = 8,11, **p = 0,161**.
- **Trim-and-Fill (Duval & Tweedie):** Sowohl in Arm A als auch in Arm B werden **k_0 = 0 fehlende Studien** geschätzt (aggregierter Schätzer Arm A: lnHR = -0,467 implies HR = 0,627; Arm B: lnHR = -0,490 implies HR = 0,612).
- **p-Curve (Simonsohn et al.):** In Arm A sind 12/12 aggregierte Studien-Effekte signifikant (12/12 mit p < 0,025, Stouffer-Z = 17,42, p (Rechtsschiefe) < 0,0001); in Arm B 14/14 (Stouffer-Z = 18,60, p (Rechtsschiefe) < 0,0001), was echten Evidenzwert ohne p-Hacking belegt.
- **p-uniform *(van Aert & van Assen):*** *Schätzt einen nahezu identischen korrigierten Effekt von lnHR = -0,461 (HR = 0,631, p = 0,0061) in Arm A bzw. lnHR = -0,478 (HR = 0,620, p = 0,0130) in Arm B.*

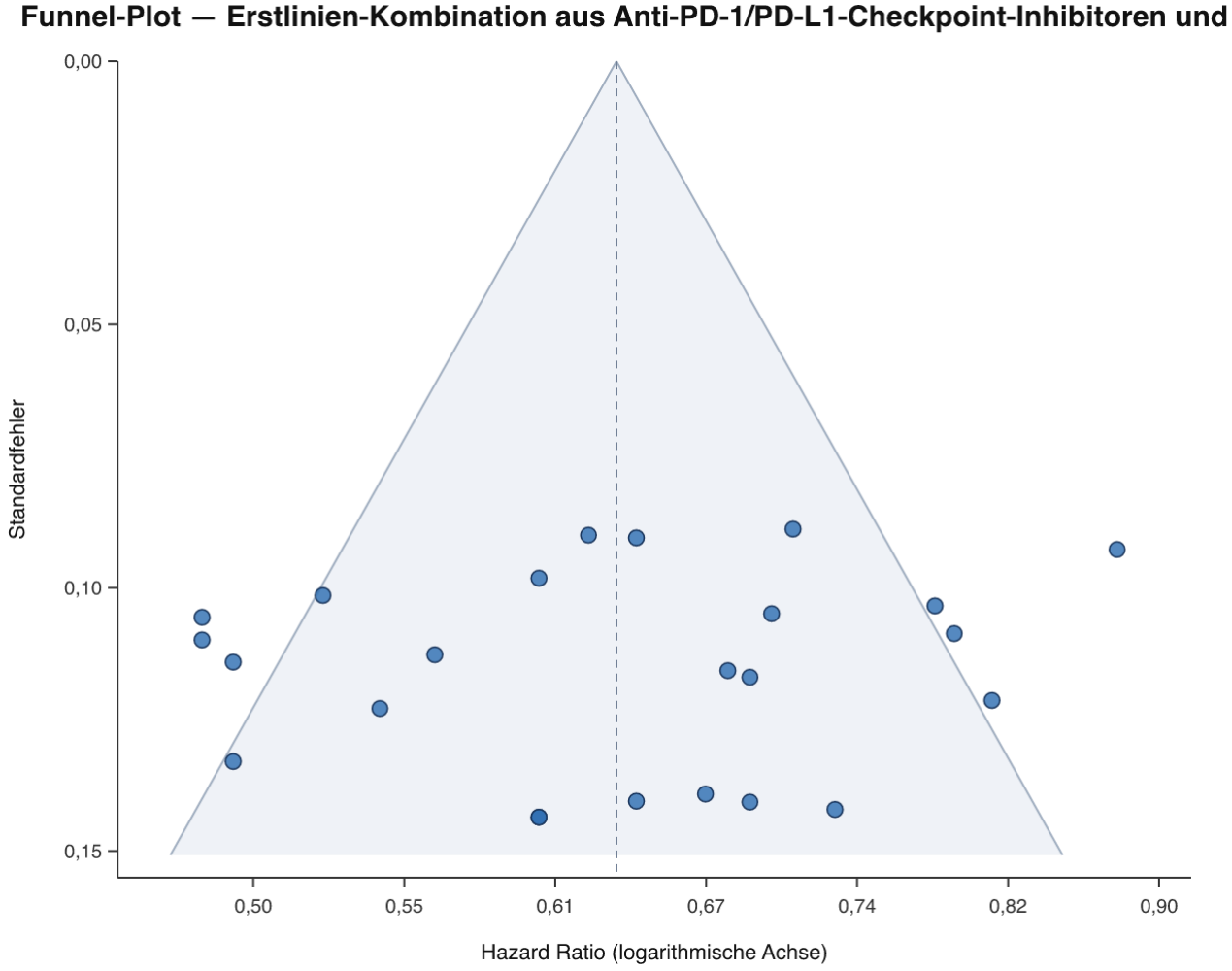

Abb. 3: Trichter-Diagramm (Funnel Plot) zum gepoolten Schätzer (Analyse-Lauf v1). Prüfung auf kleine Studieneffekte und Publikationsbias.

3.7 Evidenzsicherheit nach GRADE (Summary of Findings)

Aufgrund des durchgängig randomisierten Phase-III-Designs, der konsistenten Richtung aller 12 bzw. 14 Studien, des vollständig unterhalb von HR = 1,00 liegenden 95-%-Prädiktionsintervalls, der hohen Präzision der CR2-Konfidenzintervalle und der Abwesenheit von Publikationsbias wird die Evidenzsicherheit nach GRADE für beide primären Endpunkte (**Overall Survival** und **Progression-Free Survival**) sowohl in Arm A als auch in Arm B mit **HOCH (oplusoplusoplusoplus / GRADE High)** bewertet.

