Wesentliche neue EbM-relevante Methoden 2024/2025 – Medizinische Biometrie

ARBEITSPAPIER

Projekt: GA26-04 Version: 1.0 Stand: 20.03.2026 IQWiG-Berichte – Nr. 2210

DOI: 10.60584/GA26-04

Impressum

Herausgeber

Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen

Thema

Wesentliche neue EbM-relevante Methoden 2024/2025 – Medizinische Biometrie

Auftraggeber

Bearbeitung im Rahmen des Generalauftrags

Interne Projektnummer

GA26-04

DOI-URL

https://doi.org/10.60584/GA26-04

Anschrift des Herausgebers

Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen
Siegburger Str. 237
50679 Köln

Tel.: +49 221 35685-0
Fax: +49 221 35685-1
E-Mail: berichte@iqwig.de
Internet: www.iqwig.de

ISSN: 1864-2500

Zitiervorschlag

Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen. Wesentliche neue EbM-relevante Methoden 2024/2025 – Medizinische Biometrie; Arbeitspapier [online]. 2026 [Zugriff: TT.MM.JJJJ]. URL: https://doi.org/10.60584/GA26-04.

Schlagwörter

Statistik, Übersichtsliteratur, Evidenzsynthese

Keywords

Statistics as Topic, Review Literature as Topic, Evidence Synthesis

Für die Inhalte des Berichts ist allein das IQWiG verantwortlich.

Mitarbeiterinnen und Mitarbeiter des IQWiG

- Lars Beckmann
- Ralf Bender
- Moritz Felsch
- Catharina Brockhaus
- Charlotte Guddat
- Ulrich Grouven
- Sandra Hamacher
- Katharina Hirsch
- Corinna Kiefer
- Jona Lilienthal
- Fabian Lotz
- Martina Messow
- Max Oberste-Frielinghaus
- Mattea Patt
- Katherine Rascher
- Anke Schulz
- Wiebke Sieben
- Guido Skipka
- Sibylle Sturtz
- Kerstin van der Leck
- Frank Weber

Inhaltsverzeichnis

Abkürzungsverzeichnis

Abkürzung	**Bedeutung**
AgD	aggregierte Daten
AI	artificial Intelligence
AIPTW-IPCW	double-robust augmented IPTW estimator combined with IPCW
AMSTAR	A Measurement Tool to Assess Systematic Reviews
ATE	Average Treatment Effect
ATT	Average Treatment Effect on the Treated
CONSORT	Consolidated Standards of Reporting Trials
FCS	fully conditional Specification
GAMM	generalised additive mixed Model
G-BA	Gemeinsamer Bundesausschuss
GRADE	Grading of Recommendations Assessment, Development and Evaluation
HR	Hazard Ratio
IPD	individual Patient Data (individuelle Patientendaten)
IPCW	inverse probability of censoring weights
IPTW	inverse probability of treatment weighting
IQWiG	Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen
ITT	Intention to treat
JM	joint Modeling
LLM	large Language Model
MAIC	matching-adjusted indirect Comparison
MSE	Mean Squared Error
OS	Gesamtüberleben
OWATT	Overlap Weighted Average Treatment Effect on the Treated
PFS	progressionsfreie Überleben
PP	per Protocol
PPS	Per-Protocol-Set
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PRISMA-LSR	Preferred Reporting Items for Systematic Reviews and Meta-Analyses for Living Systematic Reviews

Abkürzung	Bedeutung
PRISMA-NMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses for Network Meta-Analyses
PRO	Patient-reported Outcome (patientenberichteter Endpunkt)
PROBAST	Prediction model Risk Of Bias Assessment Tool
PS	Propensity-Score
PWP-Modell	Prentice-Williams-Peterson-Modell
RCT	randomized controlled Trial (randomisierte kontrollierte Studie)
RMST	restricted Mean Survival Time
RoB	Risk of Bias
RoB-NMA	Risk of Bias for Network Meta-Analyses
ROBUST-RCT	Risk Of Bias instrument for Use in SysTematic reviews-for Randomised Controlled Trials
ROC	Receiver-Operating-Characteristic
RPSFTM	Rank-Preserving-Structural-Failure-Time-Modelle
SISAQOL-IMI-Initiative	Setting International Standards in Analysing Patient-Reported Outcomes and Quality of Life Endpoints - Innovative Medicines Initiative
SPIRIT	Standard Protocol Items: Recommendations for Interventional Trials
TARGET	Transparent Reporting of observational Studies emulating a Target Trial
TMLE	Targeted Maximum Likelihood Estimation

