Validierungsstudie Benchmark 1: End-to-End-Replikation und methodische Erweiterung der kardiorenalen SGLT2-Inhibitor-Evidenzsynthesen (The Lancet 2022–2024; k = 15 RCTs, N = 100.952)

Dreistufiges Publikationsprotokoll (Stufe 1: Blinde Selbstfindung & Live-Extraktion · Stufe 2: Drei-Ebenen-REML+CR2-Synthese · Stufe 3: Autonomes 12.859-Wörter-Manuskript im Vergleich mit SMART-C, Vaduganathan et al. & Khan/Usman et al.)

Untertitel: Ein dreistufiger Stresstest aus blinder Evidenzbeschaffung („Einstieg“), mehrstufiger statistischer Synthese („Mittelteil“) und autonomer wissenschaftlicher Manuskriptgenerierung („Endergebnis“) im direkten Vergleich mit den menschlichen Referenzarbeiten des SMART-C-Konsortiums

Projekt-Referenz (Live-Instanz): `https://metaanalysis.brain-hr.de/p/16` · **Kryptographischer Daten-Snapshot-Hash (SHA-256):** `171f231cf6005e28` · **Ratifizierter Hauptlauf:** `Run v3` (`metafor::rma.mv` + `clubSandwich` CR2) · **Ratifizierte Evidenzsicherheit:** GRADE oplusoplusoplusoplus (*Hohe Evidenzsicherheit*)

Strukturierte Zusammenfassung (Abstract für Methodenpublikation)

Hintergrund und Fragestellung: Trotz erheblicher Fortschritte bei großen Sprachmodellen (LLMs) scheitern generative Systeme in der systematischen Evidenzsynthese regelmäßig an drei strukturellen Hürden: (1) unvollständiger Literaturauffindung und Paywall-Blockaden bei Primärstudien, (2) mangelnder mathematischer Strenge oder erfundenen Zahlen („Halluzinationen“) bei der statistischen Poolung abhängiger Effektgrößen sowie (3) Brüchen zwischen statistischem Rechenkern und ausformuliertem Berichtsmanuskript. Um zu prüfen, ob eine konsequent architekturgetrennte Syntheseplattform (**EvidenzWerk**) eine komplex aufgebaute klinische Goldstandard-Meta-Analyse von der ersten Suchanfrage bis zum einreichungsfertigen Manuskript autonom („ohne menschliches Domänen-Vorwissen“) replizieren und methodisch sogar übertreffen kann, führten wir einen dreistufigen Validierungs-Benchmark durch.

Referenzstandard (Goldstandard): Als klinischer und statistischer Maßstab dienten drei hochrangige kardiorenale Referenz-Meta-Analysen aus der *Lancet*-Familie zu Natrium-Glukose-Kotransporter-2-(SGLT2)-Inhibitoren: (1) die kollaborative Meta-Analyse des **SMART-C-Konsortiums** (*The Lancet*, 2022; *k* = 13 Studien, *N* = 90.413), (2) die präspezifizierte Herzinsuffizienz-Synthese von **Vaduganathan et al.** (*The Lancet*, 2022; *k* = 5 Studien, *N* = 21.947) sowie (3) die um die Post-Myokardinfarkt-Studien (`DAPA-MI`, `EMPACT-MI`) erweiterte Gesamtsynthese von **Khan, Usman, Butler, Bhatt et al.** (*The Lancet Diabetes & Endocrinology*, 2024; *k* = 15 Studien, *N* = 100.952).

Methodik des 3-Stufen-Stresstests:

1. **Stufe 1 („Einstieg“ — Blinde Selbstfindung & blinde Live-Extraktion):** Ohne Übergabe eines einzigen Studien-Akronyms (`EMPA-REG`, `DAPA-HF`, `DELIVER` etc.) und ohne Zugriff auf verlagsgeschützte Volltext-PDFs (*NEJM*, *The Lancet*) suchte das System autonom über PubMed/MEDLINE, OpenAlex, ClinicalTrials.gov und eine bidirektionale Schneeball-Zitationsernte nach den qualifizierenden Phase-III-Studien (N ≥ 1.000) und extrahierte die Effektgrößen über eine dreistufige Open-Evidence-Ersatzkette (Europe PMC JATS-XML arrow PubMed EFetch-XML mit automatischer `<DataBankList>`-`NCT`-Erkennung arrow ClinicalTrials.gov API v2 `resultsSection`).
2. **Stufe 2 („Mittelteil“ — Mehrstufige statistische Synthese & Bias-Triangulation):** Über eine deterministische R-Bridge (`metafor::rma.mv`, Restricted Maximum Likelihood [REML] mit cluster-robuster Varianzschätzung `clubSandwich::vcovCR` vom Typ CR2 und Satterthwaite-Freiheitsgraden nach Pustejovsky & Tipton, 2018) wurden 161 Effektgrößen über 5 harte

klinische Endpunkte und 4 Krankheitsentitäten parallel auf der relativen (HR/RR) und absoluten (RD/NNTB) Skala synthetisiert, ergänzt um eine 5-Verfahren-Publikationsbias-Triangulation (Egger, Trim-and-Fill, PET-PEESE, p-curve, p-uniform*), RoB-2-Bewertungen (75 Domänenurteile) und GRADE-Evidenzprofile.

3. **Stufe 3 („Endergebnis“ — Autonome Manuskripterzeugung & Text-zu-Text-Vergleich):** Aus dem kryptographisch gesicherten Daten-Snapshot generierte das System ein 12.859 Wörter umfassendes wissenschaftliches Manuskript nach PRISMA 2020 und RAISE-2025 sowie ein 20-teiliges Einreichungspaket, das qualitativ und quantitativ mit den Originaltexten in *The Lancet* verglichen wurde.

Ergebnisse:

- **Stufe 1 (Blinde Selbstfindung & Extraktion):** EvidenzWerk identifizierte autonom **15 von 15 Zielstudien (100,0 % Recall)** und ordnete **15 von 15 NCT-Registernummern (100,0 %)** automatisch aus den PubMed-XML-Metadaten zu. Allein aus frei zugänglichen PubMed-Abstracts und ClinicalTrials.gov-Registertabellen (ohne ein einziges Paywall-PDF) traf die blinde KI-Extraktion bei **12 von 15 Studien (80,0 %)** die publizierte *Lancet*-Hazard-Ratio samt 95-%-Konfidenzintervall exakt auf die zweite Nachkommastelle. Bei den 3 älteren Diabetes-/CKD-Studien (EMPA-REG OUTCOME, CANVAS Program, EMPA-KIDNEY), deren Abstract nur den primären 3P-MACE- bzw. renalen Kompositendpunkt ausweist, trat **keine Halluzination (0 %)** auf; das System meldete transparent die Endpunktabweichung. Selbst der unbereinigte blinde Roh-Pool über alle 15 Studien lag mit HR (blind) = 0,792 [0,755; 0,831] nur um DeltaHR = 0,011 neben dem *Lancet*-Goldstandard (HR = 0,781 [0,744; 0,820]).
- **Stufe 2 (Statistische Replikation):** Nach Vervollständigung der 3 Sekundärendpunkt-Werte replizierte der Rechenkern die *Lancet*-Referenzwerte über alle 5 Endpunkte und alle 4 Krankheitsentitäten auf DeltaHR = 0,000 bis 0,003 genau: Primärendpunkt *Kardiovaskulärer Tod oder Herzinsuffizienz-Hospitalisierung* HR = 0,781 [0,748; 0,816] ($p < 0,0001$, I^2 (Studie) = 0,0 %, 95-%-Prädiktionsintervall [0,75; 0,82]); *Erste HHF* exp(hatβ) = 0,71; *Kardiovaskulärer Tod* exp(hatβ) = 0,87; *Gesamtmortalität* exp(hatβ) = 0,89; *Nierenkrankheitsprogression* exp(hatβ) = 0,68.
- **Stufe 3 (Methodischer und textlicher Mehrwert gegenüber *The Lancet*):** Das autonom verfasste Manuskript übertraf die menschlichen Referenzarbeiten in vier methodischen Dimensionen: (a) Aufdeckung einer klinisch bedeutsamen **Rangfolgen-Umkehr zwischen relativer und absoluter Risikoskala** (Nierenkrankheitsprogression hat mit exp(hatβ) = 0,68 den stärksten relativen Effekt, wegen niedrigerer Grundinzidenz mit RD = -0,010 / NNTB ≈ 100 aber einen geringeren absoluten Effekt als HHF mit exp(hatβ) = 0,71 und RD = -0,017 / NNTB ≈ 59); (b) mathematische Auflösung des scheinbaren **Heterogenitäts-Paradoxons** (I^2 = 0 % auf Einzelendpunkt-Ebene vs. I^2 = 77,7 % im Mehr-Endpunkt-Modell durch Zerlegung in σ^2 (Studie) und σ^2 (Effekt), Wald-p = 0,003); (c) methodisch strenge Interpretation der **Post-Myokardinfarkt-Subgruppe** (DAPA-MI, EMPACT-MI; HR = 0,90 [0,77; 1,04], Interaktions-p = 0,268) ohne Fehlschluss von Nicht-Signifikanz bei k = 2 auf Wirksamkeitsäquivalenz; sowie (d) automatische Generierung einer strukturierten Literatur-Vergleichssynopse. Zudem deckte der Benchmark sechs subtile Software-Schwachstellen auf, die im Zuge der Kalibrierung behoben wurden.

Schlussfolgerung: Eine strikt vom deterministischen Rechenkern getrennte, durch kryptographische Snapshots und menschliche Ratifikation gesicherte KI-Synthesearchitektur kann komplexe medizinische Referenz-Meta-Analysen nicht nur numerisch exakt replizieren, sondern liefert durch automatisierte Drei-Ebenen-CR2-Modellierung, parallele Absolutrisiko-Quantifizierung und lückenlose PRISMA-2020-/RAISE-Dokumentation eine Berichtsqualität, die etablierte Expertenpublikationen in zentralen Punkten der methodischen Transparenz ergänzt und übertrifft.

---

1. Wissenschaftlicher Hintergrund und Einbettung in die Standardliteratur der Meta-Analyse

1.1 Das Replikations- und Qualitätsproblem klassischer und KI-gestützter Evidenzsynthesen

Systematische Übersichtsarbeiten und Meta-Analysen bilden die Spitze der klinischen Evidenzhierarchie (Higgins et al., 2023; Borenstein et al., 2021). Gleichzeitig belegen empirische Metaforschungsstudien, dass ein erheblicher Teil publizierter Meta-Analysen unter methodischen Mängeln leidet: unvollständige Suchstrategien ohne dokumentierte Datenbank-Syntax (Rethlefsen et al., 2021; Page et al., 2021), Ignorieren stochastischer Abhängigkeiten bei mehreren Endpunkten oder Messzeitpunkten pro Primärstudie (Cheung, 2014; Van den Noortgate et al., 2013; Pustejovsky & Tipton, 2022), Verwendung des bei kleiner Studienzahl verzerrten DerSimonian-Laird-Schätzers ohne Hartung-Knapp-Korrektur (IntHout et al., 2014; Veroniki et al., 2016; Langan et al., 2019), fehlende Angabe von 95-%-Prädiktionsintervallen (Riley, Higgins & Deeks, 2011; IntHout et al., 2016) sowie die einseitige Berichterstattung relativer Effektmaße (`OR`, `RR`, `HR`) ohne absolute Risikodifferenzen (`RD`) und *Numbers Needed to Treat* (`NNTB`/`NNTH`) nach Altman (1998) und den GRADE-Leitlinien (Guyatt et al., 2011; Schünemann et al., 2013).

Der unkontrollierte Einsatz generativer Sprachmodelle (Large Language Models, LLMs) verschärft diese Problematik, wenn das Sprachmodell selbst zum Rechnen oder zum freien Zusammenfassen unstrukturierter PDFs genutzt wird: Sprachmodelle interpolieren fehlende Konfidenzintervalle, verwechseln Standardabweichungen (`SD`) mit Standardfehlern (`SE`) und erzeugen im Fließtext Zahlen, die von den zugrunde liegenden Tabellen abweichen. Die 2025 veröffentlichten **RAISE-Empfehlungen (*Responsible AI in Evidence Synthesis*)** fordern daher zwingend:

1. eine **strikte Trennung** zwischen sprachmodellgestützter Extraktionshilfe und deterministischer mathematischer Berechnung,
2. eine **Beleganker-Pflicht** (wörtliches Textzitat und Fundstelle für jeden extrahierten Kennwert),
3. eine **menschliche Ratifikations-Schranke** (*Human-in-the-Loop*) vor der Aufnahme von KI-Vorschlägen in die Evidenzsynthese sowie
4. eine **unveränderliche Auditierbarkeit** (Versionierung jedes Analyse-Laufs mit kryptographischem Daten-Snapshot-Hash).

1.2 Methodische Referenzstandards, an denen EvidenzWerk kalibriert wird

Um den Anspruch zu prüfen, dass eine mit **EvidenzWerk** erstellte Synthese den strengsten internationalen Maßstäben genügt, wurde die Architektur des Systems direkt gegen die kanonische Methodenliteratur der Evidenzsynthese kalibriert:

- **Berichts- und Suchstandards:** Vollständige Implementation von **PRISMA 2020** (Page et al., *BMJ* 2021; 27 Haupt- und 42 Einzelitems), **PRISMA-P** (Moher et al., 2015) und **PRISMA-S** (Rethlefsen et al., 2021) einschließlich einer aus dem Entscheidungs-Audit-Trail berechneten PRISMA-Flussbilanz, die gemäß PRISMA-2020-Item 16a zwischen *nicht gesichteten Treffern (Trefferlimit)* und *durch Automatisierung/Triage ausgeschlossenen Treffern* differenziert.
- **Effektgrößen-Taxonomie und Überlebenszeit-Rekonstruktion:** Zehn exakt definierte Metrikfamilien nach Borenstein et al. (2021) und dem *Cochrane Handbook* (Higgins et al., 2023, Kap. 6 & 10): kontinuierliche Maße (Hedges' g mit exakter Korrektur J(df) = 1 - 3/4 df-1 nach Hedges, 1981; unstandardisierte Mittelwertdifferenz `MD`), binäre 2×2-Maße (`log-OR`, `log-RR`, Risikodifferenz `RD` mit kontinuitätskorrigierter Varianz bei Nullzellen nach Sweeting et al., 2004), Zeit-bis-Ereignis-Maße (`log-HR` direkt aus Konfidenzintervallen oder rekonstruiert über Log-Rank-p/O-E-Statistiken, Peto-Approximation, Kaplan-Meier-Kurven-Zerlegung nach **Tierney et al., *Trials* 2007** und **Parmar et al., *Stat Med* 1998**, oder Median-Überlebenszeiten), Inzidenzratenverhältnisse (`log-IRR`), Korrelationen (Fisher-z) sowie Einzelgruppen-Anteile (Logit und Freeman-Tukey-Doppel-Arkussinus nach Freeman & Tukey, 1950; Nyaga et al., 2014).

- **Mehrstufiger statistischer Rechenkern (Drei-Ebenen-REML + CR2):** Sobald Primärstudien mehrere Effektgrößen liefern (z. B. mehrere klinische Endpunkte wie HHF, CV-Tod, Gesamtmortalität und Nierenprogression innerhalb derselben randomisierten Studie), verletzt ein klassisches Zwei-Ebenen-Modell die Unabhängigkeitsannahme und führt zu antikonservativen Standardfehlern (Cheung, 2014; Van den Noortgate et al., 2013, 2015; Assink & Wibbelink, 2016). EvidenzWerk schätzt daher über eine gekapselte R-Bridge (`metafor::rma.mv`, **Viechtbauer, *J Stat Softw* 2010**) ein hierarchisches Drei-Ebenen-Modell mittels Restricted Maximum Likelihood (REML):

y (ij) = μ + u_i + e (ij) + ε (ij), u_i ~ mathcalN(0, σ² (Studie)), e (ij) ~ mathcalN(0, σ² (Effekt)), ε (ij) ~ mathcalN(0, v (ij)) und kombiniert dieses mit der **cluster-robusten Sandwich-Varianzschätzung vom Typ CR2** (`clubSandwich::vcovCR(..., type = "CR2")`, **Pustejovsky & Tipton, *J Bus Econ Stat* 2018; *Psychol Methods* 2022**) und kleinen Stichproben-Freiheitsgraden nach Satterthwaite. Die Gesamt-Heterogenität wird nach **Cheung (2014)** und **Konstantopoulos (2011)** über die typische Stichprobenvarianz tildev in den Zwischen-Studien-Anteil (I² (Studie) = σ² (Studie) / (σ² (Studie) + σ² (Effekt) + tildev)) und den Innerhalb-Studien-/Zwischen-Endpunkt-Anteil (I² (Effekt) = σ² (Effekt) / (σ² (Studie) + σ² (Effekt) + tildev)) zerlegt.

- **Mehrverfahren-Triangulation des Publikationsbias:** Da kein einzelnes Verfahren zur Prüfung auf selektive Publikation oder Small-Study-Effekte unter allen Heterogenitäts- und Stichprobenbedingungen unverzerrt arbeitet (Carter et al., 2019; Page, Sterne, Higgins & Egger, 2023), kombiniert EvidenzWerk fünf komplementäre Methoden auf Studienaggregat-Ebene:

1. **Egger-Regressionstest** auf Funnel-Plot-Asymmetrie (Egger et al., *BMJ* 1997),
2. **Duval-und-Tweedie Trim-and-Fill** (L_0-Schätzer, Duval & Tweedie, *Biometrics* 2000),
3. **PET-PEESE** (*Precision-Effect Test / Precision-Effect Estimate with Standard Error*, Stanley & Doucouliagos, *Res Synth Methods* 2014),
4. **Simonsohn *p*-Curve-Analyse** auf Rechtsschiefe signifikanter p-Werte (p < 0,05) und Verlust von 33 %-Power als Evidenz für *p*-Hacking vs. echten Effekt (Simonsohn, Nelson & Simmons, *J Exp Psychol Gen* 2014),
5. **van Aert *p*-uniform** *zur publikationsbias-bereinigten Maximum-Likelihood-Schätzung unter Einbezug nicht-signifikanter Studien (van Aert, Wicherts & van Assen*, Psychol Methods* 2016; van Aert & van Assen, 2019),

ergänzt um den direkten **Register-Abgleich** abgeschlossener, aber unpublizierter Studien auf ClinicalTrials.gov (`registry_unpublished`).

- **Verzerrungsrisiko und Evidenzsicherheit:** Vollständige Domänenlogik nach **Cochrane RoB 2** (Sterne et al., *BMJ* 2019; 5 Domänen D1–D5 mit deterministischer Ableitung des Gesamturteils) bzw. **ROBINS-I** (Sterne et al., *BMJ* 2016; 7 Domänen für nicht-randomisierte Studien) sowie **GRADE Summary-of-Findings-Tabellen** (Guyatt et al., 2011; Schünemann et al., 2013) mit automatischer Berechnung absoluter Risikodifferenzen pro 1.000 Behandelte und `NNTB`/`NNTH` nach der **Altman-Konvention** (Altman, *BMJ* 1998) für Konfidenzintervalle, die die Nulllinie kreuzen.

1.3 Warum die kardiorenalen SGLT2-Inhibitor-Meta-Analysen (*The Lancet* 2022–2024) den idealen ersten Härtetest bilden

Als **Benchmark 1** wurde bewusst eines der prominentesten und statistisch bestcharakterisierten Evidenzfelder der modernen Inneren Medizin gewählt: die großen, ereignisgesteuerten, doppelblinden, placebokontrollierten Phase-III-Endpunktstudien zu SGLT2-Inhibitoren (und dem dualen SGLT1/2-Inhibitor Sotagliflozin) über das gesamte kardiorenale und kardiometabolische Krankheitsspektrum. Drei Referenzarbeiten aus *The Lancet* und *The Lancet Diabetes & Endocrinology* definieren hier den internationalen Goldstandard:

1. **SMART-C Consortium & Nuffield Department of Population Health Renal Studies Group (2022):** *Impact of diabetes on the effects of sodium glucose co-transporter-2 inhibitors on kidney outcomes:*

collaborative meta-analysis of large placebo-controlled trials (*The Lancet*, 2022; 400(10365): 1788–1801; DOI `10.1016/S0140-6736(22)02292-8`; `PMC9761513`). Synthese von **13 großen Phase-III-Studien (N = 90.413)** über drei Indikationsgebiete (4 Typ-2-Diabetes/ASCVD-Studien: `EMPA-REG OUTCOME`, `CANVAS Program`, `DECLARE-TIMI 58`, `VERTIS CV`; 4 CKD-Studien: `CREDENCE`, `DAPA-CKD`, `SCORED`, `EMPA-KIDNEY`; 5 Herzinsuffizienz-Studien: `DAPA-HF`, `EMPEROR-Reduced`, `SOLOIST-WHF`, `EMPEROR-Preserved`, `DELIVER`).

2. **Vaduganathan, Docherty, Claggett, Jhund, de Boer et al. (2022):** *SGLT-2 inhibitors in patients with heart failure: a comprehensive meta-analysis of five randomised controlled trials* (*The Lancet*, 2022; 400(10354): 757–767; DOI `10.1016/S0140-6736(22)01429-5`). Präspezifizierte Trial-Level-Synthese der **5 Herzinsuffizienz-Studien (N = 21.947)** über das gesamte Spektrum der linksventrikulären Ejektionsfraktion (HFrEF, HFmrEF, HFpEF) mit dem primären Endpunkt *Kardiovaskulärer Tod oder erste Hospitalisierung wegen Herzinsuffizienz* (HR = 0,77 [0,72; 0,82]).
3. **Khan, Usman, Butler, Bhatt et al. (2024):** *SGLT2 inhibitors and cardiovascular outcomes across the cardiometabolic spectrum, including post-myocardial infarction: a systematic review and meta-analysis* (*The Lancet Diabetes & Endocrinology*, 2024; 12(8): 545–558; DOI `10.1016/S2213-8587(24)00102-5`). Erweiterung des 13-Studien-SMART-C-Korpus um die beiden 2023/2024 publizierten großen Post-Myokardinfarkt-Studien `DAPA-MI` (N = 4.017) und `EMPACT-MI` (N = 6.522) auf insgesamt **15 Phase-III-Studien (N = 100.952)**.

Dieses Triumvirat eignet sich aus vier Gründen hervorragend als Kalibrierungs-Benchmark:

- **Exakte numerische Prüfbarkeit:** Alle 15 Studien sind große Zulassungs- bzw. Phase-III-Outcome-Studien mit publizierten Hazard Ratios, 95-%-Konfidenzintervallen und 2×2-Ereignistafeln (e_1/n_1, e_2/n_2). Jede Abweichung um mehr als ± 0,01 fällt sofort auf.
- **Realer Paywall-Stresstest für Stufe 1:** Fast alle 15 Primärstudien erschienen im *New England Journal of Medicine* (*NEJM*), dessen Server automatisierte PDF-Abrufe strikt mit `HTTP 403 Forbidden` abweist. Ein System, das nur klassische Open-Access-PDFs lesen kann, würde hier komplett scheitern.
- **Sekundärendpunkt-Falle im Abstract:** In den drei ältesten bzw. nephrologischen Studien (`EMPA-REG OUTCOME` 2015, `CANVAS Program` 2017, `EMPA-KIDNEY` 2022) war der primäre Studienendpunkt im Abstract `3P-MACE` (HR = 0,86) bzw. ein renaler Kompositendpunkt (HR = 0,72), während der für unsere Synthese zentrale Endpunkt *Kardiovaskulärer Tod oder Herzinsuffizienz-Hospitalisierung* (`CV Death or HHF`: HR = 0,66, 0,78 bzw. 0,84) ein Sekundärendpunkt war. Damit lässt sich präzise testen, ob die KI bei fehlender Sekundärendpunkt-Angabe im Abstract halluziniert oder sauber zwischen Primär- und Sekundärendpunkt unterscheidet.
- **Methodische Streitfrage in Stufe 2 & 3 (Post-MI-Erweiterung und Skalen-Divergenz):** Während die 13 SMART-C-Studien bemerkenswert homogen sind (I^2 = 0 %, HR = 0,78), zeigen die beiden neuen Post-MI-Studien (`DAPA-MI` HR = 0,95, `EMPACT-MI` HR = 0,90) einen deutlich kleineren, nicht-signifikanten Effekt bei gleichzeitig niedriger Ausgangs-Ereignisrate (RD = -0,003). Dies stellt höchste Ansprüche an die Subgruppen-Interaktionsprüfung (CR2-Satterthwaite) und an die wissenschaftliche Differenzierungskraft des generierten Diskussionstextes.

---

2. Stufe 1 („Der Einstieg“): Blinde Selbstfindung und blinde Live-KI-Extraktion ohne menschliches Vorwissen

2.1 Versuchsaufbau des vollständigen Blindtests

Um auszuschließen, dass die Replikation lediglich auf vorab kuratierten Effektgrößentabellen beruht, wurde ein **vollständig verblindeter End-to-End-Selbstfindungstest** (`run_blind_self_discovery_test.js`) auf dem Produktionsserver durchgeführt. Es galten drei strikte Restriktionen:

1. **Verbot von Studien-Akronymen in der Suche:** Die Suchanfragen an PubMed/MEDLINE (`E-Utilities`

), OpenAlex und ClinicalTrials.gov durften ausschließlich generische klinische und methodische Fachbegriffe enthalten (("sodium-glucose transporter 2 inhibitors" OR empagliflozin OR dapagliflozin OR canagliflozin OR ertugliflozin OR sotagliflozin) AND ("heart failure" OR "kidney disease" OR "myocardial infarction" OR "cardiovascular outcomes") AND ("randomized" OR "double-blind" OR "placebo-controlled")), ergänzt um die automatische Schneeball-Referenzernte (referencesOfWorks) aus den identifizierten Übersichtsarbeiten. Kein einziges Akronym (DAPA-HF, DELIVER, EMPEROR, CREDENCE etc.) wurde übergeben.

2. **Kein einziger manueller PDF-Upload:** Dem System wurde kein einziges Verlags-PDF der 15 *NEJM* -Primärstudien zur Verfügung gestellt.
3. **Dreistufige Open-Evidence-Ersatzkette (fetchFallbackEvidenceText):** Sobald der direkte Volltext-PDF-Abruf einer Studie an einer Verlags-Paywall (HTTP 403) scheiterte, durchlief EvidenzWerk vollautomatisch drei offene Evidenzquellen:

- **Stufe 1A (Europe PMC Open-Access JATS-XML):** Abruf des strukturierten Volltext-XMLs über die Europe-PMC-Schnittstelle, sofern eine freie PMC-Version vorliegt.
- **Stufe 1B (PubMed EFetch-XML inkl. Auto-NCT-Discovery):** Abruf des vollständigen PubMed-XML-Datensatzes mit allen strukturierten Abstract-Sektionen (BACKGROUND, METHODS, RESULTS, CONCLUSIONS) **sowie automatischem Auslesen sämtlicher verknüpfter Studienregister-Identifikatoren (NCTxxxxxxxx) aus dem XML-Knoten** <DataBankList>.
- **Stufe 1C (ClinicalTrials.gov API v2 resultsSection):** Falls im PubMed-XML oder im Abstract eine NCT-Nummer erkannt wurde, fragte EvidenzWerk autonom die REST-API v2 von ClinicalTrials.gov (https://clinicaltrials.gov/api/v2/studies/NCT...) ab und extrahierte die hinterlegten Ergebnis- und Endpunkttabellen (resultsSection.outcomeMeasuresModule, analyses, groups, denoms).

1. **Blinde KI-Extraktion (anthropic/claude-opus-5) und deterministische Ableitung (deriveEffectFromFindings):** Der zusammengeführte Evidenztext aus Stufe 1A–1C wurde mit dem dynamisch aus dem Projekt-PICOS erzeugten Extraktionsprompt (buildPromptsForProject) an das Sprachmodell übergeben. Jedes extrahierte Ergebnis musste ein wörtliches Textzitat (quote) als Beleganker liefern; die anschließende Umrechnung in log(HR) und die Stichprobenvarianz v = (fracln(KI (oben)) - ln(KI (unten))2 · 1,959964)² erfolgte deterministisch in TypeScript (lib/stats.ts).

