Studien lesen ohne Statistikstudium

Leitfaden zu Begriffen, Kennzahlen und Bewertungswerkzeugen für randomisierte Studien – am Beispiel ICRAKI

Zielgruppe: Assistenzärzt*innen · Andreas Fazekas · luftikus.org · persönliche Lehrunterlage, keine Stellungnahme des Arbeitgebers · Version 1.2 · 03.11.2026 (öffentliche Fassung) · Quellenstand: PubMed-Verifikation 03.11.2026

So ist der Leitfaden aufgebaut
Die Module bauen aufeinander auf: von der Frage, warum Studien überhaupt in die Irre führen können, über Design und Kennzahlen bis zu den Bewertungswerkzeugen. Am Ende wird ICRAKI Schritt für Schritt durchgerechnet. Jeder Begriff steht auf einer Karte mit derselben Gliederung:

Einfach erklärt	Alltagssprache, ohne Formeln
Fachlich	präzise Definition, ggf. Formel
In ICRAKI	wo der Begriff in der Beispielstudie vorkommt [1]
Achtung	typischer Denkfehler
Beleg	Quelle mit Link; Volltitel im Literaturverzeichnis mit einem Satz Erläuterung

Empfohlener Weg: Module 1–4 vor dem ersten Journal Club lesen (etwa 45 min). Module 5–9 beim Ausfüllen des Bewertungsbogens als Nachschlagewerk nutzen. Modul 10 zum Üben, Modul 11 zum Überprüfen.

Inhalt

1. Zufall und Verzerrung – die zwei Gegner
2. Wie ein RCT Verzerrung verhindert
3. Welche Frage stellt die Studie?
4. Wer wird ausgewertet?
5. Endpunkte
6. Effektmaße – wie groß ist der Unterschied?
7. Unsicherheit – wie sicher ist der Unterschied?
8. Überwachung und Frühabbruch
9. Bewertungswerkzeuge
10. Rechenwerkstatt ICRAKI
11. Selbsttest
12. Spickzettel · Literatur

Modul 1: Zufall und Verzerrung – die zwei Gegner

Lernziel: zwei Fehlerarten unterscheiden und wissen, welches Werkzeug welche Fehlerart adressiert.

Systematischer Fehler (Bias, Verzerrung) (engl. bias)	
Einfach erklärt	Eine Waage, die immer 2 kg zu viel anzeigt. Auch mit vielen Messungen wird das Ergebnis nicht richtiger.
Fachlich	Abweichung des Studienergebnisses vom wahren Effekt, die durch Design, Durchführung oder Analyse entsteht und sich mit größerer Fallzahl nicht verringert. Wichtige Formen: Selektionsbias, Performance-Bias, Attrition-Bias, Messbias, Berichtsbias.
In ICRAKI	Offenes Design, Ausschlüsse nach Randomisierung und Frühabbruch sind mögliche Quellen.
Achtung	Eine große Studie kann präzise falsch sein.
Beleg	[3] Schulz 2002 · [38] Sterne 2019

Zufallsfehler (engl. random error)	
Einfach erklärt	Würfeln: Bei 10 Würfen kommt die Sechs selten genau gleich oft wie bei 6.000 Würfen vor. Kleine Stichproben schwanken stark.
Fachlich	Unsystematische Streuung eines Schätzwerts um den wahren Wert. Wird mit steigender Fallzahl kleiner und durch Konfidenzintervall bzw. p-Wert beschrieben.
In ICRAKI	Geplant waren 1.000 Patient*innen, ausgewertet 535: Die Schätzungen sind unsicherer als geplant.
Achtung	Statistik beschreibt den Zufall, nicht den Bias. Ein kleiner p-Wert schützt nicht vor Verzerrung.
Beleg	[23] Gardner 1986 · [24] Greenland 2016

Merksatz: Fragen zur *Validität* (Bias) zuerst, Fragen zur *Präzision* (Zufall) danach. Diese Reihenfolge stammt aus den JAMA Users' Guides [37].

Modul 2: Wie ein RCT Verzerrung verhindert

Lernziel: die Bausteine Randomisierung, verdeckte Zuteilung und Verblindung erklären und im Methodenteil finden.

Randomisierte kontrollierte Studie (RCT) (engl. randomised controlled trial)	
Einfach erklärt	Der Zufall entscheidet, wer welche Behandlung bekommt. Dadurch sind die Gruppen im Mittel gleich, auch in Merkmalen, die niemand gemessen hat.
Fachlich	Prospektive Studie, in der die Zuteilung zur Intervention durch einen Zufallsprozess erfolgt. Sie schützt vor Selektions- und Confounding-Bias.
In ICRAKI	Parallelgruppen-RCT, 1:1, IHD gegen CRRT.
Achtung	Randomisierung allein reicht nicht. Die Zuteilung muss auch verdeckt sein (siehe unten).
Beleg	[2] Schulz 2002

Block- und stratifizierte Randomisierung (engl. block / stratified randomisation)	
Einfach erklärt	Blöcke sorgen dafür, dass die Gruppen gleich groß bleiben. Strata sorgen dafür, dass etwa schwer und leicht Kranke in beiden Gruppen gleich verteilt sind.
Fachlich	Blockrandomisierung: Innerhalb jedes Blocks wird gleich oft jedem Arm zugeteilt. In offenen Studien sollen die Blockgrößen variieren und größer sein, damit niemand die nächste Zuteilung erraten kann. Stratifizierung: getrennte Zufallsfolgen je Prognosefaktor. Sie ist vor allem bei kleinen Studien, bei Zwischenanalysen mit wenigen Patient*innen und bei Äquivalenzstudien nützlich. Stratifizierungsfaktoren sollen in der Analyse berücksichtigt werden.
In ICRAKI	Permutierte Blöcke mit verdeckter Größe; Strata: Zentrum, Flüssigkeitsbilanz (≥ 6 l) und Katecholamindosis (> 0,5 µg/kg/min). Die Hauptanalysen wurden für Flüssigkeitsbilanz und Katecholamindosis adjustiert.
Beleg	[2] Schulz 2002 · [4] Kernan 1999

Verdeckte Zuteilung (engl. allocation concealment)	
Einfach erklärt	Die einschließende Ärztin weiß nicht, welche Behandlung der oder die nächste Patient*in bekommt. Sonst würde sie vielleicht unbewusst „passende“ Patient*innen auswählen.
Fachlich	Schutz der Zufallsfolge bis zur Zuteilung. Unzureichende Verdeckung führt im Mittel zu überschätzten Effekten. Nicht mit Verblindung verwechseln: Verdeckung wirkt *vor* der Zuteilung, Verblindung *danach*.
In ICRAKI	Zentrale webbasierte Randomisierung nach Prüfung der Einschlusskriterien. Das gilt als gute Verdeckung.
Beleg	[3] Schulz 2002