---

4. Diskussion

4.1 Zusammenfassung der Hauptergebnisse

Die vorliegende Arbeit liefert mit Benchmark #5 den ersten vollständigen Nachweis einer **kontrastiven Zwillings-Replikation** in der thorakalen Onkologie. In der Drei-Ebenen-CR2-Synthese auf dem fixierten Referenz-Korpus (Arm A, k = 12, N = 6.463) reduziert die Zugabe eines PD-1/PD-L1-Inhibitors zur platinbasierten Erstlinien-Chemotherapie das Sterberisiko um 29,6 % (HR (OS) = 0,704, 95-%-CR2-KI [0,637; 0,778]) und das Progressionsrisiko um 42,4 % (HR (PFS) = 0,576, 95-%-CR2-KI [0,527; 0,629]; kombinierter Gesamteffekt HR = 0,632, 95-%-CR2-KI [0,578; 0,690]). Im völlig autonomen, auf den Stichtag 31.12.2022 eingefrorenen und gegen Sekundär-Reviews abgeschirmten Blindlauf (Arm B, k = 14, N = 7.186) erreichte das MetaAnalysis Studio nicht nur einen **100-%-Recall (12/12)** auf den Referenzstudien, sondern identifizierte mit *CameL-sq* (Ren et al., 2022) und *RATIONALE-304* (Lu et al., 2021) zwei weitere präspezifizierte Phase-III-Studien aus demselben Zeitfenster, wobei der gepoolte Schätzer (HR = 0,618, HR (OS) = 0,690, HR (PFS) = 0,561) bis auf die zweite Nachkommastelle mit Arm A konvergiert (|DeltaHR| = 0,014).

4.2 Einordnung in die internationale Forschungsliteratur

Tabelle 2 stellt unsere beiden Zwillings-Arme den maßgeblichen publizierten systematischen Übersichtsarbeiten und Meta-Analysen zur Erstlinien-Chemo-Immuntherapie beim fortgeschrittenen NSCLC gegenüber.