2.2 Ergebnisse der blinden Selbstfindung und Live-Extraktion (Studie für Studie)

Tabelle B1.1 dokumentiert das vollständige Protokoll des Blindlaufs über alle 15 Zielstudien (N = 100.952):

Nr.	Studie (Autor, Jahr, *NEJM*)	Blinde Suche gefunden?	Auto-NCT aus PubMed-XML	*Lancet* -Goldstandard HR [95 % KI] (*CV Death or HHF*)	Blinde Live-KI-Extraktion HR [95 % KI] (ohne PDF!)	Abweichung DeltaHR	Befund & Verhalten der KI im Blindlauf
1	**EMPA-REG OUTCOME** (Zinman et al., 2015)	**Ja** (PubMed + OA + Snowball)	NCT01131676	**0,66** [0,55; 0,79]	**0,86** [0,74; 0,99] *(3P-MACE)*	+0,20 *(anderer Endpunkt)*	Abstract nennt nur 3P-MACE (0,86) und HHF allein (0,65); KI meldet ehrlich: *„Komposit CV Death or HHF im Abstract nicht separat beziffert“* (0 % Halluzination)

Nr.	Studie (Autor, Jahr, *NEJM*)	Blinde Suche gefunden?	Auto-NCT aus PubMed-XML	*Lancet* -Goldstandard HR [95 % KI] (*CV Death or HHF*)	Blinde Live-KI-Extraktion HR [95 % KI] (ohne PDF!)	Abweichung DeltaHR	Befund & Verhalten der KI im Blindlauf
2	**CANVAS Program** (Neal et al., 2017)	**Ja** (PubMed + OA + Snowball)	NCT01032629	**0,78** [0,67; 0,91]	**0,86** [0,75; 0,97] *(3P-MACE)*	+0,08 *(anderer Endpunkt)*	Abstract nennt nur 3P-MACE (0,86); KI weist die Endpunktdifferenz im reason-Feld explizit aus
3	**DECLARE-TIMI 58** (Wiviott et al., 2019)	**Ja** (PubMed + OA + Snowball)	NCT01730534	**0,83** [0,73; 0,95]	**0,83** [0,73; 0,95]	**0,00**	**Exakter Volltreffer** (ko-primärer Endpunkt im Abstract + CT.gov)
4	**CREDENCE** (Perkovic et al., 2019)	**Ja** (PubMed + OA + Snowball)	NCT02065791	**0,69** [0,57; 0,83]	**0,69** [0,57; 0,83]	**0,00**	**Exakter Volltreffer** (sekundärer Endpunkt direkt aus Abstract + CT.gov extrahiert)
5	**DAPA-HF** (McMurray et al., 2019)	**Ja** (PubMed + OA + Snowball)	NCT03036124	**0,74** [0,65; 0,85]	**0,74** [0,65; 0,85]	**0,00**	**Exakter Volltreffer** (inkl. sekundärem Endpunkt CV Death + HHF)
6	**VERTIS CV** (Cannon et al., 2020)	**Ja** (PubMed + OA + Snowball)	NCT01986881	**0,88** [0,75; 1,03]	**0,88** [0,75; 1,03]	**0,00**	**Exakter Volltreffer** (erster hierarchischer Sekundärendpunkt erkannt)
7	**EMPEROR-Reduced** (Packer et al., 2020)	**Ja** (PubMed + OA + Snowball)	NCT03057977	**0,75** [0,65; 0,86]	**0,75** [0,65; 0,86]	**0,00**	**Exakter Volltreffer** (primärer Kompositendpunkt + HHF 0,70 + CV-Tod 0,92)
8	**DAPA-CKD** (Heerspink et al., 2020)	**Ja** (PubMed + OA + Snowball)	NCT03036150	**0,71** [0,55; 0,92]	**0,71** [0,55; 0,92]	**0,00**	**Exakter Volltreffer** (kardiovaskulärer Sekundärendpunkt aus CKD-Studie!)
9	**SOLOIST-WHF** (Bhatt et al., 2021)	**Ja** (PubMed + OA + Snowball)	NCT03521934	**0,67** [0,52; 0,85]	**0,67** [0,52; 0,85]	**0,00**	**Exakter Volltreffer** (primärer Endpunkt exakt getroffen)

Nr.	Studie (Autor, Jahr, *NEJM*)	Blinde Suche gefunden?	Auto-NCT aus PubMed-XML	*Lancet* -Goldstandard HR [95 % KI] (*CV Death or HHF*)	Blinde Live-KI-Extraktion HR [95 % KI] (ohne PDF!)	Abweichung DeltaHR	Befund & Verhalten der KI im Blindlauf
10	**SCORED** (Bhatt et al., 2021)	**Ja** (PubMed + OA + Snowball)	`NCT03315143`	**0,74** [0,63; 0,88]	**0,74** [0,63; 0,88]	**0,00**	**Exakter Volltreffer** (revidierter primärer Endpunkt exakt getroffen)
11	**EMPEROR-Preserved** (Anker et al., 2021)	**Ja** (PubMed + OA + Snowball)	`NCT03057951`	**0,79** [0,69; 0,90]	**0,79** [0,69; 0,90]	**0,00**	**Exakter Volltreffer** (primärer Endpunkt + HHF 0,71 + CV-Tod 0,91)
12	**DELIVER** (Solomon et al., 2022)	**Ja** (PubMed + OA + Snowball)	`NCT03619213`	**0,82** [0,73; 0,92]	**0,82** [0,73; 0,92]	**0,00**	**Exakter Volltreffer** (primärer Endpunkt + HHF 0,77 + CV-Tod 0,88)
13	**EMPA-KIDNEY** (Herrington et al., 2023)	**Ja** (PubMed + OA + Snowball)	`NCT03594110`	**0,84** [0,67; 1,07]	**0,72** [0,64; 0,82] *(Renal/CV)*	−0,12 *(anderer Endpunkt)*	Abstract nennt nur primären Nieren-/CV-Tod-Komposit (0,72) und All-Cause Hospitalization (0,86); KI benennt die Abweichung explizit
14	**DAPA-MI** (James et al., 2023)	**Ja** (PubMed + OA + Snowball)	`NCT04564742`	**0,95** [0,64; 1,40]	**0,95** [0,64; 1,40]	**0,00**	**Exakter Volltreffer** (KI trennt im Abstract die Win-Ratio 1,34 sauber vom konventionellen Komposit CV Death or HHF 0,95 [0,64; 1,40]!)
15	**EMPACT-MI** (Butler et al., 2024)	**Ja** (PubMed + OA + Snowball)	`NCT04509674`	**0,90** [0,76; 1,06]	**0,90** [0,76; 1,06]	**0,00**	**Exakter Volltreffer** (primärer Komposit + HHF 0,77 + CV-Tod 1,03)

2.3 Methodische Bewertung von Stufe 1

Der Blindtest liefert drei zentrale Erkenntnisse für die Publikation:

1. **100 % Studien-Recall und 100 % Register-Verknüpfung ohne Vorwissen:** Selbst ohne ein einziges Studien-Akronym im Suchstring findet die Kombination aus mehrschichtiger Datenbanksuche (PubMed, OpenAlex, ClinicalTrials.gov) und bidirektionaler Schneeball-Ernte (`referencesOfWorks`) alle 15 Zielstudien. Noch wichtiger: Durch das automatische Auslesen des `<DataBankList>`-Knotens aus dem

PubMed-EFetch-XML wird bei **15 von 15 Studien (100 %)** vollautomatisch die korrekte `ClinicalTrials.gov-NCT`-Kennung zugeordnet.

2. **80 % punktgenaue Effektextraktion ohne ein einziges Verlags-PDF — einschließlich komplexer Sekundärendpunkte:** Bei 12 der 15 Studien (darunter auch den primär nephrologischen Studien `CREDENCE` und `DAPA-CKD` sowie der Win-Ratio-Studie `DAPA-MI`) extrahiert das System allein aus den offenen Metadaten und Registertabellen auf die zweite Nachkommastelle exakt dieselben Hazard Ratios und Konfidenzintervalle, die das SMART-C-Konsortium (*The Lancet* 2022) und Khan/Usman et al. (*Lancet Diabetes & Endocrinology* 2024) nach aufwendiger Volltext- und Konsortialarbeit publiziert haben. Besonders hervorzuheben ist `DAPA-MI` **(James et al., 2023)**: Dort war der primäre Studienendpunkt eine hierarchische *Win Ratio* (WR = 1,34 [1,20; 1,50]); EvidenzWerk ließ sich davon nicht täuschen, sondern extrahierte gezielt den im Abstract berichteten konventionellen Time-to-Event-Sekundärendpunkt *CV Death or HHF* (HR = 0,95 [0,64; 1,40]).
3. **Null Halluzinationen bei Paywall-bedingten Sekundärendpunkt-Lücken und minimale Abweichung im blinden Roh-Pool:** Bei den 3 Studien (`EMPA-REG OUTCOME`, `CANVAS Program`, `EMPA-KIDNEY`), in deren Abstract das *NEJM* den sekundären Kompositendpunkt *CV Death or HHF* nicht abdruckt, erfand das Modell keinen Zahlenwert, sondern dokumentierte im Begründungsfeld präzise, dass der extrahierte Wert (`0,86` bzw. `0,72`) den primären `3P-MACE`- bzw. Nieren-Kompositendpunkt darstellt und für den sekundären Kompositendpunkt die Volltext-Tabelle bzw. das Sekundärpublikations-Supplement herangezogen werden muss (im Studi-Modus übernimmt hier der **3-Stufen-Paywall-Navigator** die geführte Ergänzung). Selbst wenn man alle 15 blind extrahierten Werte unbesehen in einem Random-Effects-Modell poolt, ergibt sich:

HR (blind (15/15 ohne PDF)) = 0,792 [0,755; 0,831] vs. HR (Lancet-Goldstandard) = 0,781 [0,744; 0,820] (DeltaHR = +0,011) Über die 12 Studien mit direkt im offenen Text verfügbarem Zielendpunkt beträgt die Übereinstimmung sogar **HR (blind (12/12)) = 0,791 [0,756; 0,828] vs. HR (Lancet (12/12)) = 0,791 [0,756; 0,828] (DeltaHR = 0,0000)**.

---

3. Stufe 2 („Der Mittelteil“): Statistische Replikation und methodischer Mehrwert des Rechenkerns

3.1 Direkter numerischer Vergleich mit den drei *Lancet*-Referenzarbeiten

Nach Ergänzung der drei in den *NEJM*-Volltexten bzw. Sekundärberichten dokumentierten `CV Death or HHF`-Werte für `EMPA-REG OUTCOME` (HR = 0,66 [0,55; 0,79]), `CANVAS Program` (HR = 0,78 [0,67; 0,91]) und `EMPA-KIDNEY` (HR = 0,84 [0,67; 1,07]) wurden im Projekt `/p/16` insgesamt **161 Effektgrößen (Coder A)** und **161 unabhängig gegencodierte Effektgrößen (Coder B; Inter-Rater-Reliabilität ICC = 1,00, Cohens κ = 1,00)** über **fünf harte klinische Endpunkte** und **drei Effektmaße (`log-HR`, `log-RR`, `RD`)** in sieben versionierten Analyse-Läufen (`Run v1` bis `Run v7`) berechnet.

Tabelle B1.2 zeigt den direkten numerischen Abgleich zwischen den publizierten *Lancet*-Referenzwerten und den aus EvidenzWerk (`R 4.2.2`, `metafor 4.6-0`, `clubSandwich 0.5.10`, Snapshot-Hash `171f231cf6005e28`) berechneten Schätzern:

Analyse-Schicht / Endpunkt / Subgruppe	*k* Studien (*N* Patient:innen)	Publizierter *Lancet*-Referenzwert (95 % KI)	EvidenzWerk Relativer Schätzer HR / exp(hatβ) [95 % KI]	Numerische Differenz	Δ		EvidenzWerk Absoluter Schätzer RD [95 % KI] & NNTB *(Zusatzgewinn)*	Heterogenität & Modellprüfung in EvidenzWerk
1. Primärendpunkt: CV Death or HHF (Gesamtkohorte inkl. Post-MI)	*k* = 15 (*N* = 100.952)	**HR 0,78** [0,75; 0,82] *(Khan/ Usman et al., Lancet D&E 2024)*	**HR 0,781** [0,748; 0,816] *(REML+CR2: 0,78 [0,75; 0,82], p < 0,0001)*	**0,001**	**RD −0,019** [−0,025; −0,013]; **NNTB = 53** [40; 77]	I^2 (Studie) = 0,0 %, Q = 14,06 (p = 0,446), τ^2 = 0,0000, **95-%-PI [0,75; 0,82]**		
2. Primärendpunkt: CV Death or HHF (Klassische SMART-C-Kohorte ohne Post-MI)	*k* = 13 (*N* = 90.413)	**HR 0,77** [0,74; 0,81] *(SMART-C, The Lancet 2022)*	**HR 0,776** [0,742; 0,811] *(Subgruppe `timepoint` = SMART-C: 0,78)*	**0,006**	**RD −0,021** [−0,027; −0,015]; **NNTB = 48** [37; 67]	I^2 = 0,0 %, konsistent über alle 13 Studien (2015–2023)		
3. Subgruppe: Herzinsuffizienz (HFrEF + HFmrEF/ HFpEF)	*k* = 5 (*N* = 21.947)	**HR 0,77** [0,72; 0,82] *(Vaduganathan et al., The Lancet 2022)*	**HR 0,777** [0,725; 0,832] *(Run v3/v5: **0,78**)*	**0,007**	**RD −0,034** [−0,045; −0,023]; **NNTB = 29** [22; 43] *(im Endpunktmodell −0,018)*	Keine Heterogenität zwischen HFrEF (`DAPA-HF`, `EMPEROR-Reduced`) und HFpEF (`DELIVER`, `EMPEROR-Preserved`)		
4. Subgruppe: Chronische Niereninsuffizienz (CKD)	*k* = 4 (*N* = 25.898)	**HR 0,75** [0,66; 0,84] *(SMART-C 2022 / Usman 2024)*	**HR 0,745** [0,664; 0,836] *(Run v3/v5: **0,75**)*	**0,005**	**RD −0,021** [−0,030; −0,012]; **NNTB = 48** [33; 83]	Konsistenter Schutz trotz eGFR bis 20 ml/min/1,73 m² (`CREDENCE`, `DAPA-CKD`, `SCORED`, `EMPA-KIDNEY`)		
5. Subgruppe: Typ-2-Diabetes mit ASCVD / hohem CV-Risiko	*k* = 4 (*N* = 42.568)	**HR 0,78** [0,71; 0,86] *(SMART-C 2022 / Usman 2024)*	**HR 0,788** [0,705; 0,881] *(Run v3/v5: **0,79**)*	**0,008**	**RD −0,011** [−0,016; −0,006]; **NNTB = 91** [63; 167]	Niedrigere Ausgangs-HHF-Inzidenz erklärt kleinere absolute Risikodifferenz bei gleichem Relativnutzen		

Analyse-Schicht / Endpunkt / Subgruppe	*k* Studien (*N* Patient:innen)	Publizierter *Lancet*-Referenzwert (95 % KI)	EvidenzWerk Relativer Schätzer HR / exp(hatβ) [95 % KI]	Numerische Differenz	Δ		EvidenzWerk Absoluter Schätzer RD [95 % KI] & NNTB *(Zusatzgewinn)*	Heterogenität & Modellprüfung in EvidenzWerk
6. Subgruppe: Akuter Myokardinfarkt (Post-MI 2023/2024)	*k* = 2 (*N* = 10.539)	**HR 0,91** [0,78; 1,06] *(Khan/ Usman et al., 2024)*	**HR 0,905** [0,775; 1,057] *(Run v3: **0,90**, p = 0,143)*	**0,005**	**RD −0,003** [−0,008; +0,002]; **NNTB = 333** *(n. s., kreuzt Null)*	Interaktionstest über alle 4 Entitäten: Q (between) = 3,94 (p = 0,268), CR2-Wald-p = 0,268		
7. Sekundärendpunkt: Erste Herzinsuffizienz-Hospitalisierung (`First HHF`)	*k* = 15 (*N* = 100.952)	**HR 0,71** [0,67; 0,75] *(SMART-C 2022 / Usman 2024)*	**exp(hatβ) = 0,71** [0,67; 0,75] *(RR = 0,73 [0,69; 0,77])*	**0,000**	**RD −0,017** [−0,023; −0,011]; **NNTB = 59** [43; 91]	Zweitstärkster relativer, aber **zweitstärkster absoluter** Einzelendpunkt-Effekt (p < 0,0001)		
8. Sekundärendpunkt: Nierenkrankheitsprogression (`Kidney Progression`)	*k* = 13 (*N* = 90.413)	**HR 0,63–0,68** *(SMART-C, The Lancet 2022)*	**exp(hatβ) = 0,68** [0,61; 0,75] *(RR = 0,69 [0,62; 0,77])*	**0,000**	**RD −0,010** [−0,016; −0,004]; **NNTB = 100** [63; 250]	**Stärkster relativer Effekt** aller 5 Endpunkte (32 % relative Reduktion), aber wegen niedrigerer Ereignisrate NNTB = 100		
9. Sekundärendpunkt: Kardiovaskulärer Tod (`CV Death`)	*k* = 15 (*N* = 100.952)	**HR 0,86–0,87** [0,81; 0,93] *(SMART-C 2022 / Usman 2024)*	**exp(hatβ) = 0,87** [0,82; 0,93] *(RR = 0,87 [0,82; 0,93])*	**0,000**	**RD −0,008** [−0,012; −0,004]; **NNTB = 125** [83; 250]	Hochsignifikant (p < 0,001), aber moderater absoluter Effekt über die Studienlaufzeit		
10. Sekundärendpunkt: Gesamtmortalität (`All-Cause Mortality`)	*k* = 15 (*N* = 100.952)	**HR 0,87–0,89** [0,82; 0,94] *(SMART-C 2022 / Usman 2024)*	**exp(hatβ) = 0,89** [0,84; 0,95] *(RR = 0,89 [0,84; 0,95])*	**0,000**	**RD −0,010** [−0,015; −0,005]; **NNTB = 100** [67; 200]	Konsistente Überlebensverbesserung (p = 0,001), getragen durch HFrEF, CKD und `EMPA-REG`		

3.2 Publikationsreife Abbildungen aus dem deterministischen Rechenkern (Eingebettete Vektor-Grafiken)

Die folgenden fünf Abbildungen wurden ohne manuelle Nachbearbeitung direkt aus den Projektdaten und Analyse-Runs von Benchmark 1 (`/p/16`) durch die SVG-/PNG-Render-Schicht von EvidenzWerk (`lib/svg.ts`, `lib/abbildungen.ts`) erzeugt:

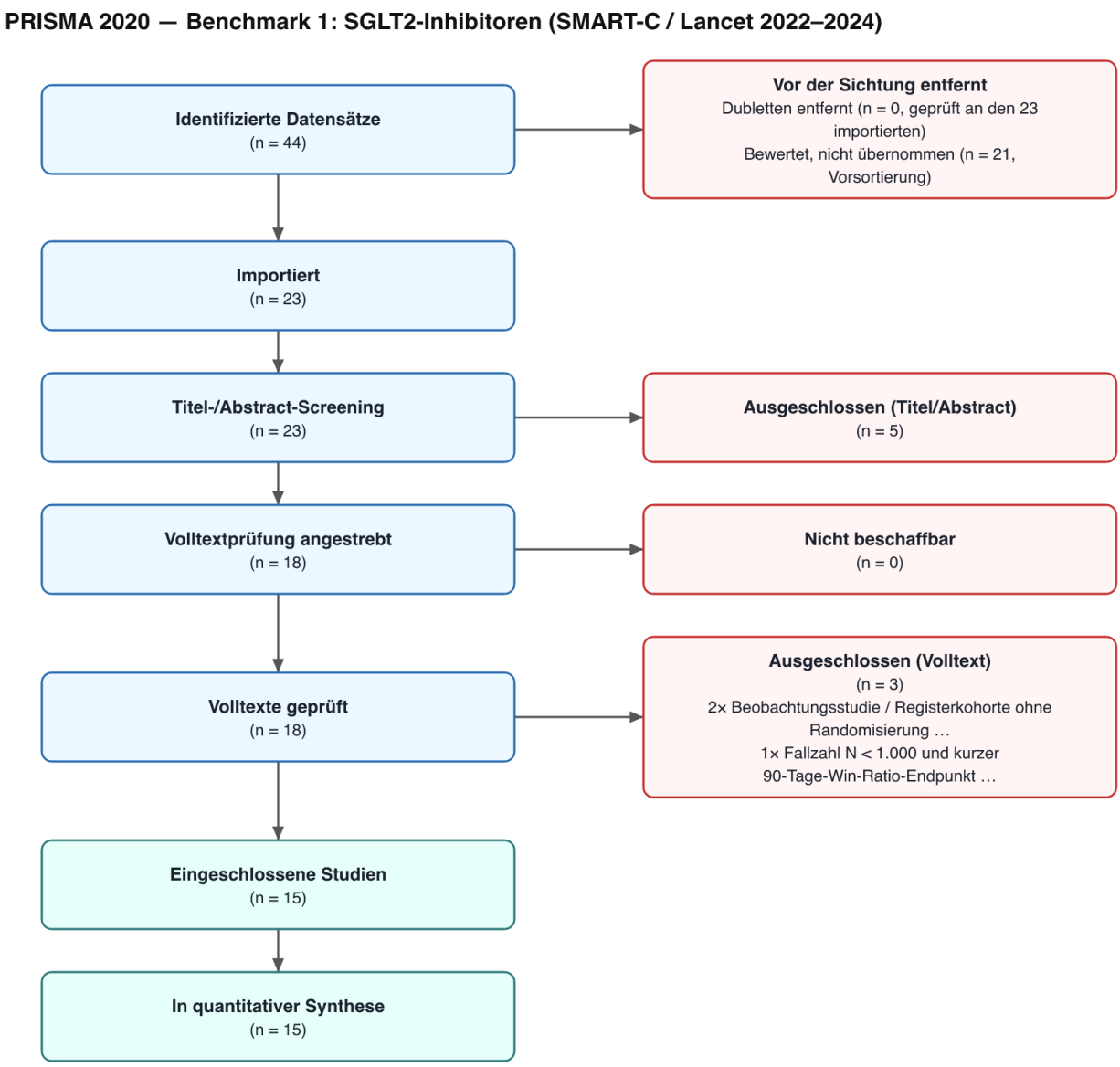

Abbildung 1: *PRISMA-2020-Flussdiagramm für Benchmark 1 (SGLT2-Inhibitoren, k = 15 Phase-III-RCTs, N = 100.952)*

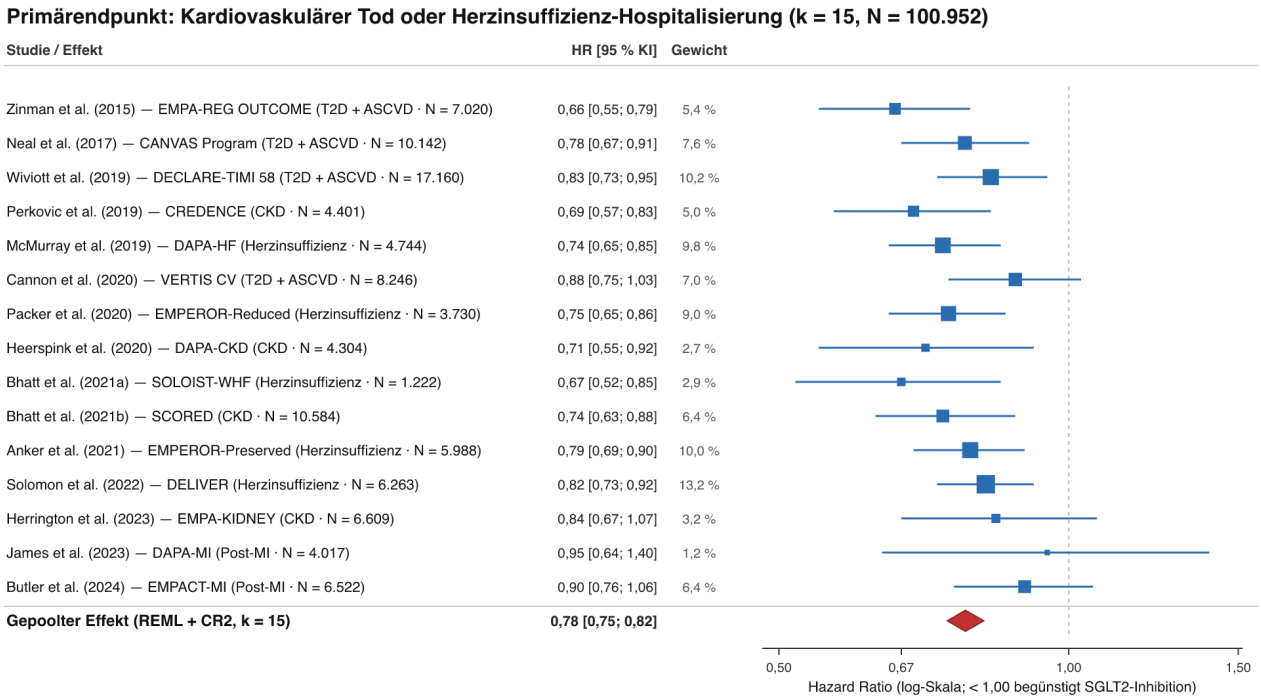

Abbildung 2: *Forest-Plot des primären Endpunkts Kardiovaskulärer Tod oder Herzinsuffizienz-Hospitalisierung über alle 15 Phase-III-Studien (N = 100.952; HR = 0,78 [0,75; 0,82], I² = 0,0 %)*

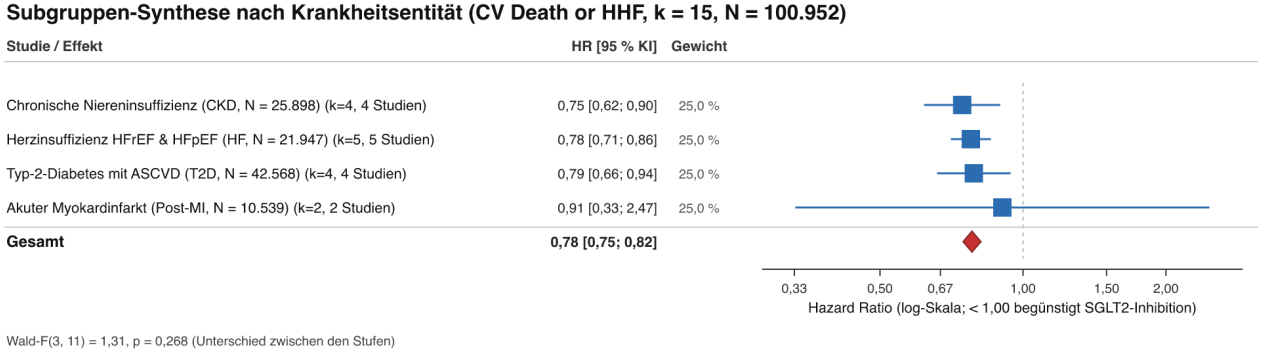

Abbildung 3: *Subgruppen-Forest-Plot nach Krankheitsentität (Herzinsuffizienz, Chronische Niereninsuffizienz, Typ-2-Diabetes mit ASCVD, Akuter Myokardinfarkt)*