Das Ressort "*Medizinische Biometrie*" hat im Jahr 2025 ein regelmäßiges Literaturscreening durchgeführt, um über neue Methoden in der Medizinischen Statistik und speziell über aktuelle methodische Entwicklungen, die für systematische Übersichten wichtig sind, laufend informiert zu sein. Dabei wurden von 15 wissenschaftlichen Zeitschriften aus den Bereichen Biostatistik, Epidemiologie und Medizin (siehe Anhang) alle Ausgaben des Publikationszeitraums Oktober 2024 bis September 2025 begutachtet. Folgende Publikationen schienen uns von besonderer Wichtigkeit und / oder für die Arbeit des IQWiG besonders hilfreich zu sein.

1 Methodische Guidelines

Sowohl für das CONSORT-Statement als auch für das SPIRIT Statement wurden neue Versionen veröffentlicht, das **CONSORT-2025-** [1] und das **SPIRIT-2025-Statement** [2]. Die Überarbeitungen dienen dazu, neuen methodischen Entwicklungen, neuer Evidenz und Rückmeldungen von Anwendern gerecht zu werden. Zu beiden Statements sind zusätzlich weitere Erklärungen, Ausführungen und Beispiele verfügbar [3,4].

Eine Erweiterung des CONSORT 2010 Statements für **Cluster-randomisierte Cross-over-Studien** wurde von McKenzie et al. [5] veröffentlicht. Dabei wurden relevante Items aus bereits bestehenden CONSORT Guidelines mit neuen Items, die sich auf die speziellen methodologischen Schwierigkeiten dieses Studiendesigns beziehen, zu einer Checkliste mit 28 Items mit 43 Unteritems vereint. Zusätzlich wurden entsprechende Checklisten für Abstracts und für die Fallzahlplanung erstellt.

Im Jahr 2023 wurden Erweiterungen des CONSORT 2010 Statements und des SPIRIT 2013 Statements für **klinischen Studien mit faktoriellen Designs** veröffentlicht. Kahan et al. [6] geben hierzu Erklärungen und weitere Ausführungen zu den damals neu hinzugekommenen oder abgewandelten Items, einschließlich illustrierender Beispiele.

Das PRISMA-2020-Statement wurde zu **PRISMA-LSR** [7] erweitert, das für Living Systematic Reviews anzuwenden ist. Es wurden 4 neue Items hinzugefügt und unter einigen der bestehenden Items neue Elemente ergänzt, um die Aspekte, die für Living Systematic Reveiws spezifisch sind, zu erfassen. Darüber hinaus werden verschiedene Arten von Flow Charts vorgeschlagen, um die Recherche mit ihren Aktualisierungen zu beschreiben.

Das **TARGET** (Transparent Reporting of observational Studies emulating a Target Trial)-Statement [8] ist eine Richtlinie zur Berichterstattung von Beobachtungsstudien, die zum Ziel haben, kausale Effekte durch Target Trial Emulation zu schätzen. Es handelt sich um eine Checkliste mit 21 Items, die entsprechend der üblichen Kapitel einer Publikation strukturiert sind. Ziel ist es, die Qualität von Publikationen zu Beobachtungsstudien, die Target Trials emulieren, zu verbessern.

Im Rahmen der **SISAQOL-IMI**-Initiative diskutieren Thomassen et al. die Auswertung Patienten-berichteter-Endpunkte (PROs) in 1-armigen [9] bzw. 2-armigen Studien [10] unter Berücksichtigung interkurrenter Ereignisse bei Verwendung unterschiedlicher Estimands. Anhand eines Beispiels werden die unterschiedlichen Strategien veranschaulicht.