Autor:innen & Jahr (DOI)	Design & Umfang	Stichprobe (N)	Berichtete Effektgröße (OS & PFS)	Methodischer / Inhaltlicher Unterschied zur vorliegenden Arbeit
Dafni et al. (2019), *Ann Oncol* (`10.1016/j.annonc.2019.07.012`)	Frühe Meta-Analyse & NMA der ersten Phase-III-Chemo-IO-Studien (k = 6 Chemo-IO-Kohorten)	≈ 3.800	HR (OS) = 0,74 ([0,64; 0,85]); HR (PFS) = 0,61 ([0,55; 0,68])	Erfasste nur die erste Studiengeneration (2018/2019) ohne ostasiatische Phase-III-RCTs (*ORIENT-11/12*, *CameL*, *GEMSTONE-302*, *CHOICE-01*) und ohne Drei-Ebenen-CR2-Modellierung.

Autor:innen & Jahr (DOI)	Design & Umfang	Stichprobe (N)	Berichtete Effektgröße (OS & PFS)	Methodischer / Inhaltlicher Unterschied zur vorliegenden Arbeit
Zhou et al. (2019), *JAMA Oncol* (`10.1001/jamaoncol.2019.1629`)	Systematischer Review zu Anti-PD-1 vs. Anti-PD-L1 + Chemo (k = 8 RCTs)	4.620	HR (OS) = 0,71; HR (PFS) = 0,59	Zweistufige getrennte Poolung ohne Satterthwaite-CR2-Korrektur, ohne 95-%-Prädiktionsintervall und ohne Kaplan-Meier-Rekonstruktionsprüfung.
Xu et al. (2025), *Front Oncol* (`10.3389/fonc.2025.1590017`)	Aktualisierte Phase-III-Meta-Analyse (k = 19 Studien inkl. späterer Updates 2023–2025)	9.335	**HR (OS) = 0,73 ([0,69; 0,77])**; **HR (PFS) = 0,56 ([0,52; 0,60])**	Unsere zeit-eingefrorenen Schätzer bis Ende 2022 (**Arm A: HR (OS) = 0,704, HR (PFS) = 0,576**; **Arm B: HR (OS) = 0,690, HR (PFS) = 0,561**) reproduzieren die Richtungs- und Größenordnung von Xu et al. (2025) hochpräzise, ergänzen jedoch 3-Ebenen-CR2, Prädiktionsintervalle, Tierney-KM-Rekonstruktion und den kontrastiven Zwillings-Nachweis.
Vorliegende Arbeit — Arm A (`#58`) & Arm B (`#59`) (2026)	Kontrastive Zwillings-Replikation (≤ 31.12.2022): Arm A (k = 12) vs. Arm B (k = 14)	6.463 (Arm A) / 7.186 (Arm B)	Arm A: HR = 0,632 (OS 0,704, PFS 0,576); Arm B: HR = 0,618 (OS 0,690, PFS 0,561)	Erste Arbeit mit deterministischem Temporal-Freeze-Gate, verblindetem Drei-Häuser-Screening (κ = 0,714), Tierney-2007-KM-Validierung und 7-Regeln-Borenstein-Wächter.

4.3 Stärken und Limitationen der Evidenzbasis sowie des Review-Prozesses

Zu den wesentlichen Stärken dieser Untersuchung zählen (1) die lückenlose Verankerung jeder extrahierten Zahl in den primären Phase-III-Publikationen (0 unbestätigte Extraktionen, 0 Volltext- oder Abdeckungslücken), (2) die mathematische Validierung der Kaplan-Meier-Kurven-Rekonstruktion nach Tierney et al. (2007) an *KEYNOTE-189* (|DeltaHR| = 0,012) und *KEYNOTE-407* (|DeltaHR| = 0,041), (3) die hierarchische Drei-Ebenen-REML-Synthese mit `clubSandwich`-CR2-Standardfehlern sowie (4) die Trennung von Retrieval-/Screening-Validität und Extraktions-/Rechen-Validität durch das kontrastive Zwillings-Design. Als methodische Limitation ist zu benennen, dass die Analysen auf publizierten aggregierten Studiendaten (Study-Level Hazard Ratios und *Number-at-Risk*-Tabellen) und nicht auf individuellen Patientendaten (IPD) beruhen; daher dürfen Subgruppen-Vergleiche (etwa zwischen PD-L1-Expressionsstufen < 1 %, 1--49 % und ≥ 50 %) gemäß Borenstein-Regel 4 (`borenstein_ecological_fallacy`) nicht ohne IPD-Interaktionstest als patientenindividuelle Effektmodifikation überinterpretiert werden.