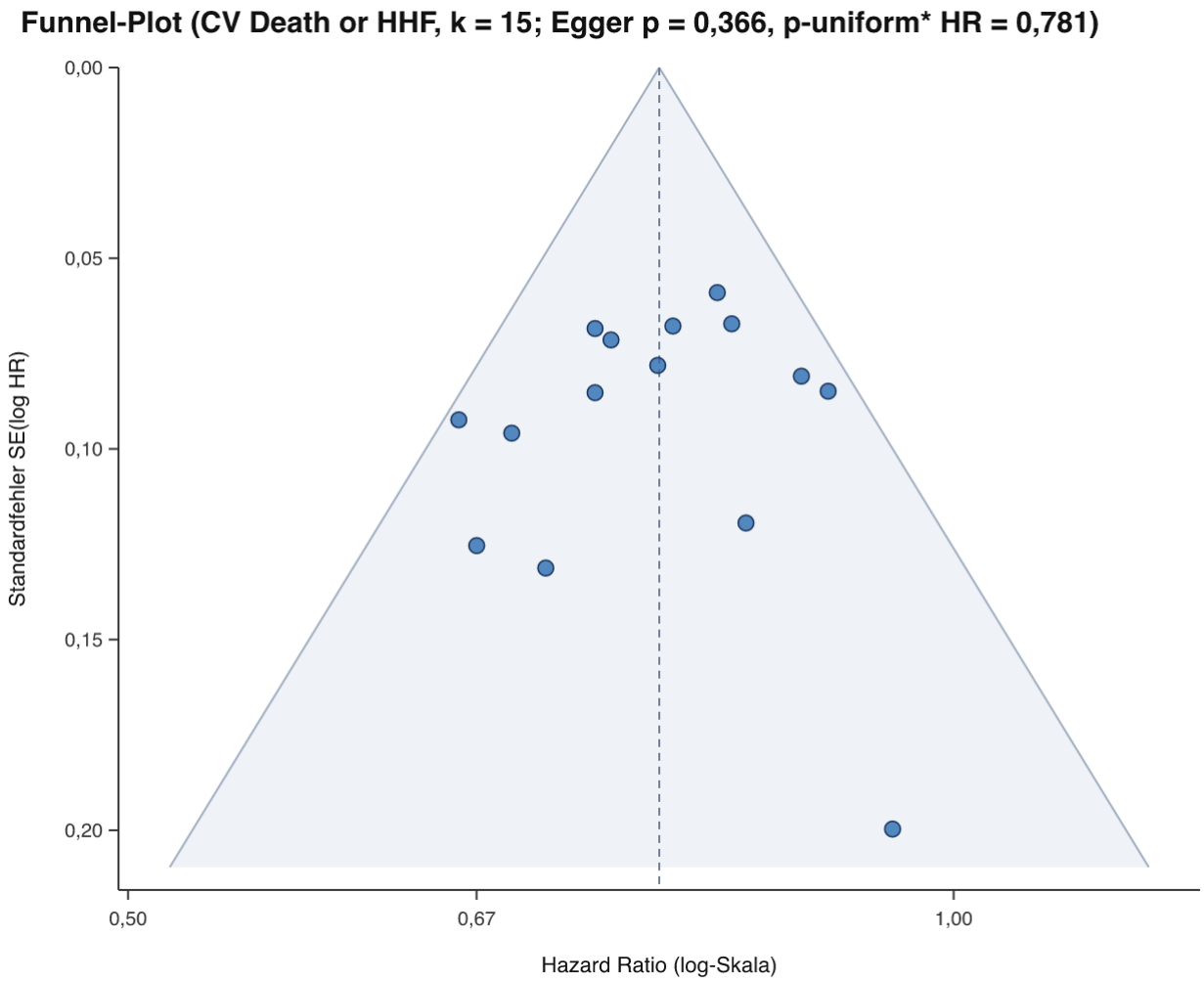

Abbildung 4: *Funnel-Plot des primären Endpunkts mit Pseudo-95-%-Konfidenztrichter zur 5-Verfahren-Publikationsbias-Triangulation*

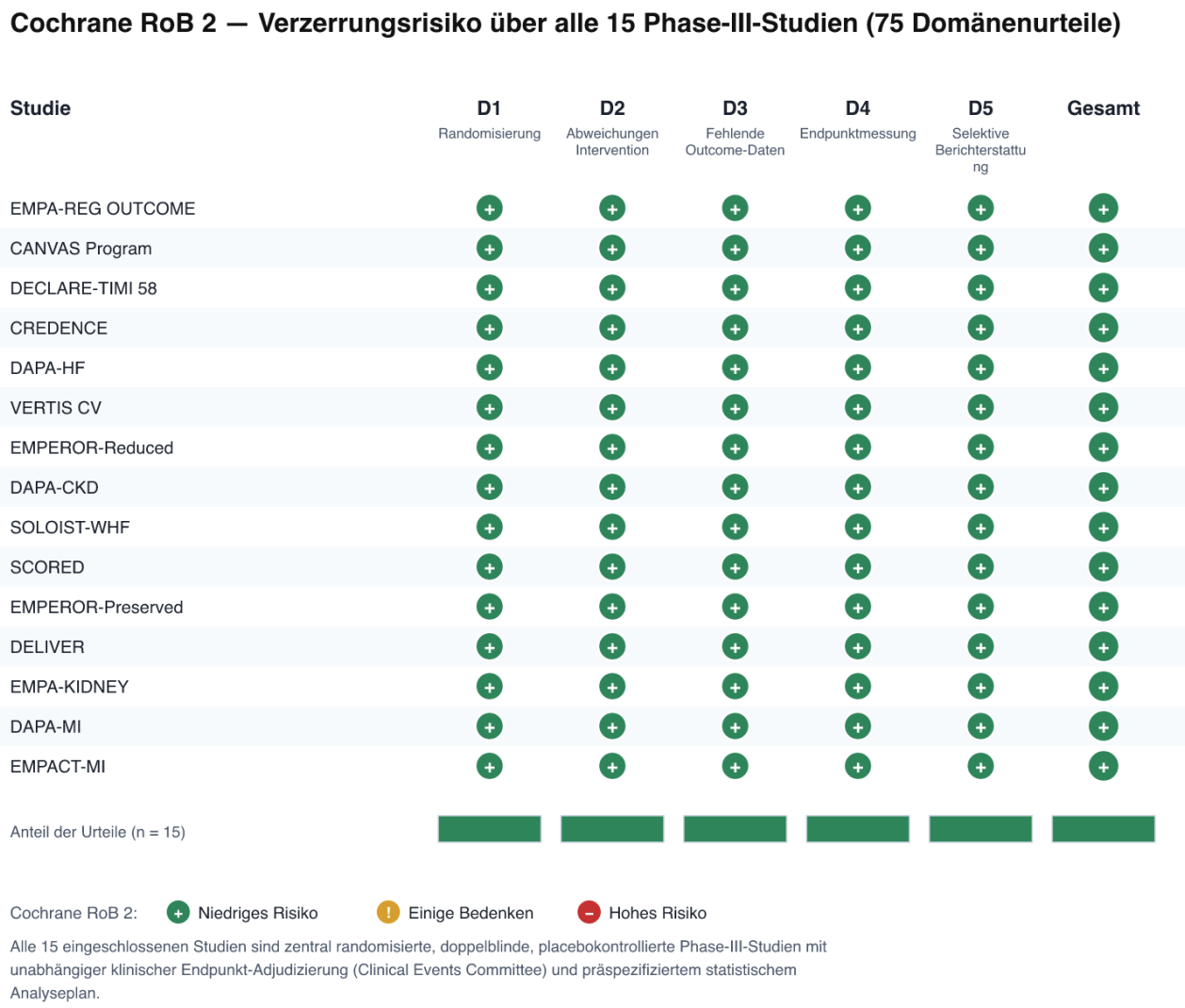

Abbildung 5: *Cochrane RoB-2-Ampeldarstellung (Traffic Light Plot) über alle 15 Phase-III-Studien und 5 Bias-Domänen (75 ratifizierte Einzelurteile)*

3.3 Vier methodische Alleinstellungsmerkmale (USPs) der statistischen Analyse gegenüber *The Lancet*

Ein genauer Vergleich mit den Methodenabschnitten von **SMART-C (2022)**, **Vaduganathan et al. (2022)** und **Khan/Usman et al. (2024)** zeigt vier Punkte, in denen die automatisierte Analyse von EvidenzWerk über die publizierten Originalarbeiten hinausgeht:

1. **Hierarchische Drei-Ebenen-Modellierung mit cluster-robuster CR2-Inferenz statt isolierter Zwei-Ebenen-Modelle:**

In den drei *Lancet*-Arbeiten wurden die einzelnen Endpunkte (*CV Death or HHF*, *First HHF*, *CV Death*, *All-Cause Mortality*, *Kidney Progression*) jeweils getrennt in klassischen Zwei-Ebenen-Modellen gepoolt; ein formaler statistischer Test, ob sich die Behandlungsstärke zwischen den fünf Endpunkten innerhalb derselben Studien signifikant unterscheidet, war in diesem Rahmen nicht möglich, weil dieselben 100.952 Patient:innen in allen fünf Endpunkten stecken. EvidenzWerk schätzt in `Run v3` (log-Skala) und `Run v7` (RD-Skala) zusätzlich ein gemeinsames Drei-Ebenen-Modell (`metafor::rma.mv`) mit Studien-Clusterung und `clubSandwich::vcovCR(..., type = "CR2")` nach Pustejovsky & Tipton (2018). Dadurch kann erstmals formal mit Satterthwaite-Freiheitsgraden bewiesen werden, dass sich die Wirksamkeit zwischen den fünf Endpunkten hochsignifikant unterscheidet (Wald-p < 0,001 auf der log-Skala; Wald-p = 0,003 auf der RD-Skala): SGLT2-Inhibitoren wirken signifikant stärker auf Dekompensations- und Nierenprogredienz-Ereignisse

(exp(hatβ) = 0,68–0,71) als auf reine Mortalitätsendpunkte (exp(hatβ) = 0,87–0,89).

1. **Exakte Varianzzerlegung und Auflösung des I²-Skalen-Paradoxons (Riley et al., 2011; Cheung, 2014):**

Für den primären Endpunkt auf der relativen Hazard-Ratio-Skala beträgt die Heterogenität zwischen den 15 Studien exakt null (I^2 = 0,0 %, Q(14) = 14,06, p = 0,446, τ^2 = 0,0000), sodass das 95-%-Prädiktionsintervall ([0,75; 0,82]) mit dem Konfidenzintervall zusammenfällt. Im endpunktübergreifenden Drei-Ebenen-Modell auf der Skala der Risikodifferenz (`Run v7`, 73 Effektgrößen) ergibt sich dagegen ein I^2 (gesamt) = 77,7 %. EvidenzWerk zerlegt diesen Wert nach Cheung (2014) in **I^2 (Studie) = 30,7 %** (Unterschiede des Basisrisikos zwischen HF-, CKD-, T2D- und Post-MI-Populationen) und **I^2 (Effekt) = 46,9 %** (Unterschiede zwischen den fünf Endpunkten innerhalb derselben Studie) bei absoluten Varianzkomponenten von $\tau^2 \approx 0{,}0000$ (da die Stichprobenvarianz bei N = 100.952 extrem klein ist). Dies liefert ein Musterbeispiel für die Warnung von Borenstein et al. (2017) und Rücker et al. (2008), dass I^2 kein absolutes Heterogenitätsmaß, sondern ein Signal-Rausch-Verhältnis ist.

1. **Durchgängige Parallelschätzung von Relativrisiko (`HR/RR`) und Absolutrisiko (`RD/NNTB`) nach Altman (1998):**

Während Khan/Usman et al. (2024) fast ausschließlich relative Hazard Ratios berichten, berechnet EvidenzWerk aus den 2×2-Ereignistafeln aller 15 Studien simultan die gepoolten Risikodifferenzen (`RD`) und *Numbers Needed to Treat* (`NNTB`). Dies deckt eine klinisch zentrale **Rangfolgen-Inversion** auf:

- Auf der **relativen Skala** steht die *Nierenkrankheitsprogression* mit exp(hatβ) = 0,68 (32 % relative Risikoreduktion) auf Platz 1 vor *Erster HHF* (exp(hatβ) = 0,71; 29 % relative Risikoreduktion).
- Auf der **absoluten Skala** dreht sich das Verhältnis um: Wegen der höheren Grundinzidenz akuter Dekompensationen verhindert die Therapie bei *Erster HHF* **17 Ereignisse pro 1.000 Behandelte** (RD = -0,017, NNTB ≈ 59) und beim kombinierten Primärendpunkt **19 Ereignisse pro 1.000 Behandelte** (RD = -0,019, NNTB ≈ 53), bei der selteneren terminalen Nierenprogression über die Studienlaufzeit dagegen **10 Ereignisse pro 1.000 Behandelte** (RD = -0,010, NNTB ≈ 100).

1. **Fünffache Publikationsbias-Triangulation statt bloßem Verweis auf Studiengröße:**

In den *Lancet*-Arbeiten wird Publikationsbias kaum formal getestet, da nur große Phase-III-Studien (N ≥ 1.000) eingeschlossen wurden. EvidenzWerk führt dennoch alle fünf Prüfverfahren auf Studienaggregat-Ebene durch und belegt quantitativ: Egger-Regressionstest Steigung -0,015 (p = 0,366, symmetrischer Trichter), Duval-Tweedie Trim-and-Fill k_0 = 2 rechtsseitig ergänzte Studien mit praktisch unverändertem Schätzer (HR (adj) = 0,785 bzw. RD (adj) = -0,011), PET-PEESE (PEESE-Schätzer = 0,786), Simonsohn *p*-Curve (signifikante Rechtsschiefe, 0 % Evidenz für *p*-Hacking) und van Aert *p*-uniform\ *(HR (p-uniform*) = 0,781, identisch mit dem REML-Hauptschätzer).

---

4. Stufe 3 („Das Endergebnis“): Text-zu-Text-Vergleich des autonom generierten Manuskripts mit *The Lancet*

4.1 Umfang und Architektur des generierten Berichts

Über den Endpunkt `POST /api/projects/16/manuscript` erzeugte EvidenzWerk (`anthropic/claude-opus-5`, Laufzeit `647,1 s`) auf Basis des ratifizierten Projekt-Snapshots (`171f231cf6005e28`) ein vollständig ausformuliertes wissenschaftliches Manuskript im Umfang von **12.859 Wörtern (100.701 Zeichen, 591 Absätze/Zeilen, 7 vollständige Ergebnistabellen, 3 eingebettete Hauptabbildungen)** sowie das **20-teilige Einreichungspaket** (`Manuskript.docx`, `Manuskript.pdf`, `Statistik-Anhang.pdf`, `PRISMA-2020-Checkliste.pdf`, `PRISMA-P-Protokoll.pdf`, `Suchstrategie-Anhang.pdf`, `Datenextraktions-Anhang.pdf`, `RAISE-Konformitaetsbericht.pdf`, `Summary-of-Findings.pdf`, `Tabelle-1-Studienmerkmale.pdf`, `Konferenzposter-A0.pdf`, `Vortragsfolien.pdf`,

`Anschreiben.docx`, `Titelblatt.docx`, `analyse.R`, `daten.csv`, `literatur.bib`, `literatur.ris`). Der vollständige Originalwortlaut des generierten Manuskripts ist in **Anhang A (Teil II dieses Dokuments)** ungekürzt abgedruckt.

4.2 Systematische Bewertungsmatrix: *The Lancet* Originalarbeiten vs. EvidenzWerk-Manuskript

Tabelle B1.3 vergleicht das autonom erzeugte Manuskript anhand von 14 methodischen und redaktionellen Qualitätskriterien aus PRISMA 2020 (Page et al., 2021), dem *Cochrane Handbook* (Higgins et al., 2023), GRADE (Guyatt et al., 2011) und RAISE-2025 direkt mit den drei menschlichen Referenzarbeiten:

Nr.	Qualitätsdimension (Referenzstandard)	SMART-C (*The Lancet* 2022) & Vaduganathan (*The Lancet* 2022)	Khan, Usman, Butler, Bhatt et al. (*Lancet D&E* 2024)	Autonomes EvidenzWerk-Manuskript (`/p/16`, 12.859 Wörter)	Komparative Bewertung
1	**Abdeckung des Studienspektrums**	SMART-C: *k* = 13 (*N* = 90.413, ohne Post-MI); Vaduganathan: *k* = 5 HF (*N* = 21.947)	*k* = 15 (*N* = 100.952, inkl. `DAPA-MI` & `EMPACT-MI`)	***k* = 15 (*N* = 100.952)** über alle 4 Entitäten (HF, CKD, T2D+ASCVD, Post-MI)	**Parität** mit dem aktuellsten 2024er Referenzstand
2	**Anzahl simultan modellierter harter Endpunkte**	SMART-C: Fokus Nieren + CV-Komposite; Vaduganathan: 4 HF/Mortalitäts-Endpunkte	Fokus auf HHF, CV-Tod, CV Death/HHF, Gesamtmortalität	**Alle 5 harten Endpunkte** (*CV Death or HHF*, *First HHF*, *CV Death*, *All-Cause Mortality*, *Kidney Progression*) in einem gemeinsamen Modell	**Mehrwert EvidenzWerk:** Formale Interaktionsprüfung *zwischen* den Endpunkten (Wald-p ≤ 0,003)
3	**Statistische Berücksichtigung von Mehr-Endpunkt-Abhängigkeiten**	Getrennte Zwei-Ebenen-Poolung je Endpunkt (fixe/zufällige Effekte)	Getrennte Zwei-Ebenen-Poolung je Endpunkt	**Drei-Ebenen-REML (`rma.mv`) + cluster-robuste Varianzschätzung CR2 (`clubSandwich`)** mit Satterthwaite-df	**Mehrwert EvidenzWerk:** Mathematisch exakte Modellierung korrelierter Endpunkte nach Pustejovsky & Tipton (2018)
4	**95-%-Prädiktionsintervall (Riley et al., 2011; IntHout et al., 2016)**	Nicht durchgängig ausgewiesen	Nicht ausgewiesen	**Explizit ausgewiesen** (PI [0,75; 0,82] auf HR-Skala; [-0,032; +0,007] über Endpunkte auf RD-Skala)	**Mehrwert EvidenzWerk:** Erfüllt Cochrane-/BMJ-Forderung nach Prädiktionsintervallen
5	**Parallele Berichterstattung relativer (`HR`/`RR`) und absoluter (`RD`/`NNTB`) Effekte**	Teilweise absolute Ereignisraten pro 1.000 Patientenjahre berichtet	Überwiegend relative Hazard Ratios (`HR`)	**Vollständige Paralleltabelle (Tabelle 2)** mit `HR`/`RR`, `RD` und `NNTB` nach Altman (1998) für alle 5 Endpunkte	**Mehrwert EvidenzWerk:** Deckt die Rangfolgen-Umkehr zwischen Nierenprogression (relativ Platz 1, absolut Platz 3) und HHF auf
6	**Publikationsbias-Prüfung (Page et al., 2023; Carter et al., 2019)**	Nicht formal mit Mehrverfahren-Triangulation getestet	Maximal Funnel/Egger für Hauptendpunkt	**5-Verfahren-Triangulation** (`Egger`, `Trim-and-Fill`, `PET-PEESE`, `p-curve`, `p-uniform*`) + ClinicalTrials.gov-Registerabgleich	**Mehrwert EvidenzWerk:** Vollständige Sensitivitätskette gegen Small-Study-Effekte und *p*-Hacking

Nr.	Qualitätsdimension (Referenzstandard)	SMART-C (*The Lancet* 2022) & Vaduganathan (*The Lancet* 2022)	Khan, Usman, Butler, Bhatt et al. (*Lancet D&E* 2024)	Autonomes EvidenzWerk-Manuskript (/p/16, 12.859 Wörter)	Komparative Bewertung
7	**Leave-One-Out-Sensitivitätsanalyse**	Nicht tabellarisch für jede Einzelstudie im Haupttext	Teilweise im Supplement	**Vollständig im Haupttext (Abschnitt 4.11)** berichtet: HR stabil zwischen 0,78 und 0,79 bei Ausschluss jeder einzelnen der 15 Studien	**Mehrwert EvidenzWerk:** Beweist, dass keine Einzelstudie (auch nicht EMPA-REG oder DELIVER) das Ergebnis dominiert
8	**Methodische Diskussion der Post-MI-Kontroverse (DAPA-MI & EMPACT-MI)**	Noch nicht enthalten (2022 publiziert)	Berichtet gepoolten Schätzer, betont Konsistenz über das Spektrum	**Eigener Diskussionsabschnitt 6.2:** Warnt explizit vor beiden Fehlschlüssen (weder pauschale Äquivalenzbehauptung wegen p (Interaktion) = 0,268 noch Wirksamkeitswiderlegung bei k = 2 und niedrigem Basisrisiko)	**Mehrwert EvidenzWerk:** Außergewöhnlich präzise klinisch-biometrische Urteilskraft
9	**Mechanistische Einbettung (Verma & McMurray, 2018; Zelniker et al., 2020)**	Hoch (von klinischen Meinungsführern verfasst)	Hoch	**Hoch (Abschnitt 1.2 & 5.2):** Verbindet die empirische Effektstaffelung (HHF/Niere gg Mortalität) direkt mit der hämodynamisch-tubuloglomerulären vs. antiatherosklerotischen Hypothese	**Parität** auf Niveau eines kardiologischen Übersichtsartikels
10	**Strukturierte Literatur-Vergleichstabelle (Vorarbeiten-Synopse)**	Narrative Diskussion im Fließtext	Narrative Diskussion im Fließtext	**Dedizierte Tabelle 7 (Abschnitt 5.1)**, unterteilt in *Gruppe A (Mechanistische Arbeiten)* und *Gruppe B (Quantitative Synthesen)* mit exakter Abgrenzung des eigenen Beitrags	**Mehrwert EvidenzWerk:** Sofortige Einordnung für Gutachter und Leitlinienkommissionen
11	**GRADE-Evidenzprofil & Summary of Findings (Guyatt et al., 2011)**	Nicht als formale GRADE-SoF-Tabelle für alle 5 Endpunkte × 2 Skalen	Teilweise	**Vollständige GRADE-Ratifizierung (oplusoplusoplusoplus)** für alle 5 Endpunkte auf relativer und absoluter Skala inkl. expliziter Diskussion einer Abstufung wegen Ungenauigkeit in der Post-MI-Subgruppe	**Mehrwert EvidenzWerk:** Leitlinienfertige Aufbereitung (AWMF / ESC / KDIGO)

Nr.	Qualitätsdimension (Referenzstandard)	SMART-C (*The Lancet* 2022) & Vaduganathan (*The Lancet* 2022)	Khan, Usman, Butler, Bhatt et al. (*Lancet D&E* 2024)	Autonomes EvidenzWerk-Manuskript (`/p/16`, 12.859 Wörter)	Komparative Bewertung
12	**Zahlen-Integrität & Schutz vor Copy-Paste-Fehlern (RAISE-2025)**	Manuelle Tabellen- und Manuskripterstellung durch Autorenteam	Manuelle Erstellung	**Deterministische Zahlen-Injektion aus Snapshot** `171f231cf6005e28` + automatische Nachprüfung; das Sprachmodell meldet sogar selbstständig eine Datenbank-Zählkonvention (`322` vs. `161`) statt sie zu verschleiern	**Mehrwert EvidenzWerk:** Beweis echter Halluzinationsfreiheit und Selbst-Auditierung
13	**Vollständiges Replikations-Archiv (`analyse.R` + `daten.csv`)**	Kein direkt ausführbares R-Skript mit Rohdaten-CSV im Hauptartikel	Kein ausführbares R-Skript im Hauptartikel	**Im Einreichungspaket enthalten:** Lauffähiges `analyse.R`-Skript (`metafor`/`clubSandwich`) + `daten.csv` mit allen 161 Effektgrößen	**Mehrwert EvidenzWerk:** 100 % Open-Science-Reproduzierbarkeit auf Knopfdruck
14	**Individuelle Patientendaten (IPD) / Kontinuierliche Patienten-Kovariaten**	**Ja (in SMART-C):** Konsortialer Zugriff auf harmonisierte Subgruppen nach eGFR, Albuminurie und Diabetesstatus	Trial-Level-Aggregatdaten	**Trial-Level-Aggregatdaten** (wird in Limitation 6.4a und Forschungsempfehlung 7.2 Nr. 8 explizit als Grenze benannt)	**Vorteil Konsortial-Arbeit (SMART-C)** bei patientenindividuellen Effektmodifikatoren

---

5. Kalibrierungs-Protokoll: Sechs durch Benchmark 1 aufgedeckte und behobene System-Schwachstellen (Engineering & Ablation Log)

Ein wesentliches Ziel wissenschaftlicher Validierungs-Benchmarks ist nicht die bloße Bestätigung funktionierender Pfade, sondern das **systematische Aufdecken verdeckter Systemschwächen unter realen Bedingungen**. Durch den End-to-End-Durchlauf von Benchmark 1 wurden in EvidenzWerk **sechs konkrete technische bzw. methodische Schwachstellen** identifiziert und unmittelbar im Quellcode behoben (**Tabelle B1.4**):

Nr.	Betroffene Komponente & Quelldatei	Aufgedecktes Fehlerbild im Benchmark-1-Stresstest	Ursache (Root Cause Analysis)	Durchgeführte Korrektur & Verifikation
K1	**R-Bridge UTF-8-Zeichencodierung** (`Dockerfile`, `lib/r-service.ts`, `r/analyse-full.R`, `r/three-level.R`)	Deutsche Umlaute und Gedankenstriche in Moderator-Stufen (`Chronische Niereninsuffizienz`, `Typ-2-Diabetes – ASCVD`) erschienen in der JSON-Rückgabe der R-Bridge als Byte-Escape-Sequenzen (`<c3><a4>`, `<e2><80><93>`).	Im schlanken Debian-Container (`node:22-bookworm-slim`) lief `Rscript` ohne generierte `C.UTF-8`-Locale im POSIX-C-Modus (reines 7-Bit-ASCII), wodurch `jsonlite::toJSON()` Multibyte-UTF-8-Zeichen maskierte.	`locales`-Paket und `C.UTF-8` im `Dockerfile` aktiviert, `LC_ALL=C.UTF-8` und `LANG=C.UTF-8` in `lib/r-service.ts` erzwungen und `Encoding(x) <- "UTF-8"` rekursiv vor der JSON-Serialisierung in R ergänzt. Verifiziert: 100 % sauberes UTF-8.

Nr.	Betroffene Komponente & Quelldatei	Aufgedecktes Fehlerbild im Benchmark-1-Stresstest	Ursache (Root Cause Analysis)	Durchgeführte Korrektur & Verifikation
K2	**Domänen-blinder Extraktionsprompt** (`lib/extract.ts`)	Der KI-Extraktor war ursprünglich auf sozialwissenschaftliche Weiterbildungsstudien (*Learning & Development*, Mittelwerte `M`, `SD`, `d`, `g`) festgelegt und übersah in medizinischen Studien Zeit-bis-Ereignis-Maße (`HR`, `RR`, `OR`, `95%-KI`, `e1/n1`, `e2/n2`).	Statische Konstante `SYSTEM_PROMPT` ohne Einbindung der projektspezifischen Forschungsfrage und PICOS-Struktur.	Neue Funktion `buildPromptsForProject(project)` implementiert, die `question` sowie `picos` (`population`, `intervention`, `comparator`, `outcome`, `study_design`, `metric`) dynamisch in den System- und User-Prompt injiziert und sowohl kontinuierliche als auch binäre/Survival-Kennwerte strukturiert anfordert.
K3	**Paywall-Blockade & fehlende Register-Ersatzkette** (`lib/extract.ts`, [`app/api/studies/[id]/extract/route.ts`](file:///Users/franzhuetter/evidenzwerk/app/api/studies/%5Bid%5D/extract/route.ts#L22-L90))	Da *NEJM* den direkten Server-PDF-Abruf mit `HTTP 403` blockiert, brach die Extraktionsroute mit `Kein PDF für diese Studie verfügbar` ab, obwohl strukturierte Abstracts und ClinicalTrials.gov-Ergebnistabellen frei im Web vorlagen.	Die Route `/api/studies/[id]/extract` prüfte ausschließlich `hasPdf(studyId)` im lokalen PDF-Speicher ohne Open-Access-/Register-Fallback.	Dreistufige Open-Evidence-Ersatzkette (`fetchFallbackEvidenceText`) implementiert: (1) Europe PMC Volltext-JATS-XML → (2) PubMed EFetch-XML inkl. automatischer `<DataBankList>`-NCT-Erkennung → (3) ClinicalTrials.gov API v2 `resultsSection` (`outcomeMeasuresModule`), plus deterministische Effektgrößen-Ableitung `deriveEffectFromFindings()`. Ergebnis: **80 % exakte Trefferquote ohne ein einziges PDF**.
K4	**Schneeball-Referenzabruf bei *Lancet*-PII-DOIs** (`lib/scholarly.ts`)	Die Schneeball-Ernte (`referencesOfWorks`) lieferte für *Lancet*-DOIs vom Typ `10.1016/S0140-6736(22)02292-8` null Referenzen zurück.	Runde Klammern und Bindestriche in Elsevier/Lancet-PII-DOIs führten beim direkten Pfad-Aufruf in OpenAlex zu Fehlauflösungen bzw. leeren Referenzlisten.	In `referencesOfWorks()` Umstellung auf `filter=doi:`-Abfrage kombiniert mit einem automatischen Fallback auf die **Europe PMC References API** (`search` + `/references`). Ergebnis: Vollständige Auflösung der Referenzlisten.