Baseline-Imbalanz (engl. baseline imbalance)	
Einfach erklärt	Trotz Zufall kann eine Gruppe zufällig etwas kränker sein, vor allem bei kleinen Studien.
Fachlich	Ungleiche Verteilung prognostischer Merkmale zu Studienbeginn. Signifikanztests auf Baseline-Unterschiede sind bei korrekter Randomisierung überflüssig; entscheidend ist, ob der Unterschied prognostisch relevant ist.
In ICRAKI	Chronische Nierenerkrankung 21,9 vs. 15,4 %, Notfall-OP 18,1 vs. 9,5 %, beides häufiger in der IHD-Gruppe. Die Autor*innen berichten eine post hoc adjustierte Sensitivitätsanalyse mit konsistentem Ergebnis.
Achtung	Tabelle 1 immer nach der Frage lesen: Benachteiligt ein Unterschied eine Gruppe systematisch?
Beleg	[3] Schulz 2002 · [5] Roberts 1999

Verblindung und offenes Design (engl. blinding, open-label)	
Einfach erklärt	Wer weiß, wer was bekommt, behandelt oder beurteilt vielleicht anders.
Fachlich	Verblindung bedeutet, dass Patient*innen, Behandler*innen oder Endpunkterheber*innen die Zuteilung nicht kennen. Begriffe wie „doppelblind“ sind uneinheitlich; Studien sollen angeben, *wer* verblindet war. Bei objektiven Endpunkten wie Tod ist wenig Raum für Messbias. Nicht doppelblinde Studien sind nicht automatisch schlechter.
In ICRAKI	Offen: Eine Dialyse lässt sich nicht verblinden. Tod ist objektiv. Die Entscheidung über die Dialyseabhängigkeit traf aber unverblindetes Personal.
Achtung	Offenes Design betrifft weiche Endpunkte stärker als harte.
Beleg	[6] Schulz 2002

Protokoll, Registrierung, Analyseplan (engl. protocol, registration, SAP)	
Einfach erklärt	Vorher aufschreiben, was man messen und wie man rechnen will. So kann niemand nachträglich das „schönste“ Ergebnis aussuchen.
Fachlich	Das Studienprotokoll (Standard: SPIRIT 2025) und der statistische Analyseplan legen Endpunkte und Analysen vorab fest; die Registrierung (z. B. ClinicalTrials.gov) macht das öffentlich überprüfbar. Abweichungen zwischen Protokoll und Publikation weisen auf selektives Berichten hin (RoB 2, Domäne 5).
In ICRAKI	Registriert als NCT06032884; Protokoll vorab publiziert; Ergebnisse gemäß Analyseplan berichtet. Abweichung in der Durchführung (kein selektives Berichten): geplant waren zwei, durchgeführt drei Zwischenanalysen.
Beleg	[7] Chan 2025 · [38] Sterne 2019

Merksatz: Randomisierung schützt vor Selektion, Verdeckung schützt die Randomisierung, Verblindung schützt die Behandlung und die Messung.

Modul 3: Welche Frage stellt die Studie?

Lernziel: Überlegenheits- und Nichtunterlegenheitsstudien unterscheiden und Fallzahlannahmen prüfen.

Überlegenheit, Nichtunterlegenheit, Äquivalenz (engl. superiority, non-inferiority, equivalence)	
Einfach erklärt	Überlegenheit: Ist A besser als B? Nichtunterlegenheit: Ist A höchstens um einen tolerierbaren Betrag schlechter als B, etwa weil A billiger oder einfacher ist? Äquivalenz: Sind A und B praktisch gleich?
Fachlich	Die Hypothese bestimmt, wie das Konfidenzintervall gelesen wird. Bei Nichtunterlegenheit wird geprüft, ob die ungünstige Grenze des KI die Margin unterschreitet. Ein einseitiges α von 0,025 entspricht dabei einem zweiseitigen 95-%-KI.
In ICRAKI	Nichtunterlegenheitsstudie: IHD sollte höchstens 7,5 Prozentpunkte mehr MAKE90-Ereignisse haben als CRRT.
Achtung	Eine gescheiterte Nichtunterlegenheit beweist nicht automatisch Unterlegenheit. Das KI zeigt, welche Effekte mit den Daten vereinbar sind.
Beleg	[8] Piaggio 2012 · [9] Mauri 2017

Nichtunterlegenheitsgrenze (Margin) (engl. non-inferiority margin, Δ)	
Einfach erklärt	Der größte Nachteil, den man für die Vorteile von A noch in Kauf nimmt.
Fachlich	Vorab festgelegter, klinisch und statistisch begründeter Grenzwert. Seine Wahl ist eine Werteentscheidung und muss begründet werden.
In ICRAKI	7,5 Prozentpunkte absolut (etwa 10 % relativ) für MAKE90; HR 1,27 für die Mortalität. Diskussionsfrage: Ist das bei einem Endpunkt, der überwiegend aus Todesfällen besteht, akzeptabel?
Beleg	[8] Piaggio 2012 · [9] Mauri 2017

Fallzahlplanung, Power, α (engl. sample size, power, alpha)	
Einfach erklärt	Vorab wird ausgerechnet, wie viele Patient*innen man braucht, um einen erwarteten Effekt mit hoher Wahrscheinlichkeit zu entdecken.
Fachlich	Power = Wahrscheinlichkeit, einen tatsächlich vorhandenen Effekt der angenommenen Größe zu entdecken (meist 80–90 %). α = akzeptierte Wahrscheinlichkeit eines falsch-positiven Befunds. Die Annahmen über Ereignisrate und Effektgröße sind subjektive klinische Urteile und beeinflussen die Fallzahl stark.
In ICRAKI	Annahme: 75 % MAKE90-Rate; 1.000 Patient*innen für 80 % Power bei einseitigem α 0,025. Erreicht wurden 554 Randomisierte, die Power für den primären Endpunkt war damit reduziert.
Beleg	[10] Schulz 2005

Modul 4: Wer wird ausgewertet?

> **Lernziel:** ITT, mITT und Per-Protocol unterscheiden und ihre Auswirkung auf das Ergebnis einschätzen.