2 Beurteilung von Studien / EBM allgemein

Whiting et al. [11] beschreiben und erläutern das **LATITUDES-Netzwerk** (www.latitudes-network.org), das eine organisierte Sammlung von Bewertungsinstrumenten für systematische Übersichten darstellt.

Die im Jahr 2011 begonnene **GRADE-Serie** des Journal of Clinical Epidemiology (siehe Berichte "*Wesentliche neue EbM-relevante Methoden 2010/2011 bis 2023/2024*") wird fortgesetzt. In der GRADE Guideline Nr. 39 [12] wird der GRADE-ADOLOPMENT-Ansatz aus 2017 (Schünemann et al., JCE 2017, 81: 101-110) zur Anpassung und Kontextualisierung von Guidelines konkreter operationalisiert und verfeinert. Murad et al. [13] beschreiben, wie mithilfe von Prädiktionsintervallen die GRADE-Domänen Ungenauigkeit und Inkonsistenz gleichzeitig bewertet werden können. Dieser Ansatz muss allerdings noch weiter untersucht und getestet werden.

Der GRADE-Ansatz zur Bewertung der Qualität der Evidenz hat sich über einen langen Zeitraum entwickelt und wurde immer wieder erweitert und weiterentwickelt. Dadurch ist eine große Zahl an Publikationen entstanden und der Ansatz ist zunehmend komplexer geworden, was dem Nutzer die Verwendung erschwert. Um dem entgegenzuwirken, hat eine Autorengruppe eine Reihe von 7 Publikationen unter der Bezeichnung **Core GRADE** erstellt, um die essenziellen Elemente von GRADE strukturiert darzustellen und damit leichter zugänglich zu machen [14]. Die Publikationsreihe soll umfassend sein und damit überflüssig machen, auf vorausgegangene GRADE-Publikationen zurückzugreifen. Core GRADE ist zwar weitgehend konsistent mit GRADE, jedoch nicht identisch. Die 7 Publikationen beinhalten zunächst eine Einführung in den Core GRADE Ansatz [15], in den weiteren werden die Themen Ziel der Bewertung der Vertrauenswürdigkeit der Evidenz und Abwertung für fehlende Genauigkeit [16], Inkonsistenz [17], Indirektheit [18], Risk of Bias, Publikationsbias und Möglichkeiten der Aufwertung der Vertrauenswürdigkeit der Evidenz [19], , Darstellung der Ergebnisse [20] und Prinzipien zur Ableitung von Empfehlungen und Entscheidungen [21].

Mit **RoB-NMA** stellen Lunny et al. [22] ein Tool vor, das dazu dienen soll, strukturiert das Verzerrungspotenzial einer einzelnen (bereits durchgeführten) Netzwerk-Metaanalyse in einem systematischen Review zu bewerten. Es besteht aus 17 Items in drei Domänen: Interventionen und Netzwerkgeometrie, Effektmodifikatoren und statistische Synthese. Innerhalb jeder Domäne gibt es eine Reihe von Aussagen, die jeweils mit „true“, „probably true“, „probably false“, „false“ und „no Information“ bewertet werden. Darauf basierend wird dann jede Domäne mit „low Risk of Bias“, „high Risk of Bias“ oder „some Concern“ bewertet.

Das PROBAST (Prediction model Risk Of Bias Assessment Tool) wurde überarbeitet, um neuen methodischen Entwicklungen im Bereich von Modellen zur Prädiktion von Endpunkten auf Patientenebene, insbesondere unter Verwendung von KI, gerecht zu werden. Das neue Tool,

PROBAST+AI [23], ersetzt PROBAST 2019. Die wichtigsten Änderungen sind: Die Bewertung von Qualität, Verzerrungspotenzial, und Anwendbarkeit erfolgt nun unabhängig von der verwendeten Methode zur Modellierung; die Modellierung wird klarer in zwei Phasen, Modellentwicklung und -validierung unterteilt; es wird unterschieden zwischen Validierung am Lerndatensatz (apparent Performance), interner Validierung und externer Validierung. Das Tool kann sowohl beim Erstellen eines systematischen Reviews von Prädiktionsmodellen, als auch bei der Auswahl eines Prädiktionsmodells zum Einsatz kommen.