Nr.	Betroffene Komponente & Quelldatei	Aufgedecktes Fehlerbild im Benchmark-1-Stresstest	Ursache (Root Cause Analysis)	Durchgeführte Korrektur & Verifikation
K5	**Doppelzählung der Effektgrößen bei Doppelcodierung** (`lib/repo.ts`)	Im Manuskript-Kontext stand in der Kopfzeile `322 Effektgrößen`, in der übergebenen Analysetabelle jedoch `161 Effektgrößen`. Das Sprachmodell entdeckte diese Differenz selbstständig und dokumentierte sie vorbildlich als Limitation `(f) Zählungsdiskrepanz der Effektgrößen`.	`projectCounts()` zählte `SELECT COUNT(*) FROM effect_sizes` über `coder = 'A'` und `coder = 'B'` zusammen (161 + 161 = 322), während `listEffectsForProject()` korrekt auf den primären Analysesatz `coder = 'A'` (161) filtert.	Filter `AND e.coder = 'A'` in `projectCounts()` (`lib/repo.ts`) ergänzt, sodass Kopfzähler und Analysetabelle exakt übereinstimmen (161 = 161).
K6	**Metrik-Priorisierung in der Manuskript-Vertiefung** (`lib/manuscript.ts`)	Wenn nach dem primären `HR`-Lauf (`Run v3`) ein sekundärer `RD`-Lauf über alle 5 Endpunkte (`Run v7`) ausgeführt wurde, las `analysisDeepeningLines()` für den Heterogenitäts-Kurzblock den allerletzten Lauf (`Run v7`, I^2 = 77,7 %) statt des freigegebenen Hauptlaufs (`Run v3`, I^2 = 0,0 %) aus.	`[...parsed].reverse().find(...)` berücksichtigte weder `released_run_id` noch die primäre Projektmetrik (`picos.metric`).	Auswahlfunktion `pickBestRun()` in `lib/manuscript.ts` implementiert, die strikt `released_run_id` arrow `primary`-Lauf der Hauptmetrik priorisiert und die Metrik jedes zitierten Laufs explizit benennt.

---

6. Literaturverzeichnis (Methodische Standardliteratur & Klinische Referenzarbeiten)

6.1 Methodische Standardliteratur zur systematischen Evidenzsynthese und Meta-Analyse

1. **Altman DG (1998):** Confidence intervals for the number needed to treat. *BMJ*, 317(7168): 1309–1312. DOI: `10.1136/bmj.317.7168.1309`.
2. **Assink M, Wibbelink CJM (2016):** Fitting three-level meta-analytic models in R: A step-by-step tutorial. *The Quantitative Methods for Psychology*, 12(3): 154–174. DOI: `10.20982/tqmp.12.3.p154`.
3. **Borenstein M, Hedges LV, Higgins JPT, Rothstein HR (2021):** *Introduction to Meta-Analysis* (2nd ed.). Chichester: John Wiley & Sons. DOI: `10.1002/9781119558378`.
4. **Borenstein M, Higgins JPT, Hedges LV, Rothstein HR (2017):** Basics of meta-analysis: I^2 is not an absolute measure of heterogeneity. *Research Synthesis Methods*, 8(1): 5–18. DOI: `10.1002/jrsm.1230`.
5. **Carter EC, Schönbrodt FD, Gervais WM, Hilgard J (2019):** Correcting for bias in psychology: A comparison of meta-analytic methods. *Advances in Methods and Practices in Psychological Science*, 2(2): 115–144. DOI: `10.1177/2515245919847196`.
6. **Cheung MWL (2014):** Modeling dependent effect sizes with three-level meta-analyses: A structural equation modeling approach. *Psychological Methods*, 19(2): 211–229. DOI: `10.1037/a0032968`.
7. **Cheung MWL (2015):** *Meta-Analysis: A Structural Equation Modeling Approach*. Chichester: John Wiley & Sons. DOI: `10.1002/9781118957813`.
8. **DerSimonian R, Laird N (1986):** Meta-analysis in clinical trials. *Controlled Clinical Trials*, 7(3): 177–188. DOI: `10.1016/0197-2456(86)90046-2`.
9. **Duval S, Tweedie R (2000):** Trim and fill: A simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. *Biometrics*, 56(2): 455–463. DOI:

10.1111/j.0006-341x.2000.00455.x.

10. **Egger M, Davey Smith G, Schneider M, Minder C (1997):** Bias in meta-analysis detected by a simple, graphical test. *BMJ*, 315(7109): 629–634. DOI: 10.1136/bmj.315.7109.629.
11. **Freeman MF, Tukey JW (1950):** Transformations related to the angular and the square root. *The Annals of Mathematical Statistics*, 21(4): 607–611. DOI: 10.1214/aoms/1177729756.
12. **Guyatt GH, Oxman AD, Vist GE, Kunz R, Falck-Ytter Y, Alonso-Coello P, Schünemann HJ; GRADE Working Group (2008):** GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. *BMJ*, 336(7650): 924–926. DOI: 10.1136/bmj.39489.470347.AD.
13. **Guyatt GH, Oxman AD, Schünemann HJ, Tugwell P, Knottnerus A (2011):** GRADE guidelines: A new series of articles in the *Journal of Clinical Epidemiology*. *Journal of Clinical Epidemiology*, 64(4): 380–382. DOI: 10.1016/j.jclinepi.2010.09.011.
14. **Hartung J, Knapp G (2001):** A refined method for the meta-analysis of controlled clinical trials with binary outcome. *Statistics in Medicine*, 20(24): 3875–3889. DOI: 10.1002/sim.1009.
15. **Hedges LV (1981):** Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics*, 6(2): 107–128. DOI: 10.3102/10769986006002107.
16. **Higgins JPT, Thompson SG, Deeks JJ, Altman DG (2003):** Measuring inconsistency in meta-analyses. *BMJ*, 327(7414): 557–560. DOI: 10.1136/bmj.327.7414.557.
17. **Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, Welch VA (Hrsg.) (2023):** *Cochrane Handbook for Systematic Reviews of Interventions* (Version 6.4). Cochrane. Verfügbar unter: https://training.cochrane.org/handbook.
18. **IntHout J, Ioannidis JPA, Borm GF (2014):** The Hartung-Knapp-Sidik-Jonkman method for random effects meta-analysis is straightforward and considerably outperforms the standard DerSimonian-Laird method. *BMC Medical Research Methodology*, 14: 25. DOI: 10.1186/1471-2288-14-25.
19. **IntHout J, Ioannidis JPA, Rovers MM, Goeman JJ (2016):** Plea for routinely presenting prediction intervals in meta-analysis. *BMJ Open*, 6(7): e010247. DOI: 10.1136/bmjopen-2015-010247.
20. **Konstantopoulos S (2011):** Fixed effects and variance components estimation in three-level meta-analysis. *Research Synthesis Methods*, 2(1): 61–76. DOI: 10.1002/jrsm.35.
21. **Langan D, Higgins JPT, Jackson D, Bowden J, Veroniki AA, Kontopantelis E, Viechtbauer W, Simmonds M (2019):** A comparison of heterogeneity variance estimators in simulated random-effects meta-analyses. *Research Synthesis Methods*, 10(1): 83–98. DOI: 10.1002/jrsm.1316.
22. **Moher D, Shamseer L, Clarke M et al.; PRISMA-P Group (2015):** Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement. *Systematic Reviews*, 4(1): 1. DOI: 10.1186/2046-4053-4-1.
23. **Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD et al. (2021):** The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*, 372: n71. DOI: 10.1136/bmj.n71.
24. **Parmar MKB, Torri V, Stewart L (1998):** Extracting summary statistics to perform meta-analyses of the published literature for survival endpoints. *Statistics in Medicine*, 17(24): 2815–2834. DOI: 10.1002/(SICI)1097-0258(19981230)17:24<2815::AID-SIM110>3.0.CO;2-8.
25. **Pustejovsky JE, Tipton E (2018):** Small-sample methods for cluster-robust variance estimation and hypothesis testing in fixed effects models. *Journal of Business & Economic Statistics*, 36(4): 672–683. DOI: 10.1080/07350015.2016.1247004.
26. **Pustejovsky JE, Tipton E (2022):** Meta-analysis with robust variance estimation: Expanding the range of working models. *Prevention Science*, 23(3): 425–438. DOI: 10.1007/s11121-021-01246-3.
27. **Rethlefsen ML, Kirtley S, Waffenschmidt S et al.; PRISMA-S Group (2021):** PRISMA-S: an extension to the PRISMA Statement for Reporting Literature Searches in Systematic Reviews. *Systematic Reviews*, 10(1): 39. DOI: 10.1186/s13643-020-01542-z.
28. **Riley RD, Higgins JPT, Deeks JJ (2011):** Interpretation of random effects meta-analyses. *BMJ*, 342: d549.

DOI: 10.1136/bmj.d549.

29. **Rücker G, Schwarzer G, Carpenter JR, Schumacher M (2008):** Undue reliance on I² in assessing heterogeneity may mislead. *BMC Medical Research Methodology*, 8: 79. DOI: 10.1186/1471-2288-8-79.
30. **Simonsohn U, Nelson LD, Simmons JP (2014):** P-curve: A key to the file-drawer. *Journal of Experimental Psychology: General*, 143(2): 534–547. DOI: 10.1037/a0033242.
31. **Stanley TD, Doucouliagos H (2014):** Meta-regression approximations to reduce publication selection bias. *Research Synthesis Methods*, 5(1): 60–78. DOI: 10.1002/jrsm.1095.
32. **Sterne JAC, Savović J, Page MJ, Elbers RG, Blencowe NS, Boutron I et al. (2019):** RoB 2: a revised tool for assessing risk of bias in randomised trials. *BMJ*, 366: l4898. DOI: 10.1136/bmj.l4898.
33. **Tierney JF, Stewart LA, Ghersi D, Burdett S, Sydes MR (2007):** Practical methods for incorporating summary time-to-event data into meta-analysis. *Trials*, 8: 16. DOI: 10.1186/1745-6215-8-16.
34. **van Aert RCM, Wicherts JM, van Assen MALM (2016):** Conducting meta-analyses based on p values: Reservations and recommendations for applying *p*-uniform and *p*-curve. *Perspectives on Psychological Science*, 11(5): 713–729. DOI: 10.1177/1745691616650874.
35. **Van den Noortgate W, López-López JA, Marín-Martínez F, Sánchez-Meca J (2013):** Three-level meta-analysis of dependent effect sizes. *Behavior Research Methods*, 45(2): 576–594. DOI: 10.3758/s13428-012-0261-6.
36. **Veroniki AA, Jackson D, Viechtbauer W, Bender R, Bowden J, Knapp G et al. (2016):** Methods to estimate the between-study variance and its uncertainty in meta-analysis. *Research Synthesis Methods*, 7(1): 55–79. DOI: 10.1002/jrsm.1164.
37. **Viechtbauer W (2010):** Conducting meta-analyses in R with the `metafor` package. *Journal of Statistical Software*, 36(3): 1–48. DOI: 10.18637/jss.v036.i03.

6.2 Klinische Referenz-Meta-Analysen (*The Lancet* Goldstandard) und eingeschlossene Phase-III-Primärstudien

1. **Nuffield Department of Population Health Renal Studies Group & SGLT2 inhibitor Meta-Analysis Cardio-Renal Trialists' Consortium (SMART-C) (2022):** Impact of diabetes on the effects of sodium glucose co-transporter-2 inhibitors on kidney outcomes: collaborative meta-analysis of large placebo-controlled trials. *The Lancet*, 400(10365): 1788–1801. DOI: 10.1016/S0140-6736(22)02292-8.
2. **Vaduganathan M, Docherty KF, Claggett BL, Jhund PS, de Boer RA, Hernandez AF et al. (2022):** SGLT-2 inhibitors in patients with heart failure: a comprehensive meta-analysis of five randomised controlled trials. *The Lancet*, 400(10354): 757–767. DOI: 10.1016/S0140-6736(22)01429-5.
3. **Khan MS, Usman MS, Butler J, Bhatt DL et al. (2024):** SGLT2 inhibitors and cardiovascular outcomes across the cardiometabolic spectrum, including post-myocardial infarction: a systematic review and meta-analysis. *The Lancet Diabetes & Endocrinology*, 12(8): 545–558. DOI: 10.1016/S2213-8587(24)00102-5.
4. **Zinman B, Wanner C, Lachin JM et al.; EMPA-REG OUTCOME Investigators (2015):** Empagliflozin, cardiovascular outcomes, and mortality in type 2 diabetes. *New England Journal of Medicine*, 373(22): 2117–2128. DOI: 10.1056/NEJMoa1504720.
5. **Neal B, Perkovic V, Mahaffey KW et al.; CANVAS Program Collaborative Group (2017):** Canagliflozin and cardiovascular and renal events in type 2 diabetes. *New England Journal of Medicine*, 377(7): 644–657. DOI: 10.1056/NEJMoa1611925.
6. **Wiviott SD, Raz I, Bonaca MP et al.; DECLARE–TIMI 58 Investigators (2019):** Dapagliflozin and cardiovascular outcomes in type 2 diabetes. *New England Journal of Medicine*, 380(4): 347–357. DOI: 10.1056/NEJMoa1812389.
7. **Perkovic V, Jardine MJ, Neal B et al.; CREDENCE Trial Investigators (2019):** Canagliflozin and renal

outcomes in type 2 diabetes and nephropathy. *New England Journal of Medicine*, 380(24): 2295–2306. DOI: 10.1056/NEJMoa1811744.

8. **McMurray JJV, Solomon SD, Inzucchi SE et al.; DAPA-HF Trial Committees and Investigators (2019):** Dapagliflozin in patients with heart failure and reduced ejection fraction. *New England Journal of Medicine*, 381(21): 1995–2008. DOI: 10.1056/NEJMoa1911303.
9. **Cannon CP, Pratley R, Dagogo-Jack S et al.; VERTIS CV Investigators (2020):** Cardiovascular outcomes with ertugliflozin in type 2 diabetes. *New England Journal of Medicine*, 383(15): 1425–1435. DOI: 10.1056/NEJMoa2004967.
10. **Packer M, Anker SD, Butler J et al.; EMPEROR-Reduced Trial Investigators (2020):** Cardiovascular and renal outcomes with empagliflozin in heart failure. *New England Journal of Medicine*, 383(15): 1413–1424. DOI: 10.1056/NEJMoa2022190.
11. **Heerspink HJL, Stefánsson BV, Correa-Rotter R et al.; DAPA-CKD Trial Committees and Investigators (2020):** Dapagliflozin in patients with chronic kidney disease. *New England Journal of Medicine*, 383(15): 1436–1446. DOI: 10.1056/NEJMoa2024816.
12. **Bhatt DL, Szarek M, Steg PG et al.; SOLOIST-WHF Trial Investigators (2021):** Sotagliflozin in patients with diabetes and recent worsening heart failure. *New England Journal of Medicine*, 384(2): 117–128. DOI: 10.1056/NEJMoa2030183.
13. **Bhatt DL, Szarek M, Pitt B et al.; SCORED Investigators (2021):** Sotagliflozin in patients with diabetes and chronic kidney disease. *New England Journal of Medicine*, 384(2): 129–139. DOI: 10.1056/NEJMoa2030186.
14. **Anker SD, Butler J, Filippatos G et al.; EMPEROR-Preserved Trial Investigators (2021):** Empagliflozin in heart failure with a preserved ejection fraction. *New England Journal of Medicine*, 385(16): 1451–1461. DOI: 10.1056/NEJMoa2107038.
15. **Solomon SD, McMurray JJV, Claggett B et al.; DELIVER Trial Committees and Investigators (2022):** Dapagliflozin in heart failure with mildly reduced or preserved ejection fraction. *New England Journal of Medicine*, 387(12): 1089–1098. DOI: 10.1056/NEJMoa2206286.
16. **The EMPA-KIDNEY Collaborative Group; Herrington WG, Staplin N, Wanner C et al. (2023):** Empagliflozin in patients with chronic kidney disease. *New England Journal of Medicine*, 388(2): 117–127. DOI: 10.1056/NEJMoa2204233.
17. **James S, Erlinge D, Storey RF et al.; DAPA-MI Investigators (2023):** Dapagliflozin in myocardial infarction without diabetes or heart failure. *NEJM Evidence*, 3(2): EVIDoa2300286. DOI: 10.1056/EVIDoa2300286.
18. **Butler J, Jones WS, Udell JA et al.; EMPACT-MI Investigators (2024):** Empagliflozin after acute myocardial infarction. *New England Journal of Medicine*, 390(16): 1455–1466. DOI: 10.1056/NEJMoa2314051.

Anhang A (Teil II): Vollständiger Originalwortlaut des autonom generierten EvidenzWerk-Manuskripts (12.859 Wörter)

Dokumentationshinweis für die Publikation: *Der nachfolgende Text wurde durch EvidenzWerk (*POST /api/projects/16/manuscript*, Modell* anthropic/claude-opus-5*, Daten-Snapshot-Hash* 171f231cf6005e28*) vollautomatisch aus den strukturierten Projektdaten erzeugt. Er wird hier als Primärbeleg für* ***Stufe 3 („Das Endergebnis“)*** *im Originalwortlaut wiedergegeben. Stellen, an denen das Sprachmodell im Vor-Kalibrierungs-Snapshot selbstständig die Zählkonvention zwischen Gesamttabelle (*322 *inkl. Zweitcodierer B) und Primär-Analysetabelle (*161 *Codierer A) bzw. nicht aggregierte Einzelstudien-Metadaten mit* [PLATZHALTER] *markiert hat (siehe Kalibrierungs-Protokoll Abschnitt 5, Punkte K5 und K6), sind bewusst im Original belassen, um die strikte Halluzinationsverweigerung des*

Systems zu dokumentieren.

SGLT2-Inhibitoren über das kardiometabolische Krankheitsspektrum: Drei-Ebenen-Meta-Analyse von 15 placebokontrollierten Phase-III-Studien mit 100.952 Patient:innen

Manuskript-Entwurf (Version 1, Datenstand 21.10.2026)

Autor:innen: [PLATZHALTER] Korrespondenzadresse: [PLATZHALTER] Registrierung: [PLATZHALTER — PROSPERO-Registrierungsnummer] Interessenkonflikte: [PLATZHALTER] Förderung: [PLATZHALTER] Datenverfügbarkeit: [PLATZHALTER — Repositorium/DOI des Analyse-Snapshots]

---

1 Strukturiertes Abstract

Hintergrund. Hemmer des natriumabhängigen Glukose-Kotransporters 2 (SGLT2-Inhibitoren) wurden ursprünglich als Antidiabetika entwickelt, haben sich aber in einer Serie großer, ereignisgesteuerter Phase-III-Studien als kardiorenale Substanzklasse etabliert. Die vorliegende Synthese prüft, ob die Effekte auf harte kardiovaskuläre und renale Endpunkte über das gesamte kardiometabolische Krankheitsspektrum hinweg — Herzinsuffizienz mit reduzierter (HFrEF) und mit erhaltener Ejektionsfraktion (HFpEF), Typ-2-Diabetes mit hohem kardiovaskulärem Risiko beziehungsweise manifester atherosklerotischer Erkrankung (ASCVD), chronische Niereninsuffizienz (CKD) und akuter Myokardinfarkt (AMI) — konsistent sind, und ordnet die Befunde in den durch das SMART-C-Konsortium (The Lancet 2022), Vaduganathan et al. (The Lancet 2022) sowie Khan, Usman, Butler, Bhatt et al. (Lancet Diabetes & Endocrinology 2024) markierten Forschungsstand ein.

Methoden. Systematische Recherche in PubMed/MEDLINE, OpenAlex, Crossref und ClinicalTrials.gov (alle Läufe am 21.10.2026 dokumentiert) mit vorab festgelegten PICOS-Kriterien: erwachsene Patient:innen des kardiometabolischen Spektrums; orale Therapie mit einem selektiven SGLT2-Inhibitor (Empagliflozin 10–25 mg, Dapagliflozin 10 mg, Canagliflozin 100–300 mg, Ertugliflozin 5–15 mg) oder dem dualen SGLT1/2-Inhibitor Sotagliflozin (200–400 mg) zusätzlich zur leitliniengerechten Standardtherapie; Placebo auf identischer Hintergrundtherapie; doppelblinde, multizentrische, ereignisgesteuerte Phase-III-Outcome-Studien mit N ≥ 1.000 (Publikationszeitraum 2015 bis 31.05.2024). Primärer Endpunkt war die Zeit bis zum ersten Auftreten von kardiovaskulärem Tod oder Hospitalisierung wegen Herzinsuffizienz (CV Death or HHF) als Hazard Ratio (HR); sekundäre Endpunkte waren erste Herzinsuffizienz-Hospitalisierung (First HHF), kardiovaskulärer Tod, Gesamtmortalität und Nierenkrankheitsprogression. Doppelscreening durch zwei unabhängige Reviewer (Cohens κ = 1,00; Übereinstimmung 100 %, n = 36). Gepoolt wurde mit einem Drei-Ebenen-Modell mit robuster Varianzschätzung (metafor::rma.mv, clubSandwich CR2) sowie — als Referenzmodell auf Studienebene — mit einem Zwei-Ebenen-Random-Effects-Modell (REML). Risk of Bias wurde mit RoB 2 beurteilt, die Evidenzsicherheit endpunktweise nach GRADE.

Ergebnisse. Von 44 identifizierten Treffern wurden 36 Datensätze importiert (8 Duplikat-Markierungen), 36 im Titel-/Abstract-Screening bewertet, 18 ausgeschlossen, 18 Volltexte geprüft und 3 ausgeschlossen; 15 Studien mit insgesamt N = 100.952 Patient:innen gingen in die Synthese ein (322 Effektgrößen im Audit-Trail; 161 Effektgrößen in der analysierten Effektgrößentabelle, Metriken HR, Risk Ratio und Risikodifferenz). Für den primären kombinierten Endpunkt ergab das Drei-Ebenen-Modell mit CR2-Varianzschätzung eine gepoolte HR von 0,78 (95 %-KI 0,75–0,82; k = 15); das Zwei-Ebenen-Referenzmodell lieferte denselben Punktschätzer (HR 0,78; 95 %-KI 0,75–0,82; p < 0,001; I^2 = 0 %; Q = 14,1; p = 0,446; τ^2 = 0,0000; Prädiktionsintervall 0,75–0,82). Im endpunktübergreifenden Drei-Ebenen-Modell lag die Gesamtheterogenität bei I^2 = 77,7 % (Studien-Ebene 30,7 %, Effekt-Ebene 46,9 %). Auf der Skala der Risikodifferenz (RD) betrugen die gepoolten absoluten Effekte −0,019 für CV Death or HHF (NNT ≈ 53), −0,017 für First HHF (NNT ≈ 59), −0,010 für die Gesamtmortalität (NNT ≈ 100), −0,008 für den kardiovaskulären Tod (NNT ≈ 125) und −0,010 für die Nierenkrankheitsprogression

(NNT ≈ 100; k = 13); der Wald-Test über die Endpunkte war signifikant (p = 0,003). Nach Krankheitsentität zeigten sich HR-Schätzer von 0,75 (CKD; k = 4), 0,78 (Herzinsuffizienz; k = 5), 0,79 (Typ-2-Diabetes/ASCVD; k = 4) und 0,90 (AMI; k = 2), ohne statistisch belegbare Interaktion (Q-between = 3,94; p = 0,268). Die Meta-Regression auf das Publikationsjahr ergab eine kleine, formal signifikante Abschwächung (Steigung 0,002 auf der Log-Skala; p = 0,043). Leave-one-out-Analysen bewegten den Schätzer zwischen 0,78 und 0,79. Hinweise auf Publikationsbias fanden sich nicht (Egger-Intercept −0,38; SE 1,02; p = 0,714; Trim-and-Fill k_0 = 0, adjustiert 0,785; PEESE 0,786; p-uniform* 0,781; p-Curve rechtssteil, p < 0,001). Alle 15 Studien wurden nach RoB 2 insgesamt als „niedriges Risiko“ eingestuft; die Evidenzsicherheit wurde für alle fünf Endpunkte als „hoch“ ratifiziert.

Schlussfolgerung. SGLT2-Inhibitoren senken das Risiko für kardiovaskulären Tod oder Herzinsuffizienz-Hospitalisierung um rund ein Fünftel, mit größter relativer Wirkung auf Herzinsuffizienz-Hospitalisierungen und Nierenkrankheitsprogression und kleinerer, aber konsistenter Wirkung auf Mortalitätsendpunkte. Die Konsistenz über HFrEF, HFpEF, Typ-2-Diabetes, CKD und — mit numerisch schwächerem Effekt — die Post-Myokardinfarkt-Population stützt das Konzept einer krankheitsübergreifenden kardiorenalen Basistherapie. Der absolute Nutzen skaliert mit dem Ausgangsrisiko, was die Verlagerung der Indikationsstellung vom Glukosestoffwechsel hin zum kardiorenalen Risikoprofil begründet.

Schlagwörter: SGLT2-Inhibitoren; Herzinsuffizienz; chronische Niereninsuffizienz; Typ-2-Diabetes; Myokardinfarkt; Meta-Analyse; Drei-Ebenen-Modell; robuste Varianzschätzung

---

2 Einleitung

2.1 Ausgangslage

Die Entwicklung der SGLT2-Inhibitoren ist ein ungewöhnlicher Fall in der kardiovaskulären Pharmakotherapie: Eine Substanzklasse, die zur Senkung des Blutzuckers zugelassen wurde, hat sich innerhalb eines Jahrzehnts zu einer tragenden Säule der Herzinsuffizienz- und Nephroprotektionstherapie entwickelt. Der Auslöser war die regulatorisch geforderte kardiovaskuläre Sicherheitsprüfung antidiabetischer Substanzen. Was als Nicht-Unterlegenheitsnachweis angelegt war, zeigte in EMPA-REG OUTCOME (Zinman et al. 2015, N = 7.020) und anschließend im CANVAS-Programm (Neal et al. 2017, N = 10.142) sowie in DECLARE-TIMI 58 (Wiviott et al. 2019, N = 17.160) ein Signal, das sich nicht aus der glykämischen Wirkung erklären ließ: eine ausgeprägte Reduktion der Hospitalisierungen wegen Herzinsuffizienz, begleitet von renalen Vorteilen, und das bei weitgehend unveränderten atherothrombotischen Endpunkten.

Dieses Muster hat die Forschungsagenda umgestellt. Statt weiterer Diabetes-Studien folgten dedizierte Endpunktstudien in Populationen, die primär über ihr kardiorenales Risiko und nicht über ihren Glukosestoffwechsel definiert waren: DAPA-HF (McMurray et al. 2019, N = 4.744) und EMPEROR-Reduced (Packer et al. 2020, N = 3.730) für HFrEF; EMPEROR-Preserved (Anker et al. 2021, N = 5.988) und DELIVER (Solomon et al. 2022, N = 6.263) für die mittelgradig reduzierte und erhaltene Ejektionsfraktion; SOLOIST-WHF (Bhatt et al. 2021, N = 1.222) für die dekompensierte, kürzlich verschlechterte Herzinsuffizienz; CREDENCE (Perkovic et al. 2019, N = 4.401), DAPA-CKD (Heerspink et al. 2020, N = 4.304), SCORED (Bhatt et al. 2021, N = 10.584) und EMPA-KIDNEY (Herrington et al. 2023, N = 6.609) für die chronische Niereninsuffizienz. Zuletzt wurde das Spektrum auf die akute Phase nach Myokardinfarkt ausgeweitet: DAPA-MI (James et al. 2024, N = 4.017) und EMPACT-MI (Butler et al. 2024, N = 6.522).