Intention-to-treat (ITT)	
Einfach erklärt	Ausgewertet wird jede*r in der Gruppe, in die sie oder er gelost wurde, egal was danach passiert ist.
Fachlich	Erhält den Vorteil der Randomisierung. Wird häufig behauptet, aber unvollständig umgesetzt: In einer Übersicht schlossen 12 von 119 „ITT“-Studien Patient*innen aus, die die Intervention nicht begonnen hatten.
Achtung	Genau lesen, wer ausgeschlossen wurde. Die Bezeichnung ITT allein garantiert nichts.
Beleg	[11] Hollis 1999

Modifizierte ITT (mITT)	
Einfach erklärt	Eine „fast“ ITT, bei der nach dem Losen einige Personen herausgenommen werden.
Fachlich	Uneinheitlich definiert. In einer systematischen Übersicht hatten 80 % der mITT-Studien Ausschlüsse nach Randomisierung. Die Autor*innen empfehlen, Ausschlüsse explizit zu benennen statt den Begriff mITT zu verwenden.
In ICRAKI	554 randomisiert, 535 ausgewertet; der Artikel nennt diese Population ITT. Ausgeschlossen wurden Patient*innen ohne Nierenersatztherapie oder ohne Einwilligung zur Datennutzung (12 IHD, 7 CRRT).
Achtung	In offenen Studien können solche Ausschlüsse von der bekannten Zuteilung beeinflusst sein.
Beleg	[12] Abraha 2010

Per-Protocol-Analyse und Behandlungswechsel (engl. per-protocol, crossover)	
Einfach erklärt	Ausgewertet werden nur jene, die die Behandlung wie geplant bekommen haben.
Fachlich	Per-Protocol-Analysen verlieren den Schutz der Randomisierung. Bei Nichtunterlegenheitsstudien gehören ITT- *und* Per-Protocol-Analyse berichtet, weil Behandlungswechsel die Gruppen einander angleichen und so Nichtunterlegenheit begünstigen können.
In ICRAKI	Wechsel der Methode bei 21,9 % (IHD) und 42,5 % (CRRT), erlaubt ab Tag 4. Per-Protocol-Ergebnis ähnlich der Hauptanalyse.
Beleg	[8] Piaggio 2012 · [9] Mauri 2017

Fehlende Daten und multiple Imputation (engl. missing data, multiple imputation)	
Einfach erklärt	Wenn Werte fehlen, werden sie mehrfach „plausibel geschätzt“. Die Ergebnisse über alle Schätzungen werden zusammengefasst, damit die zusätzliche Unsicherheit sichtbar bleibt.
Fachlich	Multiple Imputation ersetzt fehlende Werte mehrfach auf Basis der beobachteten Daten und kombiniert die Analysen. Sie ist nur so gut wie ihre Annahmen über den Grund des Fehlens; Anwendung und Annahmen sollen berichtet werden.
In ICRAKI	Kreatinin fehlte bei 20 von 213 Überlebenden ohne Dialyse und wurde imputiert. Der Vitalstatus fehlte nur bei 4 Patient*innen.
Beleg	[13] Sterne 2009

Merksatz: Vergleichen, wie viele randomisiert und wie viele ausgewertet wurden. Jede Differenz braucht eine Begründung.

Modul 5: Endpunkte

Lernziel: primäre, sekundäre und kombinierte Endpunkte unterscheiden und konkurrierende Risiken erkennen.

Primärer und sekundärer Endpunkt (engl. primary / secondary outcome)	
Einfach erklärt	Der primäre Endpunkt ist die eine Hauptfrage, auf die die Studie ausgelegt ist. Sekundäre Endpunkte sind Zusatzfragen.
Fachlich	Fallzahl und Hypothesentest beziehen sich auf den primären Endpunkt. Je mehr Endpunkte getestet werden, desto mehr Zufallsbefunde entstehen (Multiplizität). Alle analysierten Endpunkte müssen berichtet werden.
In ICRAKI	Primär: MAKE90. Mortalität war sekundär, ihr KI ist nicht für Multiplizität adjustiert. Der Mortalitätsbefund ist daher ein starkes Signal, aber kein bestätigter Hauptbefund.
Beleg	[14] Schulz 2005

Kombinierter Endpunkt (engl. composite endpoint)	
Einfach erklärt	Mehrere Ereignisse werden zu einem zusammengezählt: Wer mindestens eines hat, zählt als Ereignis.
Fachlich	Erhöht die Ereigniszahl und damit die Power, kann aber große Unterschiede zwischen den Komponenten verdecken. Gut interpretierbar, wenn die Komponenten ähnlich wichtig und ähnlich häufig sind und in dieselbe Richtung zeigen. Häufig sind die wichtigeren Komponenten seltener.
In ICRAKI	MAKE90 = Tod oder Dialyseabhängigkeit oder Kreatinin > 25 % über Ausgangswert. Getrieben von Todesfällen: 164 von 203 bzw. 147 von 197 Ereignissen.
Achtung	Immer die Komponenten einzeln ansehen.
Beleg	[15] Montori 2005 · [16] Ferreira-González 2007

Konkurrierendes Risiko (engl. competing risk)	
Einfach erklärt	Wer stirbt, kann nicht mehr dialyseabhängig werden. Der Tod „konkurriert“ mit dem Nierenendpunkt.
Fachlich	Ereignis, das das Eintreten des interessierenden Ereignisses verhindert. Die Kaplan-Meier-Methode überschätzt dann die Häufigkeit; geeignet ist die kumulative Inzidenzfunktion, als Regressionsmodell etwa das Fine-Gray-Modell (Subdistribution-HR).
In ICRAKI	Mehr Überlebende unter CRRT bedeuten mehr Patient*innen, bei denen ein Nierenendpunkt überhaupt auftreten kann. Die Dialysebeendigung wurde mit Fine-Gray analysiert: Subdistribution-HR 0,77 (0,61–0,97).
Beleg	[17] Austin 2016

Modul 6: Effektmaße – wie groß ist der Unterschied?

Lernziel: absolute und relative Effektmaße berechnen und erklären, warum beide nötig sind.

Ausgangspunkt jeder Rechnung ist eine Vierfeldertafel. Für die 90-Tage-Mortalität in ICRAKI [1]:

	verstorben	lebend	gesamt	Risiko

IHD	164 (a)	96 (b)	260	a/(a+b) = 0,631
CRRT	147 (c)	128 (d)	275	c/(c+d) = 0,535

Risiko (Ereignisrate) (engl. risk, event rate)	
Einfach erklärt	Anteil der Patient*innen einer Gruppe, bei denen das Ereignis eintritt.
Fachlich	Risiko = Ereignisse / alle Patient*innen der Gruppe, immer für einen definierten Zeitraum.
In ICRAKI	Tod bis Tag 90: 63,1 % unter IHD, 53,5 % unter CRRT (roh).
Beleg	[18] Laupacis 1988

Absolute Risikodifferenz (ARD) (engl. absolute risk difference, ARR/ARI)	
Einfach erklärt	Um wie viele Prozentpunkte ändert sich das Risiko? Die Zahl, die für die einzelne Patientin am meisten zählt.
Fachlich	ARD = $\text{Risiko}_{\text{Intervention}} - \text{Risiko}_{\text{Kontrolle}}$. Wird in Prozentpunkten (PP) angegeben, nicht in Prozent.
In ICRAKI	0,631 − 0,535 = 0,096 = 9,6 PP mehr Todesfälle unter IHD (roh). Für MAKE90 adjustiert: 6,28 PP.
Achtung	„9,6 Prozentpunkte“ und „18 Prozent mehr“ beschreiben denselben Befund, absolut bzw. relativ.
Beleg	[18] Laupacis 1988