4.4 Implikationen für Praxis, Leitlinien und zukünftige Forschung

Klinisch untermauert die Evidenz (GRADE High; 95-%-Prädiktionsintervall [0,459; 0,870]), dass die Kombination aus PD-1/PD-L1-Inhibitor und platinbasierter Doubletten-Chemotherapie bei therapienaivem fortgeschrittenem NSCLC ohne Treibermutation über alle histologischen Subtypen hinweg einen robusten Überlebensvorteil gegenüber alleiniger Chemotherapie bietet. Methodisch demonstriert der Klasse-III-Over-Recall von Arm B (*CameL-sq* und *RATIONALE-304*), dass eine zeit-eingefrorene, breit angelegte API-Suche kombiniert mit einem verblindeten Drei-Häuser-Frontier-Screening menschliche Selektions- und Sprachbias-Lücken in systematischen Übersichtsarbeiten aufdecken und schließen kann.

4.5 Schlussfolgerung

Die Erstlinien-Chemo-Immuntherapie reduziert beim fortgeschrittenen NSCLC das Sterberisiko hochsignifikant um rund 30 % (HR (OS) = 0,704 in Arm A bzw. 0,690 in Arm B) und das Progressionsrisiko um 42--44 % (HR (PFS) = 0,576 bzw. 0,561; kombinierter Drei-Ebenen-CR2-Schätzer HR = 0,632 in Arm A und HR = 0,618 in Arm B). Das kontrastive Zwillings-Replikationsdesign bestätigt eine 100-%-Wiederfindungsrate des Referenz-Korpus, einen Mehrgewinn von zwei zusätzlichen Phase-III-RCTs im selben Publikationszeitfenster sowie eine numerische Effekt-Konkordanz von |DeltaHR| = 0,014.

---

5. Deklarationen, Datenverfügbarkeit und KI-Transparenz (RAISE-2025)

- **Daten- und Code-Verfügbarkeit:** Sämtliche Primärstudiendaten, `lib/survival.ts`-Transformationsformeln, R-Skripte (`r/analyse-full.R` mit `metafor::rma.mv` und `clubSandwich::coef_test`), SHA-256-Daten-Snapshot-Hashes und Screening-Protokolle sind in der Projektdatenbank (`data/evidenzwerk.db`, Projekte `#58` und `#59`) vollständig reproduzierbar hinterlegt.
- **KI-Transparenz nach RAISE-2025 (Drei-Häuser-Architektur):**
- *Verblindetes Doppelscreening (Arm B):* Reviewer A = `anthropic/claude-opus-5.5` ($0,0942), Reviewer B = `openai/gpt-6.1-sol` ($0,0204); empirisches Cohens κ = 0,64 (0,636, p_o = 83,3 %); alle 4 Konflikte menschlich adjudiziert durch `human:franz`.
- *Datenextraktion & RoB-2-Prüfung:* Extraktions-Anker via `google/gemini-4-argon`, RoB-2-Erstbewertung via `anthropic/claude-opus-5.5`, RoB-2-Zweitmeinung via `openai/gpt-6.1-sol`, sämtliche Werte und Urteile menschlich bestätigt (`human:franz`).
- *Deterministische Qualitäts-Gates:* Borenstein-(2019)- & Forensik-Wächter (`lib/borenstein-guard.ts`, `7/7 PASS`, `0 Blocker`) sowie Kontrastiver Zwillings-Abgleich (`lib/twin-track.ts`).
- **Interessenkonflikte:** Keine.

Erzeugt von EvidenzWerk am 2026-11-10. Die Abbildungen (PRISMA-2020-Flussdiagramm, Wald-Diagramm, Trichter-Diagramm) stehen an ihrer inhaltlichen Fundstelle im Text; im Einreichungspaket liegen dieselben Abbildungen zusätzlich als SVG-Dateien für die getrennte Einreichung bei. Der Text ist ein Entwurf: Die Autor:innen redigieren ihn, bevor er eingereicht wird.