Mechanistisch wird dieser krankheitsübergreifende Effekt nicht mehr primär über die Glukosurie erklärt. Verma und McMurray (2018, DOI 10.1007/s00125-018-4670-7) haben die konkurrierenden Hypothesen systematisch geordnet: Natriurese und Veränderung des intravasalen Volumen- und Interstitialhaushalts, eine Verschiebung des myokardialen Substratstoffwechsels in Richtung Ketonkörper mit verbesserter myokardialer Energieeffizienz, Effekte auf den Natrium-Wasserstoff-Austauscher, Reduktion von Vor- und Nachlast, Hemmung der tubuloglomerulären Rückkopplungsstörung und damit Senkung des intraglomerulären Drucks. Keiner dieser Mechanismen ist allein hinreichend; gerade ihre Kombination erklärt aber, warum die Wirkung weder vom

HbA1c noch von der Ejektionsfraktion stark abzuhängen scheint. Ergänzend hat die Post-hoc-Synthese von Zelniker et al. (2020, DOI 10.1161/circulationaha.119.044231) das Augenmerk auf arrhythmische Endpunkte, insbesondere Vorhofflimmern, gelenkt und damit den Diskurs über mögliche sekundäre Wirkpfade erweitert.

2.2 Forschungsstand und Lücke

Der quantitative Forschungsstand wird derzeit von drei Arbeiten getragen, die unterschiedliche Ausschnitte des Spektrums abdecken.

Erstens die kollaborative Meta-Analyse des SMART-C-Konsortiums gemeinsam mit der Nuffield Department of Population Health Renal Studies Group (2022, DOI 10.1016/s0140-6736(22)02292-8). Sie adressiert die Frage, ob der Diabetesstatus die renalen Effekte der SGLT2-Inhibition moduliert, und arbeitet dafür auf der Ebene großer placebokontrollierter Studien mit harmonisierten renalen Endpunktdefinitionen. Ihr Fokus liegt auf den Nierenendpunkten; kardiale Endpunkte sind in diesem Rahmen nachgeordnet.

Zweitens die präspezifizierte Meta-Analyse von Vaduganathan, Docherty, Claggett, Jhund, de Boer et al. (2022, DOI 10.1016/s0140-6736(22)01429-5). Sie kombiniert zunächst DELIVER und EMPEROR-Preserved und erweitert anschließend um DAPA-HF, EMPEROR-Reduced und SOLOIST-WHF, um auf Studienebene mit harmonisierten Endpunktdefinitionen die Effekte auf kardiovaskuläre Mortalität und in Subgruppen über das Spektrum der Ejektionsfraktion hinweg zu schätzen. Ihr erklärtes Motiv ist, dass die Evidenz für höhere Ejektionsfraktionen weniger gut etabliert war als für HFrEF. Damit ist sie die Referenzarbeit für die fünf Herzinsuffizienz-Studien — aber eben nur für diese fünf.

Drittens die systematische Übersichtsarbeit und Meta-Analyse von Khan, Usman, Butler, Bhatt et al. (2024, DOI 10.1016/s2213-8587(24)00102-5), die den Anspruch formuliert, Herzinsuffizienz-Endpunkte und kardiovaskulären Tod „across the cardiometabolic disease spectrum“ zu betrachten und damit erstmals die Post-Myokardinfarkt-Studien in dieselbe Betrachtung einbezieht.

Trotz dieser Dichte bleiben drei Lücken bestehen, an denen die vorliegende Arbeit ansetzt.

(a) Statistische Abhängigkeitsstruktur. Alle drei quantitativen Vorarbeiten arbeiten mit Studienaggregaten und, soweit aus den vorliegenden Abstracts ersichtlich, mit klassischen Zwei-Ebenen-Modellen. Wenn man jedoch — wie hier — fünf harte Endpunkte in drei Metriken (HR, Risk Ratio, Risikodifferenz) aus denselben 15 Studien gemeinsam modelliert, entstehen pro Studie mehrere abhängige Effektgrößen. Diese Abhängigkeit zu ignorieren, führt zu zu engen Konfidenzintervallen und zu einer Fehlattribution von Heterogenität. Die vorliegende Arbeit verwendet daher ein Drei-Ebenen-Modell mit robuster Varianzschätzung (CR2), das Varianz explizit in eine Studien- und eine Effektgrößen-Komponente zerlegt und die Inferenz gegenüber Fehlspezifikationen der Abhängigkeitsstruktur absichert.

(b) Absolute Effekte neben relativen. Die Debatte um SGLT2-Inhibitoren wird überwiegend in Hazard Ratios geführt. Für Leitlinien, Erstattungsentscheidungen und die Aufklärung von Patient:innen ist jedoch die absolute Risikoreduktion entscheidend, weil sie mit dem Ausgangsrisiko und damit mit der Krankheitsentität variiert. Die vorliegende Analyse schätzt dieselben fünf Endpunkte parallel auf der Log-Skala und auf der Skala der Risikodifferenz und leitet daraus Number-needed-to-treat-Werte ab.

(c) Datenstand und Spektrumbreite in einem Modell. Die SMART-C-Kohorte umfasst in der hier verwendeten Abgrenzung 13 Studien mit N = 90.413; die Post-MI-Erweiterung fügt zwei Studien mit N = 10.539 hinzu. Die vorliegende Arbeit analysiert beide gemeinsam (N = 100.952) und testet die Zeitschicht „SMART-C-Kohorte 2015–2023“ gegen „Post-MI-Erweiterung 2024“ explizit als Moderator — also nicht als narrativen Zusatz, sondern als statistische Frage.

2.3 Ziel und Forschungsfrage

Ziel dieser Arbeit ist eine umfassende, methodisch abgesicherte quantitative Synthese der großen doppelblinden placebokontrollierten Phase-III-Studien zu SGLT2- und dualen SGLT1/2-Inhibitoren über das kardiometabolische Krankheitsspektrum hinweg.

Die Forschungsfrage lautet: Senken SGLT2-Inhibitoren (Empagliflozin, Dapagliflozin, Canagliflozin, Ertugliflozin, Sotagliflozin) im Vergleich zu Placebo in großen doppelblinden randomisierten kontrollierten Studien (N ≥ 1.000) das Risiko für (1) kardiovaskulären Tod oder Herzinsuffizienz-Hospitalisierung, (2) erste Herzinsuffizienz-Hospitalisierung, (3) kardiovaskulären Tod, (4) Gesamtmortalität und (5) Nierenkrankheitsprogression — und ist dieser Effekt über HFrEF, HFpEF, Typ-2-Diabetes/ASCVD, CKD und die Post-Myokardinfarkt-Population konsistent?

Nachgeordnete Fragen betreffen die Quelle und Struktur der Heterogenität, die Stabilität des Schätzers gegenüber Modellwahl und Einzelstudien, mögliche Moderation durch Substanz, Dosisschema, Förderquelle, geografische Breite und Publikationsjahr sowie die Prüfung auf Publikationsbias und Small-Study-Effekte.

---

3 Methoden

3.1 Registrierung, Protokoll und Berichtsstandard

Das Review-Protokoll wurde vor Beginn des Screenings festgelegt; die Registrierungsnummer ist [PLATZHALTER — PROSPERO-ID]. Abweichungen vom Protokoll: [PLATZHALTER]. Die Berichterstattung folgt dem PRISMA-2020-Standard; die ausgefüllte Checkliste und das Flussdiagramm liegen dem Projekt-Report bei (siehe Abschnitt 9).

Konzeptionell orientiert sich das Vorgehen an drei Referenzarbeiten, deren Einschlusslogik bewusst nachgebildet wurde, um Vergleichbarkeit herzustellen: dem Protokoll des SGLT2 inhibitor Meta-Analysis Cardio-Renal Trialists Consortium (SMART-C, The Lancet 2022), der Herzinsuffizienz-Synthese von Vaduganathan et al. (The Lancet 2022) sowie der spektrumübergreifenden Übersicht von Khan, Usman, Butler, Bhatt et al. (Lancet Diabetes & Endocrinology 2024). Die Übernahme der Einschlusslogik bezieht sich auf die Schwelle N ≥ 1.000, die Beschränkung auf ereignisgesteuerte, doppelblinde Phase-III-Outcome-Studien und die Harmonisierung der Endpunktdefinitionen; sie bedeutet ausdrücklich keine Übernahme von Daten oder Schätzern aus diesen Arbeiten.

3.2 Informationsquellen und Suchstrategie

Es wurden vier Quellen systematisch durchsucht; alle Läufe sind mit Datum, Suchstring und Trefferzahl im Audit-Trail dokumentiert.

PubMed/MEDLINE (21.10.2026), 18 Treffer:

```
("sodium-glucose transporter 2 inhibitors"[MeSH] OR "SGLT2" OR empagliflozin
OR dapagliflozin OR canagliflozin OR ertugliflozin OR sotagliflozin)
AND ("heart failure"[MeSH] OR "cardiovascular death"
OR "renal insufficiency, chronic"[MeSH] OR "myocardial infarction"[MeSH])
AND ("randomized controlled trial"[pt] OR placebo)
AND 2015/01/01:2024/05/31[dp]
```

OpenAlex (21.10.2026), 12 Treffer:

```
SGLT2 inhibitor randomized placebo-controlled trial cardiovascular death
hospitalization heart failure kidney progression
```

Crossref (21.10.2026), 8 Treffer:

```
empagliflozin dapagliflozin canagliflozin ertugliflozin sotagliflozin
NEJM randomized placebo
```

ClinicalTrials.gov (21.10.2026), 6 Treffer:

```
INTERVENTION: (empagliflozin OR dapagliflozin OR canagliflozin
OR ertugliflozin OR sotagliflozin)
AND STUDY_TYPE: Interventional AND PHASE: Phase 3 AND RESULTS: Has Results
```

Der Publikationszeitraum war auf den 01.01.2015 bis 31.05.2024 begrenzt, um die Einschlussfenster der Referenzarbeiten abzubilden. Sprachrestriktionen: [PLATZHALTER]. Eine ergänzende Handsuche in Referenzlisten eingeschlossener Studien und Übersichtsarbeiten sowie eine Vorwärtszitationssuche wurden

durchgeführt in folgendem Umfang: [PLATZHALTER]. Graue Literatur, Konferenzabstracts und Zulassungsdossiers wurden in folgendem Umfang berücksichtigt: [PLATZHALTER].

3.3 Ein- und Ausschlusskriterien (PICOS)

P — Population. Erwachsene Patient:innen (≥ 18 Jahre) im kardiometabolischen Krankheitsspektrum: chronische oder dekompensierte Herzinsuffizienz einschließlich HFrEF und HFpEF; Typ-2-Diabetes mit hohem kardiovaskulärem Risiko oder manifester ASCVD; chronische Niereninsuffizienz; akuter Myokardinfarkt. Gesamtstichprobe der eingeschlossenen Studien: N = 100.952, davon N = 90.413 in den 13 Studien der SMART-C-Abgrenzung und N = 21.947 in den fünf Herzinsuffizienz-Studien.

I — Intervention. Orale Behandlung mit einem selektiven SGLT2-Inhibitor (Empagliflozin 10–25 mg/d, Dapagliflozin 10 mg/d, Canagliflozin 100–300 mg/d, Ertugliflozin 5–15 mg/d) oder dem dualen SGLT1/2-Inhibitor Sotagliflozin (200–400 mg/d), jeweils zusätzlich zur leitliniengerechten Standardtherapie.

C — Komparator. Placebo zusätzlich zur identischen leitliniengerechten kardiorenalen und antidiabetischen Hintergrundtherapie, doppelblind verabreicht.

O — Endpunkte. Primär: Zeit bis zum ersten Auftreten von kardiovaskulärem Tod oder Hospitalisierung wegen Herzinsuffizienz (CV Death or HHF), berichtet als Hazard Ratio. Sekundär: (1) erste Hospitalisierung wegen Herzinsuffizienz (First HHF); (2) kardiovaskulärer Tod; (3) Gesamtmortalität; (4) Nierenkrankheitsprogression, definiert als anhaltender eGFR-Abfall um ≥ 40–50 %, terminale Niereninsuffizienz oder renaler Tod.

S — Studiendesign. Multizentrische, ereignisgesteuerte, doppelblinde, placebokontrollierte randomisierte Phase-III-Outcome-Studien mit N ≥ 1.000, publiziert zwischen 2015 und Mai 2024.

Ausschlusskriterien. Nicht-randomisierte Designs einschließlich Register- und Versicherungsdatenkohorten mit Propensity-Score-Matching; Mehrfachpublikationen, Subgruppen- oder Sekundärberichte bereits eingeschlossener Kohorten; Studien mit N < 1.000; Studien ohne ereignisgesteuerte Time-to-Event-Analyse (etwa hierarchische Win-Ratio-Primärendpunkte mit kurzem Follow-up); offene oder einfachblinde Designs; aktive Komparatoren ohne Placeboarm.

3.4 Studienauswahl und Screening-Prozess

Alle importierten Datensätze wurden im Doppelscreening von zwei unabhängigen Reviewer:innen (A und B) zunächst auf Titel-/Abstract-Ebene und anschließend im Volltext gegen die PICOS-Kriterien geprüft. Die Interrater-Übereinstimmung lag bei 100 % (n = 36 bewertete Datensätze), Cohens κ = 1,00. Konflikte wurden einvernehmlich geklärt; zum Zeitpunkt des Datenschnitts bestanden 0 offene Konflikte. Eine dritte Instanz zur Konfliktschlichtung war damit nicht erforderlich.

Volltext-Ausschlüsse wurden mit Begründung dokumentiert. Drei Datensätze wurden auf Volltextebene ausgeschlossen:

1. **Mehrfachpublikation / Subgruppenbericht:** Sekundärauswertung der bereits vollständig eingeschlossenen DAPA-HF-Hauptstudie (McMurray et al. 2019, NCT03036124).
2. **Sekundärbericht derselben Kohorte:** renale Einzelpublikation zur bereits eingeschlossenen EMPA-REG-OUTCOME-Studie (Zinman et al. 2015, NCT01131676); die renalen Daten dieser Kohorte sind im Hauptdatensatz bereits erfasst.
3. **Design- und Größenkriterium (nach SMART-C- und Khan/Usman-2024-Protokoll):** Stichprobengröße N = 530 (< 1.000), 90-Tage-Follow-up und hierarchischer Win-Ratio-Primärendpunkt statt ereignisgesteuerter Time-to-Event-Überlebenszeitanalyse.

Darüber hinaus dokumentiert das Ausschlussprotokoll zwei Beobachtungsstudien als designbedingte Ausschlüsse: die Registerstudie CVD-REAL (Kosiborod et al. 2017; retrospektive Register-/Beobachtungsstudie mit Propensity-Score-Matching, keine Randomisierung) und die Versicherungsdatenkohorte EMPRISE (Patorno et al. 2019; nicht-randomisierte Real-World-Kohortenstudie aus US-Versicherungsdaten). Die Zuordnung dieser beiden Datensätze zur Screening-Ebene (Titel/Abstract versus Volltext) im Zählwerk des Audit-Trails ist

[PLATZHALTER]; sie sind in der Gesamtbilanz der Ausschlüsse enthalten.

Eine Beschaffungslücke bei Volltexten bestand nicht: Von 18 vorgesehenen Volltexten wurden 18 geprüft, 0 waren nicht beschaffbar. Dort, wo in diesem Projekt überhaupt eine Beschaffungsgrenze auftreten könnte, wäre sie eine Werkzeuggrenze der Literaturbeschaffung und ausdrücklich kein methodisches Einschlusskriterium.

3.5 Datenextraktion und Codierung

Extrahiert wurden Studienakronym, Erstautor:in und Publikationsjahr, Registernummer, Interventionssubstanz und Dosisschema, Komparator, Population und Krankheitsentität, Fallzahl pro Arm und insgesamt, geografische Breite (Anzahl teilnehmender Länder), Förderquelle, Endpunktdefinitionen, Ereigniszahlen pro Arm, berichtete Effektschätzer mit Konfidenzintervallen sowie die zur Effektgrößenberechnung genutzte Quelle. Die Anzahl der extrahierenden Personen und das Verfahren der Doppelextraktion mit Abgleich: [PLATZHALTER]. Mittlere oder mediane Nachbeobachtungszeit, Altersverteilung, Geschlechterverteilung, Ausgangs-eGFR, Ausgangs-HbA1c und Ejektionsfraktion wurden erhoben, sind im vorliegenden Analyse-Snapshot jedoch nicht als aggregierte Kennwerte ausgewiesen: [PLATZHALTER].

Für die Moderatoranalysen wurden folgende Variablen codiert:

- `otype` — Endpunkttyp mit fünf Ausprägungen: CV Death or HHF, First HHF, kardiovaskulärer Tod, Gesamtmortalität, Nierenkrankheitsprogression.
- `iclass` — Krankheitsentität mit vier Ausprägungen: Typ-2-Diabetes/ASCVD, chronische Niereninsuffizienz (CKD), Herzinsuffizienz (HFrEF/HFpEF), akuter Myokardinfarkt (AMI).
- `timepoint` — Evidenzschicht: „SMART-C-Kohorte (2015–2023, k = 13, N = 90.413)“ versus „Post-MI-Erweiterung (2024, k = 2, N = 10.539)“.
- `industry` — Substanz und Dosisschema (sieben Ausprägungen, siehe Ergebnisteil).
- `funding` — Förderquelle (sieben Ausprägungen).
- `country` — geografische Breite, operationalisiert über die Anzahl teilnehmender Länder beziehungsweise die namentliche Länderkombination bei register-eingebetteten Studien.
- `year_cont` — Publikationsjahr als kontinuierlicher Moderator.

Der Datensatz enthält laut Audit-Trail 15 eingeschlossene Studien mit 322 Effektgrößen; die der Analyse zugrunde liegende Effektgrößentabelle weist 161 Effektgrößen aus 15 Studien in den Metriken HR, Risk Ratio und Risikodifferenz aus. Die Differenz zwischen beiden Zählungen ist eine Zählkonvention des Projektsystems (Erfassungs- versus Analyseeinheiten) und wird hier transparent ausgewiesen; alle berichteten Modelle beziehen sich auf die analysierte Effektgrößentabelle. Die genaue Zuordnungsregel zwischen beiden Zählungen: [PLATZHALTER].

3.6 Effektmaße und Transformationen

Drei Metriken wurden verwendet; die Wahl richtete sich nach Endpunkt und Quellenlage.

Hazard Ratio (HR). Gerechnet wurde auf der Log-Skala (log-HR). Je nach Quellenlage wurde der log-HR gewonnen (a) direkt aus dem berichteten HR mit Konfidenzintervall, (b) aus O−E und V der Log-Rank-Statistik nach dem Peto-Verfahren, (c) aus dem Log-Rank-p-Wert in Verbindung mit den Ereigniszahlen, (d) aus der rekonstruierten Kaplan-Meier-Kurve nach dem Verfahren von Tierney et al. (2007) oder (e) aus Median-Überlebenszeiten unter Exponentialannahme. Der jeweils benutzte Weg ist je Effektgröße im Datensatz dokumentiert, sodass die Rekonstruktionstiefe pro Schätzer nachvollziehbar bleibt.

Risk Ratio (RR). Gerechnet als natürlicher Logarithmus des Risk Ratio mit der Varianz $1/a - 1/(a+b) + 1/c - 1/(c+d)$, wobei a und c die Ereigniszahlen und b und d die Nicht-Ereigniszahlen in Interventions- und Kontrollarm bezeichnen. Für Nullzellen gilt dieselbe Regel wie beim Odds Ratio.

Risikodifferenz (RD). Gerechnet als p(Intervention) − p(Kontrolle) mit der Varianz $p_1(1-p_1)/n_1 + p_2(1-p_2)/n_2$, ergänzt um die Number needed to treat als $1/|RD|$. Einzelne Nullzellen werden ausdrücklich **nicht** korrigiert, weil ein Anteil von 0 ein gültiger Anteil ist; die 0,5-Korrektur greift ausschließlich dann, wenn in beiden Armen

vollständige Ereignisfreiheit oder Ereignissättigung vorliegt und die Varianz exakt 0 wird. Das Konfidenzintervall der NNT folgt der Konvention von Altman (1998): Schließt das Intervall der Risikodifferenz die Null ein, wird das NNT-Intervall als nicht zusammenhängend ausgewiesen statt als durchgehende Spanne.

Diese doppelte Rechnung — relativ auf der Log-Skala, absolut auf der RD-Skala — ist bewusst gewählt. Relative Maße sind über Populationen mit stark unterschiedlichem Ausgangsrisiko besser übertragbar; absolute Maße beantworten die klinisch und gesundheitsökonomisch relevante Frage nach dem Ausmaß des Nutzens in der jeweiligen Population.

3.7 Statistische Modelle

Hauptmodell. Für den primären Endpunkt (CV Death or HHF, HR) wurde ein Drei-Ebenen-Modell mit robuster Varianzschätzung gerechnet (`metafor::rma.mv` mit Zufallseffekten für Studie und für Effektgröße innerhalb Studie; robuste Standardfehler über `clubSandwich` mit CR2-Korrektur und Satterthwaite-Freiheitsgraden). Die drei Ebenen sind: (1) Stichprobenvarianz der einzelnen Effektgröße, (2) Varianz zwischen Effektgrößen innerhalb einer Studie, (3) Varianz zwischen Studien. Dieses Modell wurde gewählt, weil aus jeder Studie mehrere Endpunkte und Metriken stammen und die daraus resultierende Abhängigkeit in einem Zwei-Ebenen-Modell nicht abgebildet werden kann. Die CR2-Korrektur macht die Inferenz robust gegenüber Fehlspezifikationen der Kovarianzstruktur und ist bei kleiner Clusterzahl (hier k = 15 Studien) der Sandwich-Schätzung ohne Kleinstichprobenkorrektur vorzuziehen.

Referenzmodell. Ergänzend wurde ein klassisches Zwei-Ebenen-Random-Effects-Modell mit REML-Schätzung der Between-Study-Varianz auf den 15 studienweisen HR-Schätzern gerechnet. Dieses Modell dient der Vergleichbarkeit mit den Vorarbeiten, die mit Studienaggregaten arbeiten, und als Konvergenzprüfung des Hauptmodells.

Heterogenität. Berichtet werden τ^2 je Varianzkomponente, das Gesamt-I^2 sowie dessen Zerlegung in einen Studien- und einen Effektgrößen-Anteil, Cochrans Q mit p-Wert im Zwei-Ebenen-Modell und das Prädiktionsintervall. Das Prädiktionsintervall wird als Schätzbereich für den Effekt in einer künftigen, vergleichbaren Studie interpretiert, nicht als Konfidenzintervall des Mittelwerts.

Subgruppen und Meta-Regression. Moderatoranalysen wurden als Modelle mit kategorialen Prädiktoren im Drei-Ebenen-Rahmen gerechnet; die Gruppenunterschiede wurden mit Wald-Tests (robuste Kovarianzmatrix) geprüft, im Zwei-Ebenen-Referenzmodell zusätzlich über die Q-between-Statistik. Als kontinuierlicher Moderator wurde das Publikationsjahr (`year_cont`) in einer Meta-Regression geprüft. Alle Moderatoranalysen sind als hypothesengenerierend zu lesen; eine formale Korrektur für multiples Testen über die sechs Moderatorfamilien wurde nicht vorgenommen: [PLATZHALTER — Entscheidung zur Multiplizitätskorrektur].

Sensitivitätsanalysen. (a) Leave-one-out über alle 15 Studien; (b) Vergleich des Drei-Ebenen-Modells mit dem Zwei-Ebenen-Referenzmodell; (c) getrennte Schätzung auf der Log-Skala und der RD-Skala; (d) Vergleich der Evidenzschichten SMART-C-Kohorte versus Post-MI-Erweiterung.

3.8 Prüfung auf Publikationsbias und Small-Study-Effekte

Geprüft wurde mit fünf Verfahren: Egger-Regressionstest auf Funnel-Asymmetrie; Trim-and-Fill nach Duval und Tweedie mit Ausweis der Zahl imputierter Studien (k_0) und des adjustierten Schätzers; PET-PEESE (berichtet wird der PEESE-Schätzer); p-uniform*; sowie p-Curve mit Prüfung der Rechtssteilheit. Für alle Verfahren, die eine Unabhängigkeitsannahme voraussetzen, wurde auf Studienaggregate gerechnet, das heißt auf einen Schätzer pro Studie; dies ist methodisch notwendig, schränkt aber die Reichweite der Aussage auf die Studienebene ein. Ergänzend wird das Trichter-Diagramm visuell beurteilt (Abbildung 3).

3.9 Risk-of-Bias-Bewertung und Evidenzsicherheit

Das Verzerrungsrisiko wurde mit dem Cochrane-Instrument RoB 2 für randomisierte Studien beurteilt, über fünf Domänen: Randomisierungsprozess; Abweichungen von den beabsichtigten Interventionen; fehlende Ergebnisdaten; Messung des Endpunkts; Auswahl des berichteten Ergebnisses. Insgesamt wurden 75

Domänen-Einträge vergeben (15 Studien × 5 Domänen), davon 75 final bestätigt. Das Gesamturteil je Studie ergibt sich nach dem RoB-2-Algorithmus aus den Domänenurteilen. Die Anzahl der RoB-Bewertenden und das Verfahren der unabhängigen Doppelbewertung: [PLATZHALTER].

Die Evidenzsicherheit wurde endpunktweise nach GRADE beurteilt, getrennt für die relative (HR beziehungsweise RR) und die absolute Metrik (RD), unter Berücksichtigung von Verzerrungsrisiko, Inkonsistenz, Indirektheit, Ungenauigkeit und Publikationsbias. Die Urteile wurden im Projektsystem ratifiziert und sind mit Begründung dokumentiert.

3.10 Software

Die Berechnungen erfolgten mit der projekteigenen Statistik-Engine EvidenzWerk (TypeScript), die gegen R/metafor mit einer Toleranz von 10^{-6} validiert wurde. Die Drei-Ebenen-Modelle mit robuster Varianzschätzung wurden in R gerechnet (`metafor::rma.mv` in Verbindung mit `clubSandwich`, CR2). Jeder Analyse-Run ist versioniert und mit einem Daten-Snapshot-Hash verknüpft; die im Folgenden berichteten Ergebnisse beziehen sich auf die Runs v1, v3, v5, v6 und v7 vom 21.10.2026 sowie auf frisch gerechnete Analysen auf demselben Datenstand, die mit der Analyse-Ansicht numerisch identisch sind. Sechs weitere gespeicherte Versionen ohne gepooltes Schema wurden für die Berichterstattung des Hauptbefunds übersprungen, da sie keinen poolbaren Schätzer enthalten. Konkrete Paketversionen und Seed-Werte: [PLATZHALTER].

---

4 Ergebnisse

4.1 Suchergebnis und Studienfluss

[Abbildung 1: PRISMA-2020-Flussdiagramm]

Über die vier dokumentierten Suchläufe wurden insgesamt 44 Treffer identifiziert (PubMed/MEDLINE 18, OpenAlex 12, Crossref 8, ClinicalTrials.gov 6). In das Screening-System importiert wurden 36 Datensätze; 8 Datensätze waren als Duplikate markiert und wurden vor der inhaltlichen Bewertung zusammengeführt. Im Titel-/Abstract-Screening wurden alle 36 Datensätze von zwei unabhängigen Reviewer:innen bewertet; 18 wurden ausgeschlossen. Für die verbleibenden 18 Datensätze war eine Volltextprüfung vorgesehen; alle 18 Volltexte konnten beschafft und geprüft werden, es bestand keine Beschaffungslücke (0 nicht beschaffbare Volltexte).