Das Risk Of Bias instrument for Use in SysTematic reviews-for Randomised Controlled Trials (**ROBUST-RCT**) [24] wurde entwickelt, da die Autoren fanden, dass die bisher verfügbaren Instrumente für Bewertungen des Verzerrungspotenzials bei RCTs zum Teil nicht spezifisch genug sind. In bisherigen Instrumenten werden zum Teil Aspekte in die Beurteilung miteinbezogen, die nicht direkt mit dem Verzerrungspotenzial zusammenhängen. Zudem seien sie zu wenig benutzerfreundlich und legen zu viel Gewicht auf methodologische Genauigkeit. ROBUST-RCT soll die optimale Balance zwischen Einfachheit und methodologischer Strenge bieten. Es enthält 6 zentrale Items: Erzeugung der Randomisierungssequenz, Allocation Concealment, Verblindung der Teilnehmenden, Verblindung der Behandelnden Personen, Verblindung derer, die die Endpunkte erheben und die Anzahl nicht in die Auswertung eingehender Teilnehmenden. Zusätzlich gibt es acht optionale Items. Eine Einschränkung von ROBUST-RCT ist jedoch, dass es nur für RCTs mit parallelem Gruppendesign anwendbar ist und nicht für Cluster-randomisierte Studien oder Cross-over-Studien.

Viana et al. [25] untersuchten, wie sich die Anwendung des ursprünglichen Cochrane-Tools „Risk of Bias“ (RoB) im Vergleich zur **überarbeiteten Version RoB2** auf die Bewertung von Bias in randomisierten klinischen Studien (RCTs) in der Zahnmedizin auswirkt. Die Analyse ergab, dass nur rund ein Drittel der Studien mit beiden Instrumenten als „low Risk of Bias“ eingestuft wurde. Knapp 30 % der Studien, die mit RoB als „low Risk“ bewertet worden waren, wurden von RoB2 in „some Concerns“ herabgestuft, und rund 37 % sogar in „high Risk“. Auch bei den anderen Kategorien zeigten sich deutliche Verschiebungen, sodass insgesamt eine nur geringe Übereinstimmung zwischen den beiden Instrumenten festgestellt wurde. Die Ergebnisse verdeutlichen, dass RoB2 tendenziell strengere Bewertungen liefert als das ursprüngliche RoB-Tool. Dadurch könnten frühere Einschätzungen das Risiko für Bias unterschätzt haben, was Auswirkungen auf systematische Reviews und Leitlinien hat. Die Autoren empfehlen, methodische Standards und Reporting-Vorgaben wie CONSORT konsequenter umzusetzen und Studiendesigns zu wählen, die Verzerrungen minimieren.

3 Meta-Analysen

3.1 Allgemeines

Held et al. [26] stellen ein neues metaanalytisches Verfahren (Edgington-Methode) vor und vergleichen sie in einer Simulationsstudie u. a. mit Modellen mit festen Effekten und der Methode nach Knapp-Hartung. Die **Edgington-Methode** ist ein spezielles Verfahren für die Berechnung von p-Wert-Funktionen. Für 3, 5, und mehr Studien mit variierender Heterogenität zeigen die Autoren, dass die Edgington-Methode in heterogenen Situationen bessere Eigenschaften hat als die Standardverfahren für Fixed-Effect-Metaanalysen und die Konfidenzintervalle im Vergleich zu Knapp-Hartung schmaler sind. Die Konfidenzintervalle der Edgington-Methode sind nicht notwendigerweise symmetrisch und dadurch kann dieses Verfahren besser auf Heterogenität der studienspezifischen Effektschätzungen reagieren als die üblichen Standard-Verfahren für Fixed-Effect-Metaanalysen. Zudem ist eine interessante Erweiterung dieses Verfahrens möglich, die explizit eine Heterogenität zwischen den Studien berücksichtigt.