Relatives Risiko (RR) (engl. relative risk, risk ratio)	
Einfach erklärt	Das Wievielfache des Risikos hat die eine Gruppe im Vergleich zur anderen?
Fachlich	RR = Risiko_{I} / Risiko_{K}. RR = 1: kein Unterschied. Relative Risikoerhöhung = RR − 1. Relative Angaben wirken bei seltenen Ereignissen eindrucksvoll, obwohl der absolute Nutzen klein sein kann. Deshalb absolute und relative Maße gemeinsam angeben.
In ICRAKI	0,631 / 0,535 = 1,18, also 18 % relativ mehr Todesfälle. MAKE90: adjustiertes RR 1,09 (0,98–1,20).
Beleg	[18] Laupacis 1988

Odds Ratio (OR)	
Einfach erklärt	„Odds“ sind Chancen wie beim Wetten: Ereignisse geteilt durch Nicht-Ereignisse. Die OR ist das Verhältnis zweier solcher Chancen.
Fachlich	OR = (a/b) / (c/d). Bei seltenen Ereignissen ist OR ≈ RR. Bei häufigen Ereignissen liegt die OR weiter von 1 entfernt als das RR und wird dann leicht als RR fehlgedeutet.
In ICRAKI	Eigene Rechnung: (164/96) / (147/128) = 1,71 / 1,15 = 1,49, während RR = 1,18. Bei über 50 % Mortalität wäre die OR als RR gelesen eine deutliche Übertreibung.
Achtung	Eine OR nie wie ein RR sprechen.
Beleg	[20] Davies 1998

Number needed to treat / harm (NNT/NNH)	
Einfach erklärt	Wie viele Patient*innen muss man behandeln, damit eine*r zusätzlich profitiert (NNT) oder zusätzlich Schaden nimmt (NNH)?
Fachlich	NNT = 1 / |ARD|. Gehört immer mit Zeitraum und Konfidenzintervall angegeben. Die KI-Grenzen sind die Kehrwerte der KI-Grenzen der ARD. Schließt das KI der ARD die Null ein, reicht das KI der NNT „durch Unendlich“ von Nutzen bis Schaden (NNTB/NNTH).
In ICRAKI	1 / 0,096 ≈ 10. Also ein zusätzlicher Todesfall bis Tag 90 je etwa 10 Patient*innen, die initial IHD statt CRRT erhalten. 95-%-KI 6 bis 79 (eigene Rechnung, roh, wahrscheinlich überschätzt).
Beleg	[18] Laupacis 1988 · [19] Altman 1998

Kaplan-Meier-Kurve	
Einfach erklärt	Treppenkurve, die zeigt, wie viele Patient*innen zu jedem Zeitpunkt noch leben. Sie berücksichtigt, dass nicht alle gleich lange beobachtet wurden.
Fachlich	Schätzung der Überlebenswahrscheinlichkeit über die Zeit; zensierte Beobachtungen gehen bis zu ihrem Ausscheiden ein. Unter der Grafik steht die Zahl der Patient*innen unter Risiko (number at risk). Wird sie klein, ist das Kurvenende unsicher.
In ICRAKI	Abb. 1: Die Kurven trennen sich früh. Kaplan-Meier-Mortalität Tag 28: 53,6 vs. 44,8 %; Tag 90: 63,3 vs. 53,6 %.
Beleg	[21] Bland 1998

Hazard Ratio (HR) und Cox-Modell	
Einfach erklärt	Wie viel schneller treten Ereignisse in der einen Gruppe im Vergleich zur anderen auf, über die ganze Beobachtungszeit gemittelt?
Fachlich	Verhältnis der Ereignisraten (Hazards) über die Zeit, meist aus einem Cox-Proportional-Hazards-Modell. Die Annahme ist ein über die Zeit konstantes Verhältnis; sie wird z. B. mit Schoenfeld-Residuen geprüft. Die HR ist kein RR zu einem bestimmten Zeitpunkt.
In ICRAKI	HR 1,30 (1,04–1,63) für Tod bis Tag 90, adjustiert für die Stratifizierungsfaktoren; die Proportionalitätsannahme wurde laut Autor*innen geprüft.
Achtung	HR 1,30 heißt nicht „30 % mehr Tote“. Das wäre ein RR.
Beleg	[22] Spruance 2004 · [1] Gaudry 2026

Roh und adjustiert (engl. crude, adjusted)	
Einfach erklärt	Roh: einfach ausgezählt. Adjustiert: rechnerisch berücksichtigt, dass bestimmte Merkmale die Ergebnisse beeinflussen.
Fachlich	Adjustierung für die Stratifizierungsfaktoren gehört zur korrekten Analyse stratifizierter Studien. Das Cochran-Mantel-Haenszel-Verfahren kombiniert dazu Effekte über Strata. Post hoc adjustierte Analysen für Baseline-Unterschiede sind Sensitivitätsanalysen.
In ICRAKI	MAKE90: Cochran-Mantel-Haenszel, stratifiziert nach Flüssigkeitsbilanz und Katecholamindosis.
Beleg	[4] Kernan 1999 · [1] Gaudry 2026

Merksatz: Immer drei Zahlen nennen: absolute Differenz, relatives Maß, NNT/NNH, jeweils mit Konfidenzintervall und Zeitraum.

Modul 7: Unsicherheit – wie sicher ist der Unterschied?

> **Lernziel:** Konfidenzintervall und p-Wert korrekt deuten und typische Fehlschlüsse vermeiden.

Konfidenzintervall (KI) (engl. confidence interval, CI)	
Einfach erklärt	Der Bereich, in dem der wahre Effekt mit den Daten gut vereinbar ist. Breites KI: wenig Information. Schmales KI: viel Information.
Fachlich	Ein 95-%-KI enthält bei wiederholter Durchführung in 95 % der Fälle den wahren Wert. Es zeigt Größe und Präzision des Effekts und ist informativer als ein p-Wert. Schließt das KI den Nulleffekt ein (RR oder HR = 1, ARD = 0), ist der Befund auf dem 5-%-Niveau nicht signifikant.
In ICRAKI	MAKE90: ARD −1,22 bis 13,79 PP, also von „IHD minimal besser“ bis „IHD deutlich schlechter“. Mortalität: HR 1,04 bis 1,63, also schließt sie 1 aus.
Beleg	[23] Gardner 1986 · [24] Greenland 2016

p-Wert (engl. P value)	
Einfach erklärt	Wie überraschend wären diese Daten, wenn es in Wahrheit gar keinen Unterschied gäbe?
Fachlich	Wahrscheinlichkeit, unter der Nullhypothese und allen Modellannahmen ein mindestens so extremes Ergebnis zu beobachten. Er ist *nicht* die Wahrscheinlichkeit, dass die Nullhypothese stimmt, und misst nicht die Effektgröße.
In ICRAKI	MAKE90: p = 0,10. Mortalität roh (Fisher-Test, eigene Rechnung): p = 0,028.
Achtung	p = 0,049 und p = 0,051 sind praktisch dieselbe Evidenz.
Beleg	[24] Greenland 2016