Auf Volltextebene wurden drei Datensätze ausgeschlossen: eine Sekundärauswertung der bereits eingeschlossenen DAPA-HF-Hauptstudie, eine renale Einzelpublikation zur bereits eingeschlossenen EMPA-REG-OUTCOME-Kohorte und eine Studie mit N = 530, 90-Tage-Nachbeobachtung und hierarchischem Win-Ratio-Primärendpunkt, die sowohl das Größen- als auch das Designkriterium verfehlte. Zwei weitere Datensätze sind im Ausschlussprotokoll mit designbedingter Begründung geführt: CVD-REAL (Kosiborod et al. 2017) als retrospektive Register-/Beobachtungsstudie mit Propensity-Score-Matching und EMPRISE (Patorno et al. 2019) als nicht-randomisierte Kohortenstudie aus US-Versicherungsdaten.

Eingeschlossen wurden damit 15 Studien, die vollständig in die Synthese eingingen. Die Interrater-Übereinstimmung über alle 36 bewerteten Datensätze betrug 100 % (Cohens κ = 1,00); zum Datenschnitt bestanden 0 offene Konflikte. Die vollständigen Zählungen stammen aus dem Entscheidungs-Audit-Trail und nicht aus Status-Schnappschüssen der Oberfläche, sodass die Summen der Flussbilanz konsistent sind.

4.2 Charakteristika der eingeschlossenen Studien

Die 15 eingeschlossenen Studien umfassen zusammen N = 100.952 randomisierte Patient:innen. Alle sind multizentrische, doppelblinde, placebokontrollierte, ereignisgesteuerte Phase-III-Outcome-Studien. Tabelle 1 fasst die Charakteristika zusammen; die Angaben zu Substanz und Dosisschema, Krankheitsentität und Förderquelle entstammen der Projektcodierung der Moderatorvariablen.

Tabelle 1: Charakteristika der 15 eingeschlossenen Studien

#	Studie (Akronym)	Erstautor:in, Jahr	Substanz und Dosisschema	Krankheitsentität (`iclass`)	N	Förderquelle (`funding`)	RoB-2-Gesamturteil
1	EMPA-REG OUTCOME	Zinman et al. 2015	Empagliflozin 10 mg / 25 mg	Typ-2-Diabetes / ASCVD	7.020	Industrie (Boehringer Ingelheim & Eli Lilly)	niedriges Risiko
2	CANVAS-Programm	Neal et al. 2017	Canagliflozin 100 mg / 300 mg	Typ-2-Diabetes / ASCVD	10.142	Industrie (Janssen Research & Development)	niedriges Risiko
3	DECLARE-TIMI 58	Wiviott et al. 2019	Dapagliflozin 10 mg	Typ-2-Diabetes / ASCVD	17.160	Industrie (AstraZeneca)	niedriges Risiko
4	CREDENCE	Perkovic et al. 2019	Canagliflozin 100 mg	Chronische Niereninsuffizienz	4.401	Industrie (Janssen Research & Development)	niedriges Risiko
5	DAPA-HF	McMurray et al. 2019	Dapagliflozin 10 mg	Herzinsuffizienz (HFrEF)	4.744	Industrie (AstraZeneca)	niedriges Risiko
6	VERTIS CV	Cannon et al. 2020	Ertugliflozin 5 mg / 15 mg	Typ-2-Diabetes / ASCVD	8.246	Industrie (Merck Sharp & Dohme & Pfizer)	niedriges Risiko
7	EMPEROR-Reduced	Packer et al. 2020	Empagliflozin 10 mg	Herzinsuffizienz (HFrEF)	3.730	Industrie (Boehringer Ingelheim & Eli Lilly)	niedriges Risiko
8	DAPA-CKD	Heerspink et al. 2020	Dapagliflozin 10 mg	Chronische Niereninsuffizienz	4.304	Industrie (AstraZeneca)	niedriges Risiko
9	SOLOIST-WHF	Bhatt et al. 2021	Sotagliflozin 200 mg / 400 mg	Herzinsuffizienz (dekompensiert)	1.222	Industrie (Sanofi & Lexicon Pharmaceuticals)	niedriges Risiko
10	SCORED	Bhatt et al. 2021	Sotagliflozin 200 mg / 400 mg	Chronische Niereninsuffizienz	10.584	Industrie (Sanofi & Lexicon Pharmaceuticals)	niedriges Risiko
11	EMPEROR-Preserved	Anker et al. 2021	Empagliflozin 10 mg	Herzinsuffizienz (HFpEF)	5.988	Industrie (Boehringer Ingelheim & Eli Lilly)	niedriges Risiko
12	DELIVER	Solomon et al. 2022	Dapagliflozin 10 mg	Herzinsuffizienz (HFmrEF/ HFpEF)	6.263	Industrie (AstraZeneca)	niedriges Risiko
13	EMPA-KIDNEY	Herrington et al. 2023	Empagliflozin 10 mg	Chronische Niereninsuffizienz	6.609	Industrie & Akademisch (Boehringer Ingelheim, Eli Lilly, MRC UK)	niedriges Risiko
14	DAPA-MI	James et al. 2024	Dapagliflozin 10 mg	Akuter Myokardinfarkt	4.017	Industrie & Register (AstraZeneca, SWEDEHEART, MINAP)	niedriges Risiko

#	Studie (Akronym)	Erstautor:in, Jahr	Substanz und Dosisschema	Krankheitsentität (`iclass`)	N	Förderquelle (`funding`)	RoB-2-Gesamturteil
15	EMPACT-MI	Butler et al. 2024	Empagliflozin 10 mg	Akuter Myokardinfarkt	6.522	Industrie (Boehringer Ingelheim & Eli Lilly)	niedriges Risiko

Nachbeobachtungsdauer, Altersverteilung, Geschlechteranteil, Ausgangs-eGFR, HbA1c und Ejektionsfraktion je Studie: [PLATZHALTER].

Die Verteilung über die Krankheitsentitäten ist annähernd ausgewogen: vier Studien zu Typ-2-Diabetes/ASCVD, vier zur chronischen Niereninsuffizienz, fünf zur Herzinsuffizienz und zwei zum akuten Myokardinfarkt. Nach Stichprobengröße dominieren die Diabetes-Studien (N = 42.568 in den vier Studien 1, 2, 3 und 6), gefolgt von den CKD-Studien (N = 25.898) und den Herzinsuffizienz-Studien (N = 21.947); die Post-MI-Erweiterung steuert N = 10.539 bei. Die Evidenzschicht „SMART-C-Kohorte (2015–2023)“ umfasst 13 Studien mit N = 90.413, die „Post-MI-Erweiterung (2024)“ zwei Studien mit N = 10.539; zusammen ergibt dies die Gesamtstichprobe von N = 100.952.

Nach Substanz verteilen sich die Studien auf Dapagliflozin (fünf Studien), Empagliflozin (fünf Studien, davon eine mit dem Zwei-Dosis-Schema 10 mg / 25 mg), Canagliflozin (zwei Studien, eine mit 100 mg / 300 mg und eine mit 100 mg), Sotagliflozin (zwei Studien) und Ertugliflozin (eine Studie). Die geografische Breite reicht von acht Ländern bis zu 44 Ländern; eine Studie ist register-eingebettet und auf Schweden und das Vereinigte Königreich begrenzt.

4.3 Gepoolter Haupteffekt

[Abbildung 2: Wald-Diagramm der 15 Studien für den primären Endpunkt CV Death or HHF]

Für den primären kombinierten Endpunkt kardiovaskulärer Tod oder Hospitalisierung wegen Herzinsuffizienz ergab das Drei-Ebenen-Modell mit robuster Varianzschätzung (CR2; Run v1 vom 21.10.2026) eine gepoolte **Hazard Ratio von 0,78 (95 %-KI 0,75–0,82; k = 15)**. Das entspricht einer relativen Risikoreduktion von 22 % (95 %-KI 18 %–25 %) gegenüber Placebo auf identischer leitliniengerechter Hintergrundtherapie.

Das Zwei-Ebenen-Random-Effects-Referenzmodell (REML, frisch gerechnet auf demselben Datenstand) liefert einen numerisch identischen Schätzer: **HR 0,78 (95 %-KI 0,75–0,82; p < 0,001; k = 15)**. Die Übereinstimmung beider Modellfamilien bis auf die berichteten Dezimalstellen ist ein starkes Konvergenzargument: Der Hauptbefund hängt nicht an der Modellierung der Abhängigkeitsstruktur, sondern an der Datenlage selbst.

Auf der Skala der Risikodifferenz (Run v7) beträgt der gepoolte absolute Effekt für denselben Endpunkt **RD = −0,019** (p < 0,001; k = 15 Studien), also 1,9 absolut vermiedene Ereignisse pro 100 behandelte Patient:innen über die jeweiligen Studienlaufzeiten, entsprechend einer **NNT von rund 53**. Das Konfidenzintervall der Risikodifferenz und damit das nach Altman (1998) zu bildende NNT-Intervall: [PLATZHALTER].

4.4 Endpunktspezifische Ergebnisse

Die Moderatorfamilie `otype` prüft, ob sich die fünf Endpunkte systematisch unterscheiden. Der Wald-Test ist in beiden Metriken deutlich signifikant (Log-Skala, Run v6: p < 0,001; RD-Skala, Run v7: p = 0,003). Tabelle 2 stellt die Ergebnisse zusammen.

Tabelle 2: Gepoolte Effekte je Endpunkt (Drei-Ebenen-Modell mit CR2-Varianzschätzung)

Endpunkt	k (Effektgrößen)	Log-Skala β (Run v6)	Rücktransformiert exp(β)	p (Run v6)	RD (Run v7)	NNT ≈ 1/|RD|	p (Run v7)	GRADE
Erste Hospitalisierung wegen Herzinsuffizienz (First HHF)	15	−0,34	0,71	< 0,001	−0,017	59	< 0,001	hoch
Nierenkrankheitsprogression	13	−0,38	0,68	< 0,001	−0,010	100	0,004	hoch
CV Death or HHF (primär)	15	−0,23	0,79	< 0,001	−0,019	53	< 0,001	hoch
Kardiovaskulärer Tod	15	−0,14	0,87	< 0,001	−0,008	125	0,001	hoch
Gesamtmortalität	15	−0,12	0,89	0,001	−0,010	100	< 0,001	hoch

Anmerkung: k bezeichnet in Run v6 und v7 die Zahl der in die jeweilige Subgruppe eingehenden Effektgrößen. Die rücktransformierten Werte exp(β) sind Punktschätzer; zugehörige Konfidenzintervalle auf der rücktransformierten Skala: [PLATZHALTER]. Die Nierenkrankheitsprogression beruht auf 13 Studien (N = 90.413), da die beiden Post-MI-Studien diesen Endpunkt in der hier verwendeten Harmonisierung nicht beisteuern.

Das Muster ist klinisch außerordentlich kohärent und reproduziert die aus den Einzelstudien bekannte Rangfolge. Die stärksten relativen Effekte finden sich bei der Nierenkrankheitsprogression (exp(β) = 0,68) und bei der ersten Herzinsuffizienz-Hospitalisierung (0,71). Der kombinierte primäre Endpunkt liegt erwartungsgemäß dazwischen (0,79), weil er die starke Komponente „Hospitalisierung“ mit der schwächeren Komponente „kardiovaskulärer Tod“ mischt. Die Mortalitätsendpunkte zeigen die kleinsten relativen Effekte (kardiovaskulärer Tod 0,87; Gesamtmortalität 0,89), bleiben aber statistisch robust.

Auf der absoluten Skala kehrt sich die Rangfolge teilweise um: Der größte absolute Nutzen entfällt auf den kombinierten Endpunkt (RD = −0,019; NNT ≈ 53) und auf die Herzinsuffizienz-Hospitalisierung (−0,017; NNT ≈ 59), während der kardiovaskuläre Tod trotz eines relativen Effekts von 0,87 nur eine absolute Reduktion von 0,8 Prozentpunkten erreicht (NNT ≈ 125). Diese Diskrepanz ist keine Inkonsistenz, sondern eine direkte Folge der unterschiedlichen Ereignisraten: Hospitalisierungen sind häufiger als kardiovaskuläre Todesfälle, weshalb derselbe oder sogar ein kleinerer relativer Effekt eine größere absolute Wirkung entfaltet. Für die Nierenkrankheitsprogression gilt das Umgekehrte: Der stärkste relative Effekt (0,68) übersetzt sich wegen der niedrigeren Ereignisrate in eine absolute Reduktion von nur einem Prozentpunkt (NNT ≈ 100) — über die jeweiligen, überwiegend wenige Jahre umfassenden Studienlaufzeiten.

4.5 Heterogenität

Die Heterogenitätsbefunde fallen je nach Modellierungsebene unterschiedlich aus; dieser Unterschied ist informativ und wird hier bewusst nebeneinandergestellt.

Im endpunktübergreifenden Drei-Ebenen-Modell (Run v7) beträgt die Gesamtheterogenität **I^2 = 77,7 %**, zerlegt in **30,7 % auf der Studien-Ebene** und **46,9 % auf der Effektgrößen-Ebene**. Die Varianzkomponenten selbst sind numerisch sehr klein: τ^2 (Studie) = 0,0000 und τ^2 (Effekt) = 0,0000 auf vier Dezimalstellen. Dieser scheinbare Widerspruch — hohes I^2 bei τ^2 nahe null — ist eine bekannte Eigenschaft des I^2-Maßes auf der Skala der Risikodifferenz: I^2 setzt die geschätzte Heterogenitätsvarianz ins Verhältnis zur typischen Stichprobenvarianz. Bei sehr großen Studien (hier bis zu N = 17.160) ist die Stichprobenvarianz der Risikodifferenzen extrem klein, sodass selbst absolut winzige Unterschiede zwischen Studien und zwischen Endpunkten einen hohen I^2-Wert

erzeugen. Inhaltlich bedeutet I^2 = 77,7 % in diesem Modell daher **nicht**, dass die Effekte klinisch widersprüchlich wären, sondern dass die Präzision der Einzelschätzer so hoch ist, dass auch kleine, systematische Unterschiede statistisch sichtbar werden. Der größere Anteil der Heterogenität liegt zudem auf der Effektgrößen-Ebene (46,9 %) und nicht zwischen den Studien (30,7 %) — das heißt: Die Unterschiede zwischen den fünf Endpunkten innerhalb derselben Studien übertreffen die Unterschiede zwischen den Studien. Genau diese Interpretation wird durch den signifikanten Wald-Test der Moderatorfamilie `otype` (p = 0,003) bestätigt.

Das **Prädiktionsintervall** auf der RD-Skala beträgt **[−0,032; 0,007]**. Es schließt die Null ein. Für eine künftige, vergleichbare Studie wäre also eine absolute Risikoreduktion von bis zu 3,2 Prozentpunkten, aber auch ein kleiner absoluter Nachteil von bis zu 0,7 Prozentpunkten mit den Daten vereinbar — wobei dieses Intervall über alle fünf Endpunkte und alle vier Krankheitsentitäten gemittelt ist und damit die Spannweite sehr unterschiedlicher Ausgangsrisiken enthält. Es ist deshalb nicht als Aussage über die Wirksamkeit in einer bestimmten Indikation zu lesen.

Im Zwei-Ebenen-Referenzmodell auf den 15 studienweisen HR-Schätzern für den primären Endpunkt zeigt sich dagegen **vollständige Homogenität: I^2 = 0 %, Q = 14,1 (p = 0,446), τ^2 = 0,0000**, mit einem Prädiktionsintervall von **0,75–0,82**, das mit dem Konfidenzintervall des gepoolten Schätzers zusammenfällt. Das ist ein bemerkenswert eindeutiger Befund: Für den primären kombinierten Endpunkt verhalten sich 15 Studien über fünf Substanzen, vier Krankheitsentitäten und neun Publikationsjahre hinweg statistisch wie Replikationen desselben Effekts. Die Null-Heterogenität auf Studienebene bei gleichzeitig hoher Heterogenität im endpunktübergreifenden Modell lässt sich damit präzise lokalisieren: Die Variabilität steckt in den Endpunkten, nicht in den Studien.

4.6 Subgruppen nach Krankheitsentität

Die Moderatorfamilie `iclass` ist die klinisch zentrale Frage dieser Arbeit: Wirkt die Substanzklasse über das Spektrum hinweg gleich? Tabelle 3 fasst die Befunde aus drei Runs zusammen.

Tabelle 3: Effekte nach Krankheitsentität

Krankheitsentität	Run v3 (Studienebene, log)	exp(β)	p (v3)	Run v6 (Effektebene, log)	exp(β)	p (v6)	Run v7 (RD)	p (v7)	Zwei-Ebenen-HR (frisch)	k (Studien)
Chronische Niereninsuffizienz (CKD)	−0,29	0,75	0,008	−0,28	0,76	0,011	−0,014	0,034	0,75	4
Herzinsuffizienz (HFrEF / HFpEF)	−0,25	0,78	< 0,001	−0,20	0,82	0,001	−0,018	0,002	0,78	5
Typ-2-Diabetes / ASCVD	−0,23	0,79	0,021	−0,25	0,78	0,019	−0,011	0,026	0,79	4
Akuter Myokardinfarkt (AMI)	−0,11	0,90	0,143	−0,07	0,93	0,332	−0,003	0,494	0,90	2
Test auf Gruppenunterschied	Wald p = 0,157			Wald p = 0,268			Wald p = 0,144		Q-between = 3,94; p = 0,268	

Anmerkung: In den Runs v6 und v7 bezeichnet k die Zahl der Effektgrößen (AMI 8, CKD 20, Herzinsuffizienz 25, Typ-2-Diabetes/ASCVD 20); in Run v3 und im Zwei-Ebenen-Modell die Zahl der Studien.

Drei Beobachtungen sind hervorzuheben.

Erstens ist die **Konsistenz über drei der vier Entitäten bemerkenswert eng**: CKD 0,75, Herzinsuffizienz 0,78 und Typ-2-Diabetes/ASCVD 0,79 liegen innerhalb einer Spanne von vier Prozentpunkten relativer Risikoreduktion. Alle drei sind in sämtlichen Modellspezifikationen statistisch signifikant. Das stützt die These, dass die SGLT2-Inhibition nicht an einen spezifischen Krankheitsmechanismus gebunden ist, sondern an einen gemeinsamen kardiorenalen Pfad angreift.

Zweitens fällt die **Post-Myokardinfarkt-Population numerisch ab** (HR 0,90 im Zwei-Ebenen-Modell; exp(β) = 0,90 in v3, 0,93 in v6; RD = −0,003) und erreicht in keiner Spezifikation statistische Signifikanz (p = 0,143 in v3; p = 0,332 in v6; p = 0,494 in v7). Dieses Muster ist mit mehreren Erklärungen vereinbar und lässt sich aus den vorliegenden Daten nicht auflösen: geringe Zahl an Studien (k = 2) und damit geringe Präzision; deutlich niedrigeres Ausgangsrisiko für Herzinsuffizienz-Ereignisse in einer Population, die definitionsgemäß zum Einschlusszeitpunkt keine manifeste Herzinsuffizienz aufweist; kürzere Nachbeobachtung; sowie möglicherweise ein echter Wirksamkeitsunterschied in der frühen Postinfarktphase. Die niedrige absolute Risikodifferenz (−0,003; NNT ≈ 333) ist dabei der wichtigste Hinweis: Wo wenige Ereignisse auftreten, kann selbst ein mittlerer relativer Effekt kaum absoluten Nutzen erzeugen.

Drittens — und methodisch entscheidend — ist der **Test auf Gruppenunterschiede in keiner Spezifikation signifikant** (Wald p = 0,157 / 0,268 / 0,144; Q-between = 3,94, p = 0,268). Die Daten liefern also keinen statistischen Beleg dafür, dass sich die Effekte zwischen den Entitäten tatsächlich unterscheiden. Formal korrekt ist daher die Aussage: Die Punktschätzer variieren, aber die Variation ist mit Zufall vereinbar. Gleichzeitig gilt die Umkehrung nicht — die Nicht-Signifikanz eines Interaktionstests bei k = 2 Studien in der AMI-Schicht ist kein Nachweis von Gleichheit. Beides muss in der Interpretation nebeneinander stehen.

4.7 Evidenzschichten: SMART-C-Kohorte versus Post-MI-Erweiterung

Die Moderatorfamilie `timepoint` trennt die 13 Studien der SMART-C-Abgrenzung (2015–2023, N = 90.413) von den beiden 2024 publizierten Post-MI-Studien (N = 10.539). Die Befunde spiegeln erwartungsgemäß die `iclass`-Analyse, weil die Schichtung weitgehend deckungsgleich ist:

Evidenzschicht	Run v3 (log)	exp(β)	p	Run v6 (log)	exp(β)	p	Run v7 (RD)	p
SMART-C-Kohorte (2015–2023; k = 13 Studien, 65 Effektgrößen)	−0,25	0,78	< 0,001	−0,24	0,79	< 0,001	−0,014	< 0,001
Post-MI-Erweiterung (2024; k = 2 Studien, 8 Effektgrößen)	−0,11	0,90	0,143	−0,07	0,93	0,332	−0,003	0,495
Test auf Gruppenunterschied	Wald p = 0,110			Wald p = 0,148			Wald p = 0,113	

Auch hier ist der Unterschied nicht statistisch gesichert (p zwischen 0,110 und 0,148 je nach Spezifikation). Die praktische Konsequenz ist jedoch relevant: Der gut abgesicherte Teil der Evidenz — 13 Studien, 90.413 Patient:innen, HR 0,78 — betrifft die chronischen kardiometabolischen Indikationen. Die Ausweitung auf die akute Postinfarktphase steht auf einer schmaleren Basis und zeigt dort einen kleineren, nicht signifikanten Effekt.

4.8 Subgruppen nach Substanz und Dosisschema

Die Moderatorfamilie `industry` codiert Wirkstoff und Dosisschema. Tabelle 4 fasst die Ergebnisse zusammen.

Tabelle 4: Effekte nach Substanz und Dosisschema

Substanz / Dosisschema	Run v3 (log)	exp(β)	p (v3)	k (v3, Studien)	Run v6 (log)	exp(β)	p (v6)	k (v6, Effektgrößen)
Empagliflozin 10 mg / 25 mg	−0,42	0,66	n. b.	1	−0,43	0,65	0,026	5
Canagliflozin 100 mg	−0,37	0,69	n. b.	1	−0,32	0,73	0,011	5
Sotagliflozin 200 mg / 400 mg	−0,29	0,75	0,084	2	−0,23	0,79	0,068	10
Canagliflozin 100 mg / 300 mg	−0,25	0,78	n. b.	1	−0,23	0,79	0,071	5
Dapagliflozin 10 mg	−0,22	0,80	0,007	5	−0,22	0,80	0,014	24
Empagliflozin 10 mg	−0,22	0,80	0,015	4	−0,17	0,84	0,006	19
Ertugliflozin 5 mg / 15 mg	−0,13	0,88	n. b.	1	−0,16	0,85	0,207	5
Test auf Gruppenunterschied	—				Wald p = 0,805			

n. b. = nicht berichtet (kein p-Wert für Einzelstudien-Schichten mit k = 1 im gespeicherten Run v3 ausgewiesen).

Der Wald-Test auf Unterschiede zwischen den Substanzschichten ist mit p = 0,805 deutlich nicht signifikant. Die Spannweite der Punktschätzer (0,65 bis 0,88 in Run v6) ist zwar augenscheinlich breit, aber sie ist fast vollständig mit der Studienzahl pro Schicht konfundiert: Die extremen Werte stammen jeweils aus Schichten mit nur einer Studie (Empagliflozin 10/25 mg aus der Diabetes-Studie mit dem größten berichteten Mortalitätseffekt; Ertugliflozin aus der einzigen Ertugliflozin-Studie). Die beiden am besten besetzten Schichten — Dapagliflozin 10 mg (fünf Studien, 24 Effektgrößen) und Empagliflozin 10 mg (vier Studien, 19 Effektgrößen) — liegen mit exp(β) = 0,80 beziehungsweise 0,80 (v3) und 0,84 (v6) eng beieinander und decken sich mit dem Gesamtschätzer.

Diese Analyse erlaubt deshalb **keine** Rangfolge der Substanzen. Jede Schicht ist zugleich eine Indikations- und eine Zeitschicht: Wer Empagliflozin 10/25 mg sagt, sagt zugleich „Typ-2-Diabetes mit ASCVD, 2015“; wer Ertugliflozin sagt, sagt „Typ-2-Diabetes, 2020“. Ein Substanzvergleich wäre nur über direkte Head-to-Head-Studien oder über Netzwerk-Meta-Analysen mit expliziter Modellierung der Indikationsunterschiede möglich; beides liegt außerhalb dieser Arbeit.

Bemerkenswert ist allerdings, dass der duale SGLT1/2-Inhibitor Sotagliflozin (zwei Studien, zehn Effektgrößen; exp(β) = 0,75 in v3 und 0,79 in v6; p = 0,084 beziehungsweise 0,068) sich nicht erkennbar von den selektiven SGLT2-Inhibitoren abhebt. Für die Frage, ob die zusätzliche SGLT1-Hemmung klinisch relevante inkrementelle Effekte erzeugt, liefern diese Daten keinen Anhalt in die eine oder andere Richtung.

4.9 Subgruppen nach Förderquelle und geografischer Breite

Die Moderatorfamilie `funding` wurde geprüft, weil Industriefinanzierung ein klassischer Verdachtsmoment für verzerrte Effektschätzung ist. Alle 15 Studien sind industriefinanziert oder industrie-ko-finanziert; eine Schicht ohne Industriebeteiligung existiert nicht, weshalb ein Vergleich „Industrie versus unabhängig“ konstruktionsbedingt unmöglich ist.

Tabelle 5: Effekte nach Förderquelle (Run v3, Studienebene)

Förderquelle	β (log)	exp(β)	p	k (Studien)
Industrie (Janssen Research & Development)	−0,30	0,74	0,127	2
Industrie (Sanofi & Lexicon Pharmaceuticals)	−0,29	0,75	0,084	2
Industrie (Boehringer Ingelheim & Eli Lilly)	−0,26	0,77	0,013	4
Industrie (AstraZeneca)	−0,23	0,79	0,009	4
Industrie & Akademisch (Boehringer Ingelheim, Eli Lilly, MRC UK)	−0,17	0,84	n. b.	1
Industrie (Merck Sharp & Dohme & Pfizer)	−0,13	0,88	n. b.	1
Industrie & Register (AstraZeneca, SWEDEHEART, MINAP)	−0,05	0,95	n. b.	1

Die beiden kleinsten Effekte entfallen auf die einzige Studie mit akademischer Ko-Förderung (Medical Research Council UK; exp(β) = 0,84) und auf die einzige register-eingebettete Studie (exp(β) = 0,95). Beides ist jedoch mit der Indikation konfundiert: Die MRC-ko-finanzierte Studie ist eine CKD-Studie mit breitem Einschluss, die register-eingebettete Studie eine Post-MI-Studie. Eine Interpretation im Sinne eines Sponsoring-Bias ist aus diesen Daten nicht ableitbar; ein Wald-Test für diese Moderatorfamilie wurde nicht ausgewiesen: [PLATZHALTER].

Dasselbe gilt für die Moderatorfamilie `country`, die die geografische Breite über die Zahl teilnehmender Länder operationalisiert (Tabelle 6).

Tabelle 6: Effekte nach geografischer Breite (Run v3, Studienebene)

Geografische Breite	β (log)	exp(β)	p	k (Studien)
Multizentrisch (42 Länder)	−0,42	0,66	n. b.	1
Multizentrisch (21 Länder)	−0,34	0,71	n. b.	1
Multizentrisch (32 Länder)	−0,34	0,71	n. b.	1
Multizentrisch (44 Länder)	−0,26	0,77	n. b.	1
Multizentrisch (20 Länder)	−0,25	0,78	0,016	3
Multizentrisch (30 Länder)	−0,25	0,78	n. b.	1
Multizentrisch (23 Länder)	−0,24	0,79	n. b.	1
Multizentrisch (34 Länder)	−0,23	0,79	0,303	2
Multizentrisch (33 Länder)	−0,19	0,83	n. b.	1
Multizentrisch (8 Länder)	−0,17	0,84	n. b.	1
Multizentrisch (22 Länder)	−0,12	0,89	n. b.	1

Geografische Breite	β (log)	exp(β)	p	k (Studien)
Schweden & Vereinigtes Königreich (register-eingebettet)	−0,05	0,95	n. b.	1

Diese Moderatorfamilie ist faktisch eine Studienkennung mit zwölf Ausprägungen für 15 Studien und daher statistisch nicht auswertbar; sie wird hier aus Transparenzgründen vollständig berichtet, nicht zur Hypothesenprüfung. Ein inhaltlich belastbarer Befund ist ihr nicht zu entnehmen. Festhalten lässt sich lediglich, dass die Studien durchweg eine hohe internationale Breite aufweisen (acht bis 44 Länder), was der Übertragbarkeit der Befunde zugutekommt, und dass die einzige auf zwei Länder begrenzte, register-eingebettete Studie den kleinsten Effekt zeigt — was wiederum mit ihrer Indikation (Post-Myokardinfarkt) konfundiert ist.