Abdulmajeed et al. [27] diskutieren die Kritikpunkte an der **Freeman-Tukey-Double-Arcsine-Transformation** zur metanalytischen Zusammenfassung von Anteilen, die u. a. in Schwarzer et al. (RSM 2019; 10(3): 476-483, siehe Bericht "*Wesentliche neue EbM-relevante Methoden 2018/2019*“) geäußert wurden. Sie zeigen auf, dass diese durch eine geeignetere Rücktransformation behoben werden können bzw. keinen Einfluss in der praktischen Anwendung haben. In ihrer durchgeführten Simulationsstudie verhalten sich die Freeman-Tukey-Double-Arcsine-Transformation und die logit-Transformation sehr ähnlich, wobei die Freeman-Tukey-Double-Arcsine-Transformation schmalere Konfidenzintervalle und insgesamt stabilere gepoolte Anteilsschätzungen liefert.

Emprechtinger et al. [28] präsentieren die Ergebnisse einer randomisierten Studie zur Untersuchung der Nützlichkeit der Web-Application **metaHelper** (www.metaHelper.eu), mit der für Metaanalysen benötigte statistische Transformationen durchgeführt werden können. Es zeigte sich, dass metaHelper die Richtigkeit und Effizienz von statistischen Transformationen verbessert.

3.2 Reviews von Studien zur diagnostischen Güte

Nikoloulopoulos [29] schlägt ein Copula-basiertes Modell zur Durchführung einer gemeinsamen Metaanalyse für den **Vergleich von 2 diagnostischen Tests** an denselben Teilnehmern in einem gepaarten Design mit einem Goldstandard. Die Methode stellt eine Verallgemeinerung eines multinomialen generalisierten linearen gemischten Modells dar. Der Autor untersucht die Methode mithilfe von Simulationen und demonstriert die Anwendung anhand eines konkreten Datensatzes aus der Literatur zum Downsyndrom.

Zapf et al. [30] untersuchen drei frequentistische Ansätze für die Metaanalyse von **Receiver-Operating-Characteristic (ROC)-Kurven** für Studien zur diagnostischen Güte mit mehreren Schwellenwerten anhand einer Simulationsstudie. Hierbei handelte es sich um das Random-Effects-Modell von Steinhauser, Schumacher & Rücker (BMC-MRM 2016, 16: 97, siehe Bericht "*Wesentliche neue EbM-relevante Methoden 2015/2016*", das Time-to-Event Modell von Hoyer, Hirt & Kuss (RSM 2018; 9: 62-72; siehe Bericht "*Wesentliche neue EbM-relevante Methoden 2017/2018*") sowie das semiparametrische Modell von Frömke, Kirstein & Zapf (RSM 2022; 13: 612-621). Zapf et al. [30] berichten Ergebnisse aus 108 verschiedenen Simulationsszenarien, wobei alle Methoden je nach untersuchtem Szenario unterschiedlich gut abschnitten. Der Ansatz von Hoyer, Hirt & Kuss (RSM 2018; 9: 62-72; siehe Bericht "*Wesentliche neue EbM-relevante Methoden 2017/2018*") war in den meisten Szenarien bezüglich Verzerrung und Überdeckungswahrscheinlichkeit leicht überlegen und wird daher von den Autoren in der Praxis empfohlen. Im Fall einer Nicht-Konvergenz könnte man auf den Ansatz von Frömke, Kirstein & Zapf (RSM 2022; 13: 612-621) zurückgreifen.

3.3 Publikationsbias

Mohammadi [31] weist auf die Bedeutung von Publikationsbias in **Metaanalysen zur Prävalenz** hin. Dieses Thema wird laut dem Autor unterschätzt. Zur Identifizierung schlägt der Autor den Einsatz von Funnel-Plots vor. In Prävalenzstudien führt die Begrenzung der Prävalenzwerte zwischen 0 und 1 zu Problemen, da der Plot dadurch asymmetrisch erscheint. Um dennoch eine zuverlässige Bias-Diagnose zu ermöglichen, wird empfohlen, die Prävalenz zu transformieren, so dass man eine Datenverteilung über die gesamte Skala von $-\infty$ bis $+\infty$ erhält, die anschließend sinnvoll in einem Funnel-Plot darstellbar ist.