„Nicht signifikant“ ist nicht „kein Unterschied“	
Einfach erklärt	Nichts gefunden zu haben heißt nicht, bewiesen zu haben, dass nichts da ist.
Fachlich	Ein nicht signifikantes Ergebnis kann mit klinisch relevanten Effekten vereinbar sein. Was die Daten zulassen, zeigt das KI.
In ICRAKI	Die Überlegenheit der CRRT bei MAKE90 war nicht signifikant (RR 0,98–1,20). Das KI schließt aber relevante Nachteile der IHD ein.
Beleg	[25] Altman 1995

Multiplizität und hierarchisches Testen (engl. multiplicity, hierarchical testing)	
Einfach erklärt	Wer oft genug würfelt, bekommt irgendwann einen Sechserpasch. Viele Tests bringen Zufallstreffer.
Fachlich	Viele Endpunkte und Vergleiche erhöhen die Rate falsch-positiver Befunde. Statistische Korrekturen sind vor allem nötig, wenn die klinische Entscheidung allein auf der Signifikanz einzelner Endpunkte beruht. Hierarchisches Testen legt eine Reihenfolge fest: Der nächste Test ist nur zulässig, wenn der vorige bestanden ist.
In ICRAKI	Zuerst Nichtunterlegenheit; weil nicht gezeigt, danach Überlegenheit der CRRT (nicht signifikant). Sekundäre KI sind ausdrücklich nicht adjustiert.
Beleg	[14] Schulz 2005 · [1] Gaudry 2026

Subgruppenanalyse und Interaktionstest (engl. subgroup analysis, test for interaction)	
Einfach erklärt	Wirkt die Behandlung bei einer Untergruppe anders, etwa bei Sepsis? Meist ist das Zufall.
Fachlich	Glaubwürdiger, wenn vorab festgelegt, wenige, mit vorhergesagter Richtung, durch einen Interaktionstest gestützt und biologisch plausibel. Zwei Subgruppen vergleicht man über die Differenz ihrer Effekte, nicht über „signifikant hier, nicht signifikant dort“.
In ICRAKI	Abb. 2: keine Heterogenität in vorab festgelegten Subgruppen; post hoc auch nicht für die Mortalität.
Beleg	[27] Sun 2010 · [26] Altman 2003 · [28] Wang 2007

Fragility Index	
Einfach erklärt	Wie viele Patient*innen müssten ein anderes Ergebnis haben, damit der Befund nicht mehr signifikant ist?
Fachlich	In der Gruppe mit weniger Ereignissen werden schrittweise Nicht-Ereignisse in Ereignisse umgewandelt, bis $p \geq 0{,}05$. In 399 RCTs aus Top-Journals betrug der Median 8; bei 25 % lag er bei 3 oder darunter. Er ist an die 0,05-Schwelle gebunden und gilt nur für dichotome Endpunkte, daher nur ergänzend.
In ICRAKI	Mortalität roh: Fragility Index 3 (eigene Rechnung, Fisher-Test).
Beleg	[29] Walsh 2014

Bayes-Interpretation (engl. Bayesian analysis; Prior, Posterior)	
Einfach erklärt	Man startet mit einer Vorannahme (Prior), etwa „große Effekte sind selten“, und aktualisiert sie mit den Studiendaten. Heraus kommt eine direkte Wahrscheinlichkeit: Wie wahrscheinlich ist ein Schaden von mindestens X?
Fachlich	Die Posterior-Verteilung kombiniert Prior und Daten. Sie erlaubt Aussagen wie „$P(HR > 1) = 0{,}95$“. Priors müssen vorab begründet werden (skeptisch, neutral, optimistisch), sonst lassen sie sich ergebnisorientiert wählen.
In ICRAKI	Eigene Näherung: Mit flachem Prior $P(HR > 1)$ 0,99. Mit skeptischem Prior Posterior-HR 1,14 (0,97–1,33), $P(HR > 1)$ 0,95. Das ist eine Illustration, keine formale Reanalyse.
Beleg	[30] Zampieri 2021 · [31] Goligher 2018

Merksatz: Erst das KI lesen: Wo liegt der Punktschätzer, wo die ungünstige Grenze, liegt der Nulleffekt drin? Der p-Wert ergänzt nur.

Modul 8: Überwachung und Frühabbruch

> **Lernziel:** die Rolle des DSMB verstehen und die Effektüberschätzung abgebrochener Studien einordnen.

Data Safety Monitoring Board (DSMB) und Zwischenanalyse (engl. data monitoring committee, interim analysis)	
Einfach erklärt	Ein unabhängiges Gremium schaut während der Studie verdeckt auf die Daten und darf den Abbruch empfehlen, wenn eine Gruppe deutlich geschädigt wird.
Fachlich	Unabhängige Expert*innen überwachen Sicherheit und Wirksamkeit. Abbruchgrenzen sollten vorab statistisch definiert sein.
In ICRAKI	Geplant: zwei deskriptive Zwischenanalysen der 28-Tage-Mortalität ohne Entscheidungsregeln. Durchgeführt: drei; danach Abbruchempfehlung.
Beleg	[32] DeMets 2016 · [1] Gaudry 2026

Frühabbruch (truncated RCT)	
Einfach erklärt	Hört man genau dann auf, wenn der Unterschied gerade besonders groß aussieht, überschätzt man ihn.
Fachlich	Abgebrochene Studien zeigen systematisch größere Effekte als nicht abgebrochene zur gleichen Frage: Verhältnis der relativen Risiken 0,71 (0,65–0,77), am stärksten bei unter 500 Ereignissen. Frühabbrüche werden oft unzureichend berichtet. Die Daten stammen aus Abbrüchen wegen Nutzens; für Schaden gilt die Logik analog.
In ICRAKI	311 Todesfälle, Abbruch wegen eines Schadenssignals. Die Autor*innen schreiben selbst, dass die Richtung belastbarer ist als die Größe.
Beleg	[34] Bassler 2010 · [33] Montori 2005

Modul 9: Bewertungswerkzeuge

> **Lernziel:** die Werkzeuge ihrer jeweiligen Frage zuordnen.