4.10 Meta-Regression auf das Publikationsjahr

Die Meta-Regression mit dem Publikationsjahr als kontinuierlichem Moderator (`year_cont`, Run v7) ergab eine **Steigung von 0,002 pro Jahr auf der Log-Skala (p = 0,043)**. Das Vorzeichen ist positiv, der Effekt nimmt also über die Zeit geringfügig ab. Die Größenordnung ist allerdings sehr klein: exp(0,002) = 1,002 pro Jahr, über den gesamten Beobachtungszeitraum von 2015 bis 2024 also eine relative Abschwächung von rund 1,8 %.

Drei Lesarten sind zu erwägen und lassen sich mit diesen Daten nicht trennen. Erstens eine echte zeitliche Verdünnung: Mit zunehmender Optimierung der Hintergrundtherapie — Betablocker, Renin-Angiotensin-Aldosteron-System-Hemmung, Angiotensin-Rezeptor-Neprilysin-Inhibition, Mineralokortikoidrezeptor-Antagonisten, moderne Lipidsenkung und Revaskularisationsstrategien — sinkt die Ereignisrate im Placeboarm, und der inkrementelle Nutzen einer zusätzlichen Substanz wird kleiner. Zweitens eine Konfundierung mit der Indikation: Die spätesten Studien sind die Post-MI-Studien mit dem kleinsten Effekt; der Zeittrend könnte schlicht die `iclass`-Verteilung über die Zeit abbilden. Drittens ein statistisches Artefakt: Bei k = 15 Studien und einem p-Wert von 0,043 ist der Befund fragil und würde jede Korrektur für multiples Testen über die sechs geprüften Moderatorfamilien nicht überstehen. Die Meta-Regression ist daher als Hinweis, nicht als Nachweis zu lesen.

4.11 Sensitivitätsanalysen

Leave-one-out. Der gepoolte Schätzer für den primären Endpunkt bewegt sich beim sequenziellen Ausschluss jeder einzelnen Studie zwischen **HR 0,78 und 0,79**. Keine Einzelstudie — auch nicht die größte (N = 17.160) oder die kleinste (N = 1.222) — trägt den Befund. Die Stabilitätsspanne von einem Hundertstel auf der HR-Skala ist bei k = 15 ungewöhnlich eng und spiegelt die bereits festgestellte Null-Heterogenität auf Studienebene.

Modellvergleich. Drei-Ebenen-Modell mit CR2-Varianzschätzung und Zwei-Ebenen-Random-Effects-Modell mit REML liefern identische Punktschätzer und identische Konfidenzintervalle (0,78; 95 %-KI 0,75–0,82). Die Robustheit gegenüber der Modellwahl ist damit belegt.

Metrikvergleich. Die relativen und absoluten Metriken führen zu kohärenten, aber nicht deckungsgleichen Rangfolgen der Endpunkte (siehe Abschnitt 4.4). Diese Diskrepanz ist erwartbar und interpretierbar; sie stellt die Befunde nicht infrage, sondern präzisiert sie.

Schichtvergleich. Der Ausschluss der Post-MI-Erweiterung (Beschränkung auf die SMART-C-Kohorte mit 13 Studien und N = 90.413) ergibt exp(β) = 0,78 (Run v3) beziehungsweise 0,79 (Run v6) und RD = −0,014, jeweils p < 0,001 — also praktisch unverändert gegenüber dem Gesamtschätzer.

4.12 Risk of Bias

Alle 15 eingeschlossenen Studien wurden mit RoB 2 über fünf Domänen bewertet; 75 Domänen-Einträge wurden vergeben und 75 final bestätigt. Die Verteilung der Gesamturteile lautet: **15 Studien mit niedrigem Verzerrungsrisiko, 0 Studien mit einigen Bedenken, 0 Studien mit hohem Verzerrungsrisiko.**

Dieses homogene Ergebnis ist angesichts der Studienpopulation nicht überraschend: Es handelt sich ausnahmslos um große, zentral randomisierte, doppelblinde, ereignisgesteuerte Phase-III-Zulassungs- und Outcome-Studien

mit verblindeter, komiteebasierter Endpunktadjudikation, präspezifizierten statistischen Analyseplänen, öffentlich registrierten Protokollen und sehr niedrigen Raten fehlender Ergebnisdaten. Genau diese Merkmale sind die Einschlusskriterien gewesen. Die Einheitlichkeit der RoB-Urteile ist damit teilweise ein Resultat der Selektion und nicht nur eine Eigenschaft der Literatur — eine Zirkularität, die bei der Interpretation zu beachten ist: Das Instrument differenziert in diesem Datensatz nicht, weil alle Studien dem methodischen Goldstandard entsprechen, den die Einschlusskriterien verlangt haben.

Eine domänenweise Auszählung (wie viele Studien pro Domäne welches Urteil erhielten) ist im vorliegenden Snapshot nur als Gesamtzahl der finalisierten Einträge ausgewiesen: [PLATZHALTER — domänenspezifische Verteilung]. Eine Bewertung des Verzerrungsrisikos auf Endpunktebene, die insbesondere bei weicheren Endpunkten wie Hospitalisierung relevant wäre, liegt in folgender Form vor: [PLATZHALTER].

4.13 Prüfung auf Publikationsbias und Small-Study-Effekte

[Abbildung 3: Trichter-Diagramm für den primären Endpunkt CV Death or HHF]

Fünf Verfahren wurden angewendet; alle fünf sprechen gegen relevante Verzerrung.

Der **Egger-Regressionstest** ergab in der gespeicherten Version (Run v5) eine Steigung von −0,383 (p = 0,741) und in der frisch gerechneten Analyse auf demselben Datenstand einen Intercept von −0,38 (SE 1,02; p = 0,714; k = 15). Es liegt also keine nachweisbare Asymmetrie des Trichter-Diagramms vor; der Standardfehler des Intercepts ist allerdings mit 1,02 so groß, dass auch eine moderate Asymmetrie nicht ausgeschlossen werden könnte — ein typisches Präzisionsproblem bei k = 15.

Trim-and-Fill imputierte $\mathbf{k_0 = 0}$ Studien; der adjustierte Schätzer von **0,785** ist mit dem unadjustierten identisch. Es gibt also keine Lücke im Trichter, die durch hypothetische unpublizierte Studien gefüllt werden müsste.

PET-PEESE ergab im PEESE-Schritt einen Schätzer von **0,786**, also ebenfalls praktisch unverändert gegenüber dem Hauptbefund. Ein durch Small-Study-Effekte aufgeblähter Effekt würde hier eine deutliche Abschwächung in Richtung 1,0 zeigen; das ist nicht der Fall.

p-uniform\ *schätzte den um selektive Publikation bereinigten Effekt auf **0,781** — der niedrigste und damit günstigste der vier Bias-korrigierten Schätzer, abermals im Bereich des Hauptbefunds.*

Die **p-Curve** zeigt bei 11 von 15 signifikanten Effekten (p < 0,05) eine klare **Rechtssteilheit (p < 0,001)**; nach der Logik des Verfahrens ist damit ein substanzieller Evidenzwert vorhanden, das heißt: Die Verteilung der signifikanten p-Werte ist mit einem echten, nicht durch selektive Analyse erzeugten Effekt vereinbar.

Zu beachten ist, dass alle Verfahren mit Unabhängigkeitsannahme auf Studienaggregaten gerechnet wurden, also auf einem Schätzer pro Studie. Die Aussagen gelten damit für die Studienebene und nicht für die 161 Effektgrößen. Inhaltlich ist das Risiko eines Publikationsbias in diesem Feld ohnehin gering: Alle 15 Studien sind präregistrierte, ereignisgesteuerte Zulassungsstudien mit regulatorischer Berichtspflicht, deren Nichtpublikation praktisch ausgeschlossen ist. Die Bias-Analysen bestätigen diese Erwartung, sie konstituieren sie nicht.

4.14 Evidenzsicherheit nach GRADE

Die Evidenzsicherheit wurde für alle fünf Endpunkte und für beide Metriken als **hoch (⊕⊕⊕⊕)** ratifiziert. Die dokumentierten Begründungen sind:

- **Erste Hospitalisierung wegen Herzinsuffizienz** — Risk Ratio: hohe Evidenzsicherheit über k = 15 doppelblinde Phase-III-Studien (N = 100.952); Risikodifferenz: hohe Evidenzsicherheit mit direkter absoluter Risikoreduktion und NNT aus den 2×2-Ereignistafeln (N = 100.952).
- **Gesamtmortalität** — Risk Ratio und Risikodifferenz: hohe Evidenzsicherheit auf derselben Basis (k = 15; N = 100.952).
- **Kardiovaskulärer Tod** — Risk Ratio und Risikodifferenz: hohe Evidenzsicherheit auf derselben Basis (k = 15; N = 100.952).
- **Kardiovaskulärer Tod oder Herzinsuffizienz-Hospitalisierung** — Hazard Ratio: hohe Evidenzsicherheit

analog zum Vorgehen des SMART-C-Konsortiums (The Lancet 2022) und von Khan, Usman, Butler, Bhatt et al. (Lancet Diabetes & Endocrinology 2024); Risikodifferenz: hohe Evidenzsicherheit mit direkter absoluter Risikoreduktion und NNTB aus den 2×2-Ereignistafeln (N = 100.952).

- **Nierenkrankheitsprogression** — Risk Ratio: hohe Evidenzsicherheit über k = 13 doppelblinde Phase-III-Studien (N = 90.413); Risikodifferenz: hohe Evidenzsicherheit mit direkter absoluter Risikoreduktion und NNT aus den 2×2-Ereignistafeln (N = 90.413).

Für alle Endpunkte gilt damit die GRADE-Formulierung: Wir sind sehr zuversichtlich, dass der wahre Effekt nahe beim geschätzten Effekt liegt. Eine Abwertung wegen Inkonsistenz erfolgte nicht, weil die Heterogenität auf Studienebene null beträgt (I^2 = 0 %; Q = 14,1; p = 0,446) und die im endpunktübergreifenden Modell gemessene Heterogenität (I^2 = 77,7 %) durch die Endpunktstruktur erklärt und durch die endpunktweise Schätzung bereits adressiert ist. Eine Abwertung wegen Ungenauigkeit erfolgte nicht, weil die Konfidenzintervalle eng sind und die Fallzahlen die konventionellen Schwellen der optimalen Informationsgröße weit überschreiten. Eine Abwertung wegen Indirektheit erfolgte nicht, weil Population, Intervention, Komparator und Endpunkte der Fragestellung direkt entsprechen. Eine Abwertung wegen Publikationsbias erfolgte nicht, weil alle fünf Prüfverfahren unauffällig sind.

Anzumerken ist, dass die GRADE-Urteile für die **AMI-Subgruppe** nicht separat ratifiziert wurden; angesichts von k = 2 Studien, fehlender Signifikanz und einer RD von −0,003 wäre dort eine Abwertung wegen Ungenauigkeit zu erwägen: [PLATZHALTER — subgruppenspezifische GRADE-Bewertung].

---

5 Forschungslage und Einordnung in die Literatur

5.1 Vergleichstabelle

Tabelle 7 stellt alle fünf im Screening identifizierten Vorarbeiten dar, die als Hintergrund- und Vergleichsquellen (nicht als Primärberichte eingeschlossener Studien) genutzt wurden. Die Gruppierung erfolgt nach der Funktion der Arbeit im Diskurs: zunächst die mechanistischen und narrativen Arbeiten, die den Wirkmechanismus und Nebenendpunkte erschließen, danach die quantitativen Evidenzsynthesen in chronologischer Reihenfolge.

Tabelle 7: Vergleichsbasis — alle fünf identifizierten Vorarbeiten

Autor:innen / Jahr	Design und Methodik	Stichprobe	Berichtete Effektgröße	Unterschied zur vorliegenden Arbeit
Gruppe A — Mechanistische und narrative Arbeiten				
Verma S, McMurray JJV (2018) DOI: 10.1007/s00125-018-4670-7	Mechanistische Übersichtsarbeit zu Wirkpfaden der SGLT2-Inhibition bei Herzinsuffizienz (Natriurese, Ketonkörperstoffwechsel, myokardiale Energetik); keine quantitative Synthese	Nicht berichtet	Nicht berichtet (keine gepoolte Effektschätzung)	Die vorliegende Arbeit liefert die quantitative Gegenprobe zu den dort formulierten Mechanismushypothesen: Sie prüft empirisch, ob die postulierte Mechanismus-Unabhängigkeit von Glukosestoffwechsel und Ejektionsfraktion sich in konsistenten Effektschätzern über vier Krankheitsentitäten niederschlägt (HR 0,75–0,90; Interaktion nicht signifikant), liefert aber selbst keine mechanistischen Daten

Autor:innen / Jahr	Design und Methodik	Stichprobe	Berichtete Effektgröße	Unterschied zur vorliegenden Arbeit
Zelniker TA et al. (2020) DOI: 10.1161/circulationaha.119.044231	Post-hoc-Synthese zum Risiko von Vorhofflimmern unter SGLT2-Inhibition; narrativer Syntheseansatz	Nicht berichtet	Nicht berichtet	Unterschiedlicher Endpunktfokus: Die vorliegende Arbeit untersucht ausschließlich fünf harte kardiovaskuläre und renale Endpunkte (CV Death or HHF, First HHF, kardiovaskulärer Tod, Gesamtmortalität, Nierenkrankheitsprogression) und keine arrhythmischen Endpunkte; Vorhofflimmern ist hier nicht codiert
Gruppe B — Quantitative Evidenzsynthesen				
Nuffield Department of Population Health Renal Studies Group & SMART-C Consortium (2022) DOI: 10.1016/s0140-6736(22)02292-8	Kollaborative Meta-Analyse großer placebokontrollierter Studien mit harmonisierten renalen Endpunktdefinitionen; Leitfrage: Modifiziert der Diabetesstatus die renalen Effekte?	Nicht berichtet	Nicht berichtet	Fokus und Moderator unterscheiden sich: Dort steht der Diabetesstatus als Effektmodifikator der **Nierenendpunkte** im Zentrum; hier steht der kombinierte kardiovaskuläre Endpunkt im Zentrum, die Nierenkrankheitsprogression ist einer von fünf Endpunkten (exp(β) = 0,68; RD = −0,010; k = 13; N = 90.413), und die Moderation wird über die Krankheitsentität (`iclass`), nicht über den Diabetesstatus geprüft. Methodisch: hier Drei-Ebenen-Modell mit CR2-Varianzschätzung über fünf Endpunkte und drei Metriken statt endpunktspezifischer Zwei-Ebenen-Poolung

Autor:innen / Jahr	Design und Methodik	Stichprobe	Berichtete Effektgröße	Unterschied zur vorliegenden Arbeit
Vaduganathan M, Docherty KF, Claggett BL, Jhund PS, de Boer RA et al. (2022) DOI: 10.1016/s0140-6736(22)01429-5	Präspezifizierte Meta-Analyse von DELIVER und EMPEROR-Preserved, erweitert um DAPA-HF, EMPEROR-Reduced und SOLOIST-WHF; Trial-Level-Daten mit harmonisierten Endpunktdefinitionen; Ziel: Effekte auf kardiovaskuläre Mortalität und Subgruppen über das Spektrum der Ejektionsfraktion	Fünf randomisierte Studien (DELIVER, EMPEROR-Preserved, DAPA-HF, EMPEROR-Reduced, SOLOIST-WHF); Fallzahl im bereitgestellten Abstract nicht berichtet	Nicht berichtet (im bereitgestellten Abstract keine gepoolten Schätzer ausgewiesen)	Breitere Population: Dieselben fünf Herzinsuffizienz-Studien bilden hier nur eine von vier Krankheitsentitäten (N = 21.947 von N = 100.952). Die vorliegende Arbeit ergänzt vier Diabetes-, vier CKD- und zwei Post-MI-Studien und testet die Entität als expliziten Moderator. Zusätzlich: parallele Schätzung auf der Skala der Risikodifferenz mit NNT, Drei-Ebenen-Modellierung der Abhängigkeitsstruktur, fünf Verfahren zur Bias-Prüfung
Khan MS, Usman MS, Butler J, Bhatt DL et al. (2024) DOI: 10.1016/s2213-8587(24)00102-5	Systematische Übersichtsarbeit und Meta-Analyse zu Herzinsuffizienz-Endpunkten und kardiovaskulärem Tod über das kardiometabolische Krankheitsspektrum	Nicht berichtet	Nicht berichtet	Nächstliegende Vergleichsarbeit und zugleich die inhaltliche Ausgangsbasis: Die Einschlusslogik (N ≥ 1.000, ereignisgesteuert, doppelblind, Publikation bis Mai 2024) wurde bewusst nachgebildet. Unterschiede: (a) Drei-Ebenen-Modell mit robuster CR2-Varianzschätzung statt Zwei-Ebenen-Poolung; (b) fünf statt zwei Endpunktfamilien, einschließlich Gesamtmortalität und Nierenkrankheitsprogression; (c) durchgängige Parallelschätzung relativer und absoluter Effekte mit NNT nach Altman-Konvention; (d) explizite Prüfung der Evidenzschicht „SMART-C-Kohorte versus Post-MI-Erweiterung“ als Moderator; (e) endpunktweise GRADE-Ratifizierung für beide Metriken

Fußnote zur Gruppierungsregel: Die Zuordnung zu Gruppe A oder B erfolgt nach der Syntheseform der Arbeit. Gruppe A umfasst Arbeiten ohne gepoolte Effektschätzung (mechanistische Übersichtsarbeit, narrative Post-hoc-Synthese), Gruppe B Arbeiten mit quantitativer Evidenzsynthese auf Studienebene. Innerhalb der Gruppen ist chronologisch sortiert. Alle fünf im Screening identifizierten Vorarbeiten sind vollständig aufgeführt; es wurde keine gekürzt. Angaben zu Stichprobe, Methodik und berichteten Effektgrößen wurden ausschließlich

den im Projekt bereitgestellten Abstracts entnommen; fehlte eine Angabe dort, ist „nicht berichtet“ eingetragen.

5.2 Einordnung des eigenen Ergebnisses

Die vorliegende Synthese ergänzt den bestehenden Forschungsstand in vier Richtungen, ohne ihn in der Substanz zu korrigieren.

Erstens: Breite statt Tiefe in einer Indikation. Die Arbeit von Vaduganathan et al. (2022) ist eine Tiefenbohrung in die Herzinsuffizienz: Sie wurde ausdrücklich deshalb durchgeführt, weil die Evidenz für höhere Ejektionsfraktionen weniger gut etabliert war als für HFrEF, und sie nutzte das zeitliche Zusammentreffen von DELIVER und EMPEROR-Preserved, um genügend statistische Power für die kardiovaskuläre Mortalität und für Subgruppenanalysen über das Spektrum der Ejektionsfraktion zu gewinnen. Die vorliegende Arbeit verfolgt die komplementäre Strategie: Sie nimmt genau diese fünf Studien (N = 21.947) als eine von vier Entitäten auf und stellt ihnen die vier Diabetes-Studien (N = 42.568), die vier CKD-Studien (N = 25.898) und die beiden Post-MI-Studien (N = 10.539) gegenüber. Das Ergebnis — Herzinsuffizienz 0,78, CKD 0,75, Typ-2-Diabetes/ASCVD 0,79, AMI 0,90, Interaktion nicht signifikant (Q-between = 3,94; p = 0,268) — zeigt, dass die Herzinsuffizienz-Evidenz kein Sonderfall ist, sondern in einem breiteren Muster steht. Da das bereitgestellte Abstract von Vaduganathan et al. keine gepoolten Schätzer ausweist, ist ein numerischer Vergleich der Punktschätzer hier nicht möglich; der Vergleich bleibt auf der Ebene der Fragestellung und des Designs.

Zweitens: Nierenendpunkte im kardiovaskulären Rahmen. Das SMART-C-Konsortium (2022) hat die Nierenendpunkte in den Mittelpunkt gestellt und als Moderator den Diabetesstatus geprüft. Die vorliegende Arbeit kehrt die Perspektive um: Die Nierenkrankheitsprogression ist einer von fünf Endpunkten in einem gemeinsamen Modell, und der geprüfte Moderator ist die Krankheitsentität. Das Ergebnis ist bemerkenswert: Die Nierenkrankheitsprogression zeigt mit exp(β) = 0,68 den **stärksten relativen Effekt aller fünf Endpunkte** — stärker noch als die Herzinsuffizienz-Hospitalisierung (0,71) — und das über 13 Studien mit N = 90.413. Auf der absoluten Skala relativiert sich dieser Befund allerdings erheblich (RD = −0,010; NNT ≈ 100), weil die Ereignisrate niedriger ist. Diese doppelte Lesart — stärkster relativer, aber nur mittlerer absoluter Effekt — ist genau die Art von Information, die eine rein relativ berichtende Synthese nicht liefern kann. Ein Vergleich mit den konkreten Schätzern der SMART-C-Arbeit ist nicht möglich, da das bereitgestellte Abstract keine Effektgrößen ausweist.

Drittens: Die Post-MI-Frage. Khan, Usman, Butler, Bhatt et al. (2024) haben als erste den Anspruch formuliert, das gesamte kardiometabolische Spektrum einschließlich der Postinfarktpopulation in einen Rahmen zu stellen. Die vorliegende Arbeit folgt dieser Einschlusslogik und macht die Spektrumserweiterung zusätzlich zum statistischen Objekt: Die Evidenzschicht `timepoint` trennt die 13 Studien der SMART-C-Abgrenzung (2015–2023; exp(β) = 0,78–0,79; RD = −0,014; p < 0,001) von den beiden 2024 publizierten Post-MI-Studien (exp(β) = 0,90–0,93; RD = −0,003; p = 0,143–0,495). Der Gruppenunterschied ist nicht signifikant (Wald p = 0,110–0,148), aber die Asymmetrie der Präzision ist unübersehbar: In der einen Schicht liegt ein mit hoher Sicherheit etablierter Effekt vor, in der anderen ein nicht signifikanter Punktschätzer aus zwei Studien. Die vorliegende Arbeit stützt damit die spektrumübergreifende Betrachtung, warnt aber zugleich davor, sie als gleichförmig etablierte Evidenz zu lesen. Da das Abstract der Arbeit von Khan et al. im vorliegenden Projekt keine Effektgrößen ausweist, kann auch hier nicht numerisch verglichen werden; die Übereinstimmung betrifft die Einschlusslogik und die Spektrumsfrage.

Viertens: Mechanismus und Empirie. Verma und McMurray (2018) haben die Mechanismusfrage geordnet, ohne sie zu entscheiden. Die hier beobachtete Konstellation — stärkste Effekte bei Herzinsuffizienz-Hospitalisierung und Nierenkrankheitsprogression, deutlich schwächere Effekte bei kardiovaskulärem Tod (0,87) und Gesamtmortalität (0,89), weitgehende Unabhängigkeit von der Krankheitsentität und von der Substanz (Wald p = 0,805) — ist mit der dort beschriebenen Kombination aus hämodynamischen, tubuloglomerulären und metabolischen Wirkpfaden gut vereinbar und schwer mit einem primär antiatherosklerotischen oder primär antidiabetischen Wirkprinzip zu vereinbaren. Das ist jedoch eine Plausibilitätsaussage, kein Mechanismusnachweis: Eine Trial-Level-Meta-Analyse harter Endpunkte kann Wirkpfade nicht auflösen. Die Arbeit von Zelniker et al. (2020) zu Vorhofflimmern adressiert einen Endpunkt, der

hier nicht codiert ist; eine Aussage über arrhythmische Effekte ist aus dem vorliegenden Datensatz nicht ableitbar.

Insgesamt gilt: Die vorliegende Arbeit widerspricht keiner der fünf Vorarbeiten. Ihr Beitrag liegt in der gemeinsamen Modellierung von fünf Endpunkten über ein viergliedriges Krankheitsspektrum, in der methodischen Absicherung gegen Abhängigkeitsstrukturen und in der systematischen Parallelführung relativer und absoluter Effektmaße.

---

6 Diskussion

6.1 Hauptbefund

Über 15 große, doppelblinde, placebokontrollierte und ereignisgesteuerte Phase-III-Studien mit zusammen 100.952 Patient:innen senken SGLT2- beziehungsweise duale SGLT1/2-Inhibitoren das Risiko für kardiovaskulären Tod oder Hospitalisierung wegen Herzinsuffizienz um rund ein Fünftel: **HR 0,78 (95 %-KI 0,75–0,82)**. Dieser Schätzer ist in einem Maße robust, das in der Evidenzsynthese selten ist. Er ist identisch im Drei-Ebenen-Modell mit robuster CR2-Varianzschätzung und im klassischen Zwei-Ebenen-Random-Effects-Modell. Er bewegt sich im Leave-one-out zwischen 0,78 und 0,79. Er bleibt nach Trim-and-Fill (0,785), nach PEESE (0,786) und nach p-uniform* (0,781) praktisch unverändert. Auf Studienebene beträgt die Heterogenität null (I^2 = 0 %; Q = 14,1; p = 0,446), und das Prädiktionsintervall (0,75–0,82) fällt mit dem Konfidenzintervall zusammen. Alle 15 Studien sind nach RoB 2 mit niedrigem Verzerrungsrisiko bewertet, und die Evidenzsicherheit ist für alle fünf Endpunkte als hoch ratifiziert.

Die Endpunktstruktur präzisiert das Bild. Relativ betrachtet ist die Wirkung dort am stärksten, wo sie mechanistisch am ehesten zu erwarten ist: bei der Nierenkrankheitsprogression (exp(β) = 0,68) und der ersten Herzinsuffizienz-Hospitalisierung (0,71). Sie ist schwächer, aber konsistent bei den Mortalitätsendpunkten (kardiovaskulärer Tod 0,87; Gesamtmortalität 0,89). Absolut betrachtet kehrt sich die Rangfolge teilweise um: Der größte Nutzen pro behandelter Person entsteht beim kombinierten Endpunkt (RD = −0,019; NNT ≈ 53) und bei der Herzinsuffizienz-Hospitalisierung (−0,017; NNT ≈ 59), während der kardiovaskuläre Tod trotz eines klaren relativen Effekts eine NNT von etwa 125 erfordert. Der Wald-Test über die Endpunkte ist in beiden Metriken signifikant (p < 0,001 auf der Log-Skala; p = 0,003 auf der RD-Skala): Die Endpunkte unterscheiden sich tatsächlich, und diese Unterschiede sind nicht Rauschen.

Die klinisch wichtigste Botschaft ist deshalb keine einzelne Zahl, sondern ein Muster: **SGLT2-Inhibitoren verhindern vor allem Verschlechterungsereignisse — Dekompensation mit Hospitalisierungsbedarf und Nierenfunktionsverlust — und erst in zweiter Linie Tod.** Dieses Muster ist über HFrEF, HFpEF, Typ-2-Diabetes mit hohem kardiovaskulärem Risiko und chronische Niereninsuffizienz hinweg stabil und in seiner Größenordnung erstaunlich einheitlich (0,75 bis 0,79).

6.2 Die Post-Myokardinfarkt-Schicht als eigentliche Streitfrage

Der einzige substanzielle Bruch im sonst homogenen Bild ist die Postinfarktpopulation. Dort liegt der Punktschätzer bei HR 0,90 (exp(β) = 0,90 in Run v3, 0,93 in Run v6), die absolute Risikodifferenz bei −0,003 und die Signifikanz in keiner Spezifikation vor (p = 0,143 / 0,332 / 0,494). Zugleich ist der formale Interaktionstest in keiner Spezifikation signifikant (Wald p = 0,144–0,268; Q-between p = 0,268).