Werkzeug	Beantwortet	Einfach erklärt	Fachlich	Beleg
CONSORT 2025	Vollständig berichtet?	Checkliste: Was muss in einem RCT-Artikel stehen?	Checkliste und Flussdiagramm; 2025 sieben neue Punkte und ein Abschnitt Open Science. Misst Berichtsqualität, nicht Bias.	[35]
CONSORT-NI	NI korrekt berichtet?	Zusatz für Nichtunterlegenheitsstudien.	Margin-Begründung, ITT und Per-Protocol, KI gegen Margin.	[8]
SPIRIT 2025	Protokoll vollständig?	Checkliste für den Plan vor der Studie.	Grundlage für den Abgleich Protokoll – Publikation.	[7]
JAMA Users' Guides / CASP	Valide? Ergebnis? Anwendbar?	Leitfragen in fester Reihenfolge, ohne Punkte.	CASP basiert auf den Users' Guides, ausdrücklich ohne Score; Antworten Ja / Nein / Unklar.	[36, 37]

RoB 2	Wie verzerrt, pro Endpunkt?	Fünf Prüfbereiche mit Ampelurteil.	Randomisierung, Abweichungen, fehlende Daten, Messung, selektives Berichten. Signalfragen und Algorithmus, Werkzeug online.	[38]
JBI-RCT-Tool	Wie verzerrt?	Einfachere Alternative zu RoB 2.	Überarbeitete Fassung 2023, ebenfalls nach Bias-Domänen geordnet.	[39]
PRECIS-2	Wie alltagsnah?	Ist die Studie eher Labor oder eher Stationsalltag?	Neun Domänen auf einer Skala von explanatorisch bis pragmatisch; hilft bei der Übertragbarkeit.	[40]
GRADE	Wie sicher ist die Gesamtevidenz?	Note für das Vertrauen in einen Effekt: hoch bis sehr niedrig.	RCT starten hoch; Abwertung für Bias, Inkonsistenz, Indirektheit, Unpräzision, Publikationsbias. Bewertet den Endpunkt über alle Studien.	[41]

> **Merksatz:** CONSORT fragt „Steht es drin?“, RoB 2 fragt „Stimmt es?“, PRECIS-2 fragt „Passt es zu uns?“ und GRADE fragt „Wie sicher sind wir insgesamt?“.

Modul 10: Rechenwerkstatt ICRAKI

Lernziel: die zentralen Kennzahlen einer Studie selbst nachrechnen. Alle Rohzahlen stammen aus Tabelle 2 [1], die Rechnungen sind eigene.

#	Schritt	Rechnung	Bedeutung
1	Risiken	IHD 164/260 = 0,631; CRRT 147/275 = 0,535	63,1 vs. 53,5 % verstorben bis Tag 90
2	Absolute Differenz	0,631 − 0,535 = 0,096 → 9,6 PP; 95-%-KI 1,3 bis 17,8 PP (Newcombe)	KI schließt 0 aus
3	Relatives Risiko	0,631 / 0,535 = 1,18	18 % relativ mehr Todesfälle
4	Odds Ratio	(164/96) / (147/128) = 1,71 / 1,15 = 1,49	OR > RR, weil das Ereignis häufig ist [20]
5	NNH mit KI	1 / 0,096 ≈ 10; KI: 1/0,178 = 5,6 bis 1/0,013 = 79	Ein zusätzlicher Todesfall pro etwa 10 Behandelten
6	Nichtunterlegenheit MAKE90	Obere KI-Grenze 13,79 > Margin 7,5	NI nicht gezeigt (siehe Abbildung)
7	Überlegenheit CRRT MAKE90	Untere KI-Grenze −1,22 < 0	Überlegenheit nicht gezeigt → inkonklusiv
8	Mortalität HR	HR 1,30 (1,04–1,63): untere Grenze > 1; obere Grenze > NI-Margin 1,27	Signal für Schaden unter IHD (sekundär, nicht adjustiert)
9	Fragility Index	3 zusätzliche Todesfälle im CRRT-Arm → Fisher p ≥ 0,05	Fragiler Befund [29]
10	Bayes-Näherung	log-HR 0,262; SE = (ln 1,63 − ln 1,04) / 3,92 = 0,115. Skeptischer Prior: Mittelwert 0, SD = ln 1,25 / 1,96	Posterior-HR 1,14 (0,97–1,33); P(HR > 1) 0,95

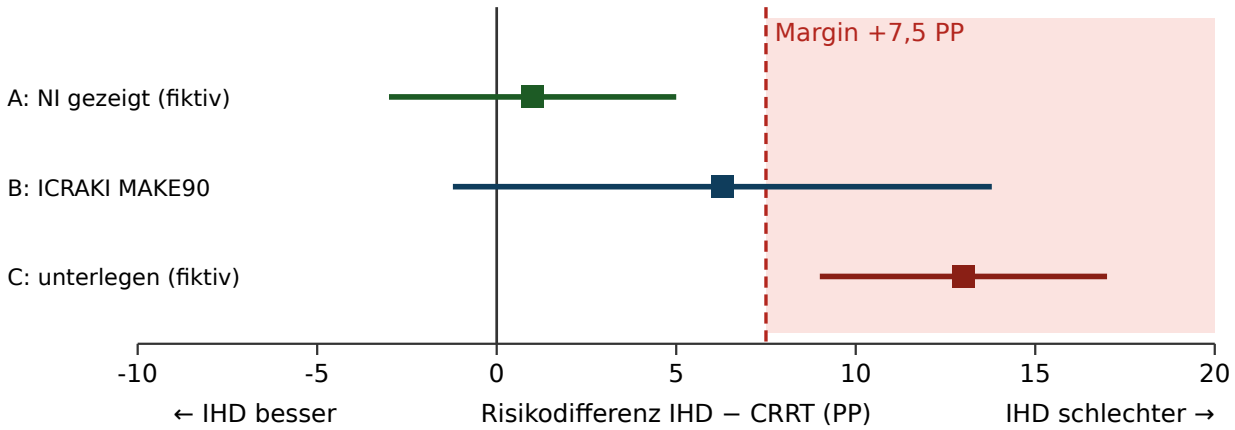


Abbildung: Lesen einer Nichtunterlegenheitsstudie. Entscheidend ist die rechte (ungünstige) KI-Grenze gegenüber der Margin. A: ganz links der Margin, also NI gezeigt. B (ICRAKI): Das KI überschreitet die Margin und schließt 0 ein, also inkonklusiv. C: ganz rechts der Margin und rechts von 0, also unterlegen. Schema nach [8, 9].

Modul 11: Selbsttest

> **Lernziel:** das Gelernte an ICRAKI anwenden. Erst selbst antworten, dann die Lösungen auf der nächsten Seite lesen.

1. MAKE90 hatte p = 0,10. Sind IHD und CRRT damit gleichwertig?
2. Was bedeutet NNH 10 (6–79) hier konkret?
3. Warum ist die OR (1,49) größer als das RR (1,18)?
4. Ist das offene Design für den Endpunkt Tod ein großes Problem?
5. Warum ist „mITT“ ein Warnsignal?
6. Eine Studie wird nach 311 Todesfällen wegen Schadens abgebrochen. Was erwarten Sie für die wahre Effektgröße?
7. In der Subgruppe „Sepsis“ ist ein Effekt signifikant, bei „keine Sepsis“ nicht. Wirkt die Behandlung nur bei Sepsis?
8. Ein Paper erfüllt alle CONSORT-Punkte. Ist es unverzerrt?