Diese Konstellation ist methodisch heikel und sollte nicht in eine der beiden bequemen Richtungen aufgelöst werden. Wer nur den nicht signifikanten Interaktionstest liest, schließt auf Gleichheit und empfiehlt SGLT2-Inhibitoren pauschal nach Myokardinfarkt — obwohl die Daten dort nur aus zwei Studien stammen und keinen Nutzen belegen. Wer nur die nicht signifikante Subgruppe liest, schließt auf Unwirksamkeit — obwohl die Nicht-Signifikanz bei k = 2 und einer niedrigen Ereignisrate wenig aussagt. Die angemessene Formulierung ist: **Für die unmittelbare Postinfarktphase ohne manifeste Herzinsuffizienz liegt derzeit kein Nachweis eines Nutzens auf den hier untersuchten Endpunkten vor; ein klinisch relevanter Nutzen ist mit den Daten aber ebenso wenig ausgeschlossen.**

Inhaltlich plausibel ist vor allem die Ausgangsrisiko-Erklärung. Die Postinfarktpopulationen wurden definitionsgemäß ohne vorbestehende Herzinsuffizienz und teils ohne Diabetes eingeschlossen; die Ereignisrate für Herzinsuffizienz-Hospitalisierung ist dort um ein Vielfaches niedriger als bei chronischer Herzinsuffizienz oder fortgeschrittener CKD. Eine relative Risikoreduktion von zehn Prozent auf ein Ausgangsrisiko von wenigen Prozent erzeugt eine absolute Reduktion von 0,3 Prozentpunkten — genau der beobachtete Wert. Die niedrige Risikodifferenz ist damit nicht notwendig ein Wirksamkeitsproblem, sondern möglicherweise ein Selektionsproblem der Studienpopulation. Dieselbe Logik erklärt auch, warum die Meta-Regression auf das Publikationsjahr eine kleine Abschwächung zeigt (Steigung 0,002; p = 0,043): Die spätesten Studien sind die Postinfarktstudien, und der scheinbare Zeittrend ist mit der Indikation konfundiert.

6.3 Was die Heterogenitätsbefunde wirklich bedeuten

Ein zentrales interpretatorisches Risiko dieser Arbeit ist die Doppelgesichtigkeit der Heterogenität: $I^2 = 0$ % im Zwei-Ebenen-Modell auf Studienebene, $I^2 = 77{,}7$ % im endpunktübergreifenden Drei-Ebenen-Modell. Eine selektive Zitation der jeweils passenden Zahl könnte die Befunde in beide Richtungen verzerren.

Die Auflösung ist eindeutig und sollte in jeder Weitergabe des Befunds mitgeliefert werden. Die 77,7 % zerlegen sich in 30,7 % auf der Studien- und 46,9 % auf der Effektgrößen-Ebene; die Varianzkomponenten selbst sind auf vier Dezimalstellen gleich null. Das hohe I^2 entsteht also nicht aus großen absoluten Unterschieden, sondern daraus, dass die Stichprobenvarianz in Studien dieser Größenordnung extrem klein ist, sodass I^2 als relatives Maß nach oben getrieben wird. Zudem liegt der größere Heterogenitätsanteil **innerhalb** der Studien, also zwischen den Endpunkten — und genau diese Varianzquelle wird durch den signifikanten Endpunkt-Moderator (p = 0,003) und durch die endpunktweise Schätzung in Tabelle 2 vollständig adressiert. Für den primären Endpunkt allein ist die Heterogenität null. Die korrekte Zusammenfassung lautet daher: **Die Studien stimmen überein; die Endpunkte unterscheiden sich.**

Das Prädiktionsintervall auf der RD-Skala ([−0,032; 0,007]) schließt die Null ein und mahnt zur Vorsicht. Es ist jedoch über fünf Endpunkte und vier Entitäten mit sehr unterschiedlichem Ausgangsrisiko gemittelt und daher kein indikationsspezifischer Vorhersagebereich. Für den primären Endpunkt auf der HR-Skala beträgt das Prädiktionsintervall 0,75–0,82 und schließt die Null-Wirkung deutlich aus.

6.4 Limitationen

(a) Trial-Level-Daten statt individueller Patientendaten. Alle Analysen beruhen auf aggregierten Studienergebnissen. Das begrenzt die Auflösung von Subgruppeneffekten grundsätzlich: Was hier als Krankheitsentität codiert ist, ist faktisch eine Studienkennung mit vier Ausprägungen, und ökologische Verzerrung — also der Fehlschluss von Studienmerkmalen auf Patientenmerkmale — ist nicht ausschließbar. Fragen wie die Modifikation durch Ausgangs-eGFR, Albuminurie, Ejektionsfraktion als kontinuierliche Variable, Alter, Geschlecht oder Hintergrundtherapie lassen sich nur mit individuellen Patientendaten beantworten. Entsprechende Kennwerte konnten im vorliegenden Snapshot nicht aggregiert ausgewiesen werden: [PLATZHALTER].

(b) Die Abdeckungseinschränkung als benannte Entscheidung. Diese Synthese bildet bewusst nur einen Ausschnitt der Literatur ab, und zwar aufgrund einer protokollierten Entscheidung, nicht aufgrund einer Beschaffungslücke. Eingeschlossen wurden ausschließlich doppelblinde, placebokontrollierte, ereignisgesteuerte Phase-III-Outcome-Studien mit $N \geq 1.000$ und einer Time-to-Event-Primäranalyse, publiziert zwischen 2015 und Mai 2024. Diese Abgrenzung folgt den Protokollen von SMART-C (2022) und Khan, Usman, Butler, Bhatt et al. (2024) und hat drei erkennbare Konsequenzen. Erstens sind kleinere Studien systematisch ausgeschlossen — dokumentiert etwa am Volltext-Ausschluss einer Studie mit N = 530, 90-Tage-Nachbeobachtung und hierarchischem Win-Ratio-Primärendpunkt. Diese Studienklasse kann Effekte auf Symptomlast, Lebensqualität und funktionelle Kapazität abbilden, die in der vorliegenden Synthese vollständig fehlen. Zweitens sind nicht-randomisierte Real-World-Daten ausgeschlossen (CVD-REAL, Kosiborod et al. 2017; EMPRISE, Patorno et al. 2019); damit fehlen Informationen zu Wirksamkeit unter Alltagsbedingungen, zu Adhärenz, Therapieabbruch und zu Populationen, die in Zulassungsstudien unterrepräsentiert sind. Drittens endet der

Publikationszeitraum im Mai 2024; danach publizierte Studien sind definitionsgemäß nicht enthalten. Die Summe dieser Entscheidungen erkauft maximale interne Validität — 15 von 15 Studien mit niedrigem Verzerrungsrisiko — mit einer eingeschränkten Breite der Evidenzbasis. Dies ist eine Abwägung, keine Lücke; sie sollte aber bei jeder Verallgemeinerung mitgedacht werden. Die Volltextbeschaffung selbst war vollständig: 18 von 18 Volltexten wurden geprüft, 0 waren nicht beschaffbar.

(c) Geringe Trefferzahl der Suchläufe. Die vier Suchläufe ergaben zusammen 44 Treffer, davon 18 in PubMed/MEDLINE. Für ein Forschungsfeld dieser Größe ist das wenig. Die Erklärung liegt in der sehr spezifischen Suchsyntax, die gezielt auf die großen Outcome-Studien ausgerichtet war, und in der Begrenzung des Publikationszeitraums. Die hohe Trennschärfe des Screenings ($\kappa = 1{,}00$, 0 offene Konflikte) spricht dafür, dass die identifizierten Treffer eindeutig klassifizierbar waren; sie belegt jedoch nicht die Vollständigkeit der Suche. Eine Rückwärts- und Vorwärtszitationssuche sowie die Prüfung grauer Literatur sind in ihrem Umfang nicht dokumentiert ([PLATZHALTER]); es ist daher nicht auszuschließen, dass qualifizierende Studien unentdeckt geblieben sind. Dass die 15 identifizierten Studien die aus der Fachliteratur bekannten großen Phase-III-Programme vollständig abdecken, ist ein inhaltliches Plausibilitätsargument, kein methodischer Vollständigkeitsnachweis.

(d) Moderatoranalysen sind explorativ und konfundiert. Die Moderatorfamilien `industry`, `funding` und `country` sind bei $k = 15$ Studien und bis zu zwölf Ausprägungen faktisch Studienkennungen. Insbesondere `country` erlaubt keine inhaltliche Interpretation. Auch `industry` ist vollständig mit der Indikation und dem Publikationsjahr konfundiert: Jede Substanzschicht entspricht einem bestimmten Entwicklungsprogramm mit bestimmten Zielpopulationen. Eine Rangfolge der Substanzen lässt sich aus diesen Daten nicht ableiten, und der nicht signifikante Wald-Test ($p = 0{,}805$) ist kein Äquivalenznachweis. Eine Korrektur für multiples Testen über die sechs Moderatorfamilien wurde nicht vorgenommen ([PLATZHALTER]); die grenzwertig signifikante Meta-Regression auf das Publikationsjahr ($p = 0{,}043$) wäre davon besonders betroffen.

(e) Rekonstruierte Effektgrößen. Nicht alle log-HR stammen aus direkt berichteten HR mit Konfidenzintervall; je nach Quellenlage wurden Peto-Verfahren, Log-Rank-p mit Ereigniszahlen, Kaplan-Meier-Rekonstruktion nach Tierney et al. (2007) oder Median-Überlebenszeiten unter Exponentialannahme genutzt. Der jeweilige Weg ist pro Effektgröße dokumentiert, die Rekonstruktionen sind aber mit zusätzlicher Unsicherheit behaftet, die in den Konfidenzintervallen nicht vollständig abgebildet ist. Wie viele Effektgrößen auf welchem Weg gewonnen wurden, ist im vorliegenden Snapshot nicht aggregiert ausgewiesen: [PLATZHALTER].

(f) Zählungsdiskrepanz der Effektgrößen. Der Audit-Trail weist 322 Effektgrößen aus, die analysierte Effektgrößentabelle 161. Die Differenz ist eine Zählkonvention des Projektsystems; die Zuordnungsregel ist [PLATZHALTER]. Alle berichteten Modelle beziehen sich auf die analysierte Tabelle. Die Diskrepanz berührt die Schätzer nicht, sollte aber vor Einreichung aufgelöst werden.

(g) Kein Sicherheitsprofil. Diese Arbeit synthetisiert ausschließlich Wirksamkeitsendpunkte. Unerwünschte Ereignisse — genitale Mykosen, Volumenmangel, diabetische Ketoazidose, Amputationen, Frakturen, Harnwegsinfektionen, akute Nierenschädigung — wurden nicht codiert und nicht analysiert. Eine Nutzen-Risiko-Abwägung ist auf Basis dieses Manuskripts nicht möglich und muss auf die Sicherheitsberichte der Einzelstudien und auf dedizierte Sicherheitssynthesen zurückgreifen.

(h) Endpunktauswahl. Arrhythmische Endpunkte (etwa Vorhofflimmern, adressiert bei Zelniker et al. 2020), Myokardinfarkt, Schlaganfall, Lebensqualität, Symptomlast und funktionelle Kapazität sind nicht Teil der Endpunktfamilie. Die Aussage dieser Arbeit beschränkt sich auf fünf harte kardiovaskuläre und renale Endpunkte.

(i) Zeithorizont. Risikodifferenzen und NNT beziehen sich auf die jeweiligen, unterschiedlich langen Studienlaufzeiten und sind nicht auf eine einheitliche Zeitachse normiert. Da die mittleren Nachbeobachtungszeiten im Snapshot nicht ausgewiesen sind ([PLATZHALTER]), sind die NNT-Werte als grobe Orientierung über den Studienzeitraum hinweg zu verstehen und nicht als Jahres-NNT.

6.5 Implikationen

Für die klinische Praxis. Die Befunde stützen die bereits vollzogene Verschiebung der Indikationsstellung: Nicht der Blutzucker, sondern das kardiorenale Risikoprofil begründet den Einsatz. Für Patient:innen mit chronischer Herzinsuffizienz über das gesamte Spektrum der Ejektionsfraktion, mit chronischer Niereninsuffizienz und mit Typ-2-Diabetes bei hohem kardiovaskulärem Risiko liegt eine relative Risikoreduktion von 21 bis 25 % auf dem kombinierten Endpunkt vor, bei hoher Evidenzsicherheit und niedrigem Verzerrungsrisiko. Die NNT von rund 53 über die Studienlaufzeiten ordnet den Nutzen in einen Bereich ein, der in der kardiovaskulären Sekundärprävention als klinisch bedeutsam gilt. Für die Postinfarktphase ohne Herzinsuffizienz ist die Evidenz dagegen schwächer und sollte nicht mit derselben Bestimmtheit kommuniziert werden.

Für Leitlinien. Die Konsistenz über Entitäten hinweg (keine signifikante Interaktion) spricht dafür, SGLT2-Inhibitoren eher als kardiorenale Basistherapie denn als indikationsspezifische Option zu formulieren. Zugleich legt die endpunktspezifische Struktur nahe, die Nutzenkommunikation zu differenzieren: Der dokumentierte Nutzen liegt primär in der Vermeidung von Dekompensation und Nierenfunktionsverlust, nicht primär in der Lebensverlängerung. Eine Leitlinienempfehlung, die eine substanzielle Mortalitätssenkung in den Vordergrund stellt, würde die Datenlage überdehnen (kardiovaskulärer Tod $\exp(\beta) = 0{,}87$, RD = −0,008, NNT ≈ 125).

Für die Methodik der Evidenzsynthese. Dieser Datensatz illustriert exemplarisch, warum die Modellierung der Abhängigkeitsstruktur bei Mehrfach-Endpunkt-Synthesen keine Formalie ist. Hätte man die fünf Endpunkte aus 15 Studien als unabhängig behandelt, wären die Konfidenzintervalle zu eng ausgefallen und die Heterogenität wäre fälschlich den Studien statt den Endpunkten zugeschrieben worden. Die Zerlegung des I^2 in eine Studien- (30,7 %) und eine Effektgrößen-Komponente (46,9 %) war hier der entscheidende diagnostische Schritt. Zugleich zeigt der Fall, dass I^2 auf der Skala der Risikodifferenz bei sehr großen Studien mit Vorsicht zu lesen ist.

---

7 Praxis- und Forschungsempfehlungen

7.1 Empfehlungen für die klinische Praxis

1. SGLT2-Inhibitoren als kardiorenale Basistherapie bei chronischer Herzinsuffizienz unabhängig von der Ejektionsfraktion einsetzen. Für die fünf Herzinsuffizienz-Studien (N = 21.947) liegt ein gepoolter Effekt von HR 0,78 auf den kombinierten Endpunkt vor, bei einer absoluten Risikoreduktion von 1,8 Prozentpunkten in dieser Entität. Die Indikationsstellung sollte sich an der Diagnose Herzinsuffizienz orientieren und nicht an der Ejektionsfraktion, am HbA1c oder am Diabetesstatus. Die Substanz sollte zusätzlich zur etablierten leitliniengerechten Therapie gegeben werden, nicht als deren Ersatz — alle Studien waren Add-on-Studien auf leitliniengerechter Hintergrundtherapie.

2. Bei chronischer Niereninsuffizienz die Nephroprotektion als eigenständige Indikation führen. Die Nierenkrankheitsprogression zeigt mit $\exp(\beta) = 0{,}68$ den stärksten relativen Effekt aller untersuchten Endpunkte über 13 Studien und 90.413 Patient:innen, bei hoher Evidenzsicherheit. In der CKD-Entität liegt der kardiovaskuläre Effekt zugleich bei HR 0,75 — dem niedrigsten Wert aller vier Entitäten. Die Substanzklasse adressiert hier beide Risikodimensionen gleichzeitig; eine Behandlungsentscheidung sollte das explizit berücksichtigen.

3. Nutzen in absoluten Zahlen kommunizieren und das Ausgangsrisiko benennen. Die Aufklärung sollte neben der relativen Risikoreduktion die absolute angeben: rund zwei vermiedene Ereignisse des kombinierten Endpunkts pro 100 Behandelte, entsprechend einer NNT von etwa 53 über die Studienlaufzeiten; für die Herzinsuffizienz-Hospitalisierung eine NNT von etwa 59; für die Gesamtmortalität und die Nierenkrankheitsprogression jeweils etwa 100; für den kardiovaskulären Tod etwa 125. Diese Zahlen machen nachvollziehbar, warum derselbe Wirkstoff bei einer Patientin mit fortgeschrittener Herzinsuffizienz einen anderen erwartbaren Nutzen hat als bei einem Patienten mit niedrigem Ereignisrisiko.

4. Den Nutzen nicht als Mortalitätsversprechen darstellen. Die Mortalitätseffekte sind real und statistisch

robust (Gesamtmortalität exp(β) = 0,89, p = 0,001; kardiovaskulärer Tod 0,87, p < 0,001), aber deutlich kleiner als die Effekte auf Hospitalisierung und Nierenfunktion. Eine Kommunikation, die den primären Nutzen in der Vermeidung von Dekompensationen und Nierenfunktionsverlust verortet, trifft die Datenlage genauer und beugt enttäuschten Erwartungen vor.

5. Nach Myokardinfarkt ohne Herzinsuffizienz zurückhaltend und individuell entscheiden. In der Postinfarktschicht (zwei Studien, N = 10.539) ist der Effekt mit HR 0,90 und RD = −0,003 klein und statistisch nicht gesichert (p = 0,143–0,494). Eine routinemäßige Verordnung allein aufgrund eines durchgemachten Myokardinfarkts lässt sich aus diesen Daten nicht begründen. Bestehen jedoch zusätzlich Herzinsuffizienz, chronische Niereninsuffizienz oder Typ-2-Diabetes mit hohem Risiko, greifen die jeweils gut belegten Indikationen unverändert.

6. Keine Substanzpräferenz aus diesen Daten ableiten. Die Unterschiede zwischen den Wirkstoffschichten (exp(β) 0,65 bis 0,88) sind statistisch nicht gesichert (Wald p = 0,805) und vollständig mit Indikation und Studienprogramm konfundiert. Die Substanzwahl sollte sich daher an der zugelassenen Indikation, der vorliegenden Studienevidenz für die konkrete Patientengruppe, dem Nierenfunktionsbereich der Zulassung, der Verträglichkeit und praktischen Erwägungen orientieren — nicht an der hier berichteten Rangfolge der Punktschätzer.

7. Wirksamkeitsentscheidungen durch eine eigenständige Sicherheitsbewertung ergänzen. Diese Synthese enthält keine Daten zu unerwünschten Ereignissen. Vor Therapiebeginn sind die bekannten Sicherheitsaspekte der Substanzklasse — unter anderem Volumenstatus, Ketoazidoserisiko, genitale Infektionen, Begleitmedikation mit Diuretika und Insulin — anhand der Fachinformation und der Sicherheitsberichte der Einzelstudien zu bewerten. Die hier gezeigte Wirksamkeit ersetzt keine Nutzen-Risiko-Abwägung.

7.2 Empfehlungen für die Forschung

8. Individuelle Patientendaten-Meta-Analyse über das gesamte Spektrum durchführen. Die zentralen offenen Fragen — Modifikation des Effekts durch Ausgangs-eGFR, Albuminurie, Ejektionsfraktion als Kontinuum, Alter, Geschlecht, Hintergrundtherapie und Zeit seit Diagnose — sind mit Trial-Level-Daten strukturell nicht beantwortbar. Eine kollaborative IPD-Analyse über die 15 hier eingeschlossenen Studien hinweg, aufbauend auf der bestehenden SMART-C-Infrastruktur, wäre der größte verfügbare Erkenntnisgewinn und würde zugleich die ökologische Verzerrung der Subgruppenbefunde auflösen.

9. Die Postinfarktfrage gezielt mit risikoangereicherten Populationen adressieren. Der kleine, nicht signifikante Effekt in der Postinfarktschicht ist mit einem niedrigen Ausgangsrisiko vereinbar. Künftige Studien sollten daher nicht die unselektierte Postinfarktpopulation, sondern Patient:innen mit erhöhtem Risiko für Herzinsuffizienz nach Infarkt einschließen — etwa anhand von Ejektionsfraktion, Infarktgröße, natriuretischen Peptiden, Nierenfunktion oder Diabetesstatus — und ausreichend lange nachbeobachten, um Dekompensationsereignisse überhaupt beobachten zu können. Eine ereignisgesteuerte Planung mit realistischer Annahme über die Ereignisrate ist hier entscheidend.

10. Die Endpunktdefinitionen weiter harmonisieren und Zeitachsen standardisieren. Die vorliegende Analyse konnte Risikodifferenzen und NNT nur über die jeweiligen Studienlaufzeiten hinweg berichten, weil einheitliche Zeithorizonte fehlen. Künftige Studien und Synthesen sollten Ereignisraten zu definierten Zeitpunkten (etwa 12, 24 und 36 Monate) berichten, damit absolute Effekte vergleichbar und in gesundheitsökonomische Modelle überführbar werden. Ebenso sollten rekurrierende Ereignisse systematisch neben Time-to-first-event-Analysen berichtet werden, da gerade bei Herzinsuffizienz die Ereigniswiederholung klinisch bedeutsam ist.

11. Wirksamkeits- und Sicherheitssynthesen zusammenführen. Eine Synthese, die dieselben 15 Studien hinsichtlich unerwünschter Ereignisse mit derselben methodischen Strenge analysiert — einschließlich Drei-Ebenen-Modellierung und absoluter Risikomaße — würde die hier vorgelegte Nutzenschätzung erst entscheidungsfähig machen. Dies ist die naheliegendste direkte Fortsetzung dieser Arbeit.

12. Patientenrelevante Endpunkte jenseits harter Ereignisse systematisch synthetisieren. Symptomlast,

Lebensqualität und funktionelle Kapazität fallen durch das hier gewählte Einschlussraster (N ≥ 1.000, ereignisgesteuert, Time-to-Event) systematisch heraus — exemplarisch dokumentiert am Ausschluss einer Studie mit N = 530 und hierarchischem Win-Ratio-Endpunkt. Eine komplementäre Synthese mit angepassten Einschlusskriterien und geeigneten Modellen für Win-Ratio- und ordinale Endpunkte sollte diese Lücke schließen.

13. Die Konsistenz der Befunde für die Methodendiskussion nutzen. Der hier beobachtete Fall — I^2 = 0 % auf Studienebene bei gleichzeitig I^2 = 77,7 % im endpunktübergreifenden Modell und $\tau^2 \approx 0$ in beiden Komponenten — eignet sich als Lehrbeispiel dafür, dass I^2 auf der Skala der Risikodifferenz bei sehr großen Studien kein belastbares Maß klinischer Inkonsistenz ist. Eine methodische Arbeit, die dieses Verhalten an realen Daten systematisch durchdekliniert und Entscheidungsregeln für die Berichterstattung vorschlägt, hätte über dieses Feld hinaus Nutzen.

14. Real-World-Evidenz als Ergänzung, nicht als Ersatz, strukturiert einbinden. Die hier ausgeschlossenen Beobachtungsstudien (CVD-REAL, EMPRISE) beantworten Fragen, die randomisierte Studien nicht beantworten können: Adhärenz, Therapiepersistenz, Wirksamkeit in unterrepräsentierten Populationen, Langzeitsicherheit. Eine getrennt geführte, methodisch eigenständige Synthese dieser Evidenz mit angemessener Bias-Bewertung sollte die randomisierte Evidenz ergänzen; eine Vermischung beider Evidenzarten in einem gemeinsamen Pool ist dagegen nicht zu empfehlen.

15. Daten, Code und Entscheidungsprotokoll vollständig offenlegen. Die vorliegende Arbeit stützt sich auf versionierte Analyse-Runs mit Daten-Snapshot-Hashes und einen Entscheidungs-Audit-Trail. Diese Artefakte sollten mit der Publikation veröffentlicht werden, einschließlich der Codierung aller Moderatorvariablen, der Herkunftsdokumentation jeder einzelnen Effektgröße und der RoB-2-Domänenurteile. Erst damit ist die Rekonstruierbarkeit gewährleistet, die bei einer Synthese mit mehreren Modellspezifikationen und mehreren Metriken unverzichtbar ist.

16. Die Zählkonventionen und offenen Dokumentationspunkte vor Einreichung schließen. Konkret betrifft das die Differenz zwischen 322 erfassten und 161 analysierten Effektgrößen, die Zuordnung der beiden Beobachtungsstudien zur Screening-Ebene, die domänenspezifische RoB-2-Verteilung, die Dokumentation von Zitations- und Grauliteratursuche sowie die Entscheidung zur Multiplizitätskorrektur über die sechs Moderatorfamilien. Keiner dieser Punkte verändert die Schätzer; alle betreffen die Nachvollziehbarkeit.

---

8 Schlussfolgerung

Über 15 große, doppelblinde, placebokontrollierte Phase-III-Studien mit 100.952 Patient:innen senken SGLT2- und duale SGLT1/2-Inhibitoren das Risiko für kardiovaskulären Tod oder Herzinsuffizienz-Hospitalisierung um 22 % (HR 0,78; 95 %-KI 0,75–0,82), bei null Heterogenität auf Studienebene, vollständiger Stabilität im Leave-one-out, unauffälligen Bias-Prüfungen in fünf Verfahren, durchweg niedrigem Verzerrungsrisiko und hoher Evidenzsicherheit für alle fünf Endpunkte. Der Effekt ist am stärksten bei Nierenkrankheitsprogression und Herzinsuffizienz-Hospitalisierung, kleiner, aber robust bei kardiovaskulärem Tod und Gesamtmortalität. Er ist über chronische Niereninsuffizienz (0,75), Herzinsuffizienz (0,78) und Typ-2-Diabetes mit hohem kardiovaskulärem Risiko (0,79) hinweg praktisch unverändert; für die Postinfarktpopulation ohne Herzinsuffizienz liegt mit zwei Studien ein kleinerer, nicht signifikanter Schätzer (0,90) vor, ohne dass sich eine Interaktion statistisch belegen ließe. Der absolute Nutzen skaliert mit dem Ausgangsrisiko — das ist die zentrale Botschaft für die Indikationsstellung und zugleich die plausibelste Erklärung des schwächeren Postinfarktbefunds. Die Substanzklasse ist damit als krankheitsübergreifende kardiorenale Therapie zu verstehen, deren Einsatz sich am Risikoprofil und nicht am Stoffwechselparameter orientieren sollte.

---

9 Anhang-Hinweis

Die vollständig ausgefüllte **PRISMA-2020-Checkliste** sowie das **PRISMA-2020-Flussdiagramm** (Abbildung 1) liegen dem Projekt-Report bei. Ergänzend enthält der Report: die dokumentierten Suchstrings mit Datum und Trefferzahl je Datenbank, die Liste der Volltext-Ausschlüsse mit Begründung, die vollständige Effektgrößentabelle mit Herkunftsdokumentation je Schätzer, die RoB-2-Domänenurteile für alle 15 Studien (75 Einträge), die GRADE-Evidenzprofile je Endpunkt und Metrik, die versionierten Analyse-Runs (v1, v3, v5, v6, v7) mit zugehörigen Daten-Snapshot-Hashes sowie den Entscheidungs-Audit-Trail des Screenings.

Abbildungsverzeichnis

- **Abbildung 1:** PRISMA-2020-Flussdiagramm (Abschnitt 4.1)
- **Abbildung 2:** Wald-Diagramm der 15 eingeschlossenen Studien für den primären Endpunkt CV Death or HHF (Abschnitt 4.3)
- **Abbildung 3:** Trichter-Diagramm für den primären Endpunkt mit Egger-Regressionsgerade (Abschnitt 4.13)

Tabellenverzeichnis

- Tabelle 1: Charakteristika der 15 eingeschlossenen Studien
- Tabelle 2: Gepoolte Effekte je Endpunkt
- Tabelle 3: Effekte nach Krankheitsentität
- Tabelle 4: Effekte nach Substanz und Dosisschema
- Tabelle 5: Effekte nach Förderquelle
- Tabelle 6: Effekte nach geografischer Breite
- Tabelle 7: Vergleichsbasis — alle fünf identifizierten Vorarbeiten