Spickzettel: 10 Faustregeln

1. Erst Validität, dann Ergebnisse, dann Anwendbarkeit [36, 37].
2. Verdeckte Zuteilung ist wichtiger als die Bezeichnung „doppelblind“ [3, 6].
3. Randomisiert minus ausgewertet: jede Differenz begründen lassen [11, 12].
4. Absolute *und* relative Effekte nennen, dazu NNT/NNH mit KI [18, 19].
5. Konfidenzintervall vor p-Wert [23, 24].
6. Nicht signifikant ist nicht gleich wie kein Unterschied [25].
7. Bei Nichtunterlegenheit: ungünstige KI-Grenze gegen die Margin [8, 9].
8. Kombinierte Endpunkte in ihre Komponenten zerlegen [15, 16].
9. Frühabbruch heißt: Effekt wahrscheinlich überschätzt [34].
10. Subgruppen nur mit Interaktionstest ernst nehmen [26, 27].

Lösungen zum Selbsttest

1. Nein. Nicht signifikant heißt nicht gleich. Das KI (−1,22 bis 13,79 PP) schließt relevante Nachteile der IHD ein. [25]
2. Pro etwa 10 Patient*innen mit initialer IHD statt CRRT ein zusätzlicher Todesfall bis Tag 90. Plausibel ist alles zwischen 1 von 6 und 1 von 79; der Wert ist roh und wahrscheinlich überschätzt. [19]
3. Bei häufigen Ereignissen (Mortalität über 50 %) entfernt sich die OR stärker von 1 als das RR. [20]
4. Eher wenig, weil Tod objektiv ist. Für die Dialyseabhängigkeit, über die unverblindete Ärzt*innen entschieden, ist es relevanter. [6]
5. Es bedeutet Ausschlüsse nach Randomisierung. Ihre Begründung und Verteilung prüfen (hier 12 vs. 7). [12]
6. Eher kleiner als beobachtet; abgebrochene Studien mit unter 500 Ereignissen überschätzen Effekte am stärksten. [34]
7. Nicht daraus ableitbar. Entscheidend ist der Interaktionstest bzw. der direkte Vergleich der Effekte. [26, 27]
8. Nein. CONSORT prüft die Vollständigkeit des Berichts. Das Verzerrungsrisiko prüft RoB 2. [35, 38]

Hinweis: Dieser Leitfaden vermittelt Methodenwissen. Studienbewertungen und abgeleitete Kennzahlen ersetzen keine individuelle klinische Beurteilung. Die Zahlen zu ICRAKI sind teils eigene Berechnungen aus publizierten Rohdaten, unadjustiert und als solche gekennzeichnet.

Literatur mit Erläuterung

1. Gaudry S, Boubaya M, Louis G, et al. Intermittent or continuous renal-replacement therapy in acute kidney injury. N Engl J Med. 2026;395(17):1677-90. doi:10.1056/NEJMoa2608397
   Beispielstudie dieses Leitfadens.
2. Schulz KF, Grimes DA. Generation of allocation sequences in randomised trials: chance, not choice. Lancet. 2002;359(9305):515-9. doi:10.1016/S0140-6736(02)07683-3
   Wie eine Zufallsfolge erzeugt wird; einfache, Block- und stratifizierte Randomisierung.
3. Schulz KF, Grimes DA. Allocation concealment in randomised trials: defending against deciphering. Lancet. 2002;359(9306):614-8. doi:10.1016/S0140-6736(02)07750-4
   Warum die nächste Zuteilung geheim bleiben muss.
4. Kernan WN, Viscoli CM, Makuch RW, et al. Stratified randomization for clinical trials. J Clin Epidemiol. 1999;52(1):19-26. doi:10.1016/s0895-4356(98)00138-3
   Wann Stratifizierung nützt.
5. Roberts C, Torgerson DJ. Understanding controlled trials: baseline imbalance in randomised controlled trials. BMJ. 1999;319(7203):185. doi:10.1136/bmj.319.7203.185
   Kurzer Artikel zu Ungleichgewichten trotz Randomisierung.
6. Schulz KF, Grimes DA. Blinding in randomised trials: hiding who got what. Lancet. 2002;359(9307):696-700. doi:10.1016/S0140-6736(02)07816-9
   Wer verblindet wird und warum harte Endpunkte robust sind.
7. Chan AW, Boutron I, Hopewell S, et al. SPIRIT 2025 statement: updated guideline for protocols of randomized trials. Nat Med. 2025;31(6):1784-92. doi:10.1038/s41591-025-03668-w
   Standard für Studienprotokolle.
8. Piaggio G, Elbourne DR, Pocock SJ, et al. Reporting of noninferiority and equivalence randomized trials: extension of the CONSORT 2010 statement. JAMA. 2012;308(24):2594-604. doi:10.1001/jama.2012.87802
   Berichtsstandard für Nichtunterlegenheitsstudien.
9. Mauri L, D'Agostino RB. Challenges in the design and interpretation of noninferiority trials. N Engl J Med. 2017;377(14):1357-67. doi:10.1056/NEJMra1510063
   Gut lesbare Einführung in Nichtunterlegenheit.
10. Schulz KF, Grimes DA. Sample size calculations in randomised trials: mandatory and mystical. Lancet. 2005;365(9467):1348-53. doi:10.1016/S0140-6736(05)61034-3
    Fallzahl, Power und ihre subjektiven Annahmen.

11. Hollis S, Campbell F. What is meant by intention to treat analysis? Survey of published randomised controlled trials. BMJ. 1999;319(7211):670-4. doi:10.1136/bmj.319.7211.670
Was ITT ist und wie oft es falsch angewendet wird.

12. Abraha I, Montedori A. Modified intention to treat reporting in randomised controlled trials: systematic review. BMJ. 2010;340:c2697. doi:10.1136/bmj.c2697
mITT: uneinheitlich definiert, häufig mit Ausschlüssen nach Randomisierung.

13. Sterne JAC, White IR, Carlin JB, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. BMJ. 2009;338:b2393. doi:10.1136/bmj.b2393
Umgang mit fehlenden Daten.

14. Schulz KF, Grimes DA. Multiplicity in randomised trials I: endpoints and treatments. Lancet. 2005;365(9470):1591-5. doi:10.1016/S0140-6736(05)66461-6
Viele Vergleiche erzeugen Zufallsbefunde.

15. Montori VM, Permanyer-Miralda G, Ferreira-González I, et al. Validity of composite end points in clinical trials. BMJ. 2005;330(7491):594-6. doi:10.1136/bmj.330.7491.594
Wie man kombinierte Endpunkte liest.

16. Ferreira-González I, Busse JW, Heels-Ansdell D, et al. Problems with use of composite end points in cardiovascular trials: systematic review of randomised controlled trials. BMJ. 2007;334(7597):786. doi:10.1136/bmj.39136.682083.AE
Empirische Probleme kombinierter Endpunkte.

17. Austin PC, Lee DS, Fine JP. Introduction to the analysis of survival data in the presence of competing risks. Circulation. 2016;133(6):601-9. doi:10.1161/CIRCULATIONAHA.115.017719
Konkurrierende Risiken und Fine-Gray-Modell.

18. Laupacis A, Sackett DL, Roberts RS. An assessment of clinically useful measures of the consequences of treatment. N Engl J Med. 1988;318(26):1728-33. doi:10.1056/NEJM198806303182605
Originalarbeit zur Number needed to treat.

19. Altman DG. Confidence intervals for the number needed to treat. BMJ. 1998;317(7168):1309-12. doi:10.1136/bmj.317.7168.1309
Konfidenzintervalle für NNT, NNTB und NNTH.

20. Davies HT, Crombie IK, Tavakoli M. When can odds ratios mislead? BMJ. 1998;316(7136):989-91. doi:10.1136/bmj.316.7136.989
Odds Ratio und relatives Risiko bei häufigen Ereignissen.

21. Bland JM, Altman DG. Survival probabilities (the Kaplan-Meier method). BMJ. 1998;317(7172):1572. doi:10.1136/bmj.317.7172.1572
Kaplan-Meier auf einer Seite.

22. Spruance SL, Reid JE, Grace M, Samore M. Hazard ratio in clinical trials. Antimicrob Agents Chemother. 2004;48(8):2787-92. doi:10.1128/AAC.48.8.2787-2792.2004
Was die Hazard Ratio misst und was nicht.

23. Gardner MJ, Altman DG. Confidence intervals rather than P values: estimation rather than hypothesis testing. Br Med J (Clin Res Ed). 1986;292(6522):746-50. doi:10.1136/bmj.292.6522.746
Klassiker: Konfidenzintervalle statt p-Werte.

24. Greenland S, Senn SJ, Rothman KJ, et al. Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. Eur J Epidemiol. 2016;31(4):337-50. doi:10.1007/s10654-016-0149-3
25 häufige Fehlinterpretationen von p, KI und Power.

25. Altman DG, Bland JM. Absence of evidence is not evidence of absence. BMJ. 1995;311(7003):485. doi:10.1136/bmj.311.7003.485
„Nicht signifikant“ heißt nicht „kein Unterschied“.

26. Altman DG, Bland JM. Interaction revisited: the difference between two estimates. BMJ. 2003;326(7382):219. doi:10.1136/bmj.326.7382.219
Subgruppen richtig vergleichen.

27. Sun X, Briel M, Walter SD, Guyatt GH. Is a subgroup effect believable? Updating criteria to evaluate the credibility of subgroup analyses. BMJ. 2010;340:c117. doi:10.1136/bmj.c117
Kriterien für glaubwürdige Subgruppeneffekte.

28. Wang R, Lagakos SW, Ware JH, et al. Statistics in medicine – reporting of subgroup analyses in clinical trials. N Engl J Med. 2007;357(21):2189-94. doi:10.1056/NEJMsr077003
NEJM-Vorgaben für Subgruppen.

29. Walsh M, Srinathan SK, McAuley DF, et al. The statistical significance of randomized controlled trial results is frequently fragile: a case for a Fragility Index. J Clin Epidemiol. 2014;67(6):622-8. doi:10.1016/j.jclinepi.2013.10.019
Definition des Fragility Index.

30. Zampieri FG, Casey JD, Shankar-Hari M, et al. Using Bayesian methods to augment the interpretation of critical care trials. An overview of theory and example reanalysis of the Alveolar Recruitment for Acute Respiratory Distress

Syndrome Trial. Am J Respir Crit Care Med. 2021;203(5):543-52. doi:10.1164/rccm.202006-2381CP
Bayes-Methoden für Intensivmediziner*innen erklärt.

31. Goligher EC, Tomlinson G, Hajage D, et al. Extracorporeal membrane oxygenation for severe acute respiratory distress syndrome and posterior probability of mortality benefit in a post hoc Bayesian analysis of a randomized clinical trial. JAMA. 2018;320(21):2251-9. doi:10.1001/jama.2018.14276
EOLIA-Reanalyse als Bayes-Lehrbeispiel.

32. DeMets DL, Ellenberg SS. Data monitoring committees – expect the unexpected. N Engl J Med. 2016;375(14):1365-71. doi:10.1056/NEJMra1510066
Aufgaben unabhängiger Monitoring-Komitees.

33. Montori VM, Devereaux PJ, Adhikari NK, et al. Randomized trials stopped early for benefit: a systematic review. JAMA. 2005;294(17):2203-9. doi:10.1001/jama.294.17.2203
Wie häufig und wie schlecht berichtet Frühabbrüche sind.

34. Bassler D, Briel M, Montori VM, et al. Stopping randomized trials early for benefit and estimation of treatment effects: systematic review and meta-regression analysis. JAMA. 2010;303(12):1180-7. doi:10.1001/jama.2010.310
Ausmaß der Effektüberschätzung bei Frühabbruch.

35. Hopewell S, Chan AW, Collins GS, et al. CONSORT 2025 statement: updated guideline for reporting randomised trials. BMJ. 2025;389:e081123. doi:10.1136/bmj-2024-081123
Aktuelle Berichtscheckliste für RCTs.

36. Critical Appraisal Skills Programme. CASP Checklist: Randomised Controlled Trials. 2024. casp-uk.net (PDF)
Didaktische Checkliste ohne Score.

37. Guyatt GH, Sackett DL, Cook DJ. Users' guides to the medical literature. II. How to use an article about therapy or prevention. A. Are the results of the study valid? JAMA. 1993;270(21):2598-601. doi:10.1001/jama.270.21.2598
Ursprung der Struktur Validität – Ergebnisse – Anwendbarkeit.

38. Sterne JAC, Savović J, Page MJ, et al. RoB 2: a revised tool for assessing risk of bias in randomised trials. BMJ. 2019;366:l4898. Werkzeug: riskofbias.info. doi:10.1136/bmj.l4898 · riskofbias.info
Referenzstandard für das Verzerrungsrisiko.

39. Barker TH, Stone JC, Sears K, et al. The revised JBI critical appraisal tool for the assessment of risk of bias for randomized controlled trials. JBI Evid Synth. 2023;21(3):494-506. doi:10.11124/JBIES-22-00430
Einfachere Alternative zu RoB 2.

40. Loudon K, Treweek S, Sullivan F, et al. The PRECIS-2 tool: designing trials that are fit for purpose. BMJ. 2015;350:h2147. doi:10.1136/bmj.h2147
Pragmatisch oder explanatorisch: neun Domänen.

41. Guyatt GH, Oxman AD, Vist GE, et al. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. BMJ. 2008;336(7650):924-6. doi:10.1136/bmj.39489.470347.AD
Einführung in GRADE.

Verifikation: [1] Volltext; [2–35, 37–41] über PubMed (03.11.2026); [36] über die Website von CASP. Links führen zu DOI bzw. Herausgeber*innen-Seiten; der Volltextzugang hängt von den Lizenzen der Klinik ab.