Der **inverse Publication Bias** wird von Xing et al. [32] näher beleuchtet. Inverser Publication Bias tritt auf, wenn ein ähnliches Sicherheitsprofil hinsichtlich unerwünschter Ereignisse (UEs) bevorzugt wird, und kann mithilfe verschiedener Methoden aufgedeckt werden. Dazu zählen die Suche nach grauer Literatur, visuelle Tools wie Funnel-Plots und quantitative Methoden wie Eggers´ Test und Peters´ Test, letzterer wird für die Anwendung bei seltenen Ereignissen bevorzugt. Anhand von 3 Fallbeispielen wird die Anwendung der Methoden illustriert.

3.4 Überlebenszeiten

Tierney et al. [33] beschreiben, wie Hazard Ratios (HRs) und Maße für deren Varianz aus publizierten (aggregierten) Time-to-event-Daten geschätzt werden können, wenn diese nicht explizit berichtet werden und keine Individualdaten vorliegen. Der **Leitfaden** basiert auf der ursprünglichen, sehr häufig zitierten Arbeit von Tierney et al. (Trials 2007; 8: 16). Aufbauend auf Anwendungserfahrung, Feedback aus der Cochrane-Community und einer Nutzerbefragung wird der Leitfaden aus dem Originalartikel erweitert, präzisiert und systematisch strukturiert. Der Leitfaden beschreibt, wie HRs auf Basis von Ereigniszahlen, Gruppengrößen,

p-Werten aus Log-Rank-Tests oder Cox-Modellen, Vergleichen von medianen Überlebenszeiten sowie aus Informationen aus Kaplan–Meier-Kurven mit zusätzlichen Angaben wie „Number at Risk“ geschätzt werden können. Ergänzt wird der Leitfaden durch ein überarbeitetes Excel-Tool, das die verschiedenen Datenszenarien automatisch erkennt, geeignete Berechnungen vornimmt und auf mögliche Einschränkungen hinweist.

3.5 Indirekte Vergleiche und Netzwerk-Metaanalysen

Gianola et al. [34] untersuchen in einer metaepidemiologischen Studie die Qualität publizierter Netzwerk-Metaanalysen aus dem Januar 2023 mithilfe von **PRISMA-NMA** und **AMSTAR-2**. Es wird erneut bestätigt, dass die verwendeten Methoden und die Berichterstattung publizierter Netzwerk-Metaanalysen unzureichend sind. Es wird darauf hingewiesen, dass zur Durchführung von Netzwerk-Metaanalysen ein großes Autorenteam erforderlich ist, das auch einen Statistiker enthält. Die Guidelines für Netzwerk-Metaanalysen sollten besser beachtet werden und die Dokumentation von Netzwerk-Metaanalysen sollte verbessert werden.

Eine wichtige Aufgabe in Netzwerk-Metaanalysen ist die **Untersuchung von Inkonsistenz**, d. h. Unterschieden zwischen direkter und indirekter Evidenz. Dazu gibt es zwei gängige Verfahren: das Design-by-Treatment-Interaction- und das Side-Splitting-Modell. Qin et al. [35] vergleichen diese beiden Modelle analytisch und per Simulationen. Sie zeigen, dass das Side-Splitting ein Spezialfall des Design-by-Treatment-Interaction-Modells ist, welches zusätzliche Annahmen trifft. Es wird empfohlen, grundsätzlich das Design-by-Treatment-Interaction-Modell zu verwenden. Das Side-Splitting-Modell könne ergänzend einbezogen werden, insbesondere bei wenigen Studien. Die Publikation gibt generell einen guten Überblick über verschiedene Netzwerkstrukturen und Modelle sowie einige praktische Handlungsanweisungen zur Untersuchung von Inkonsistenzen.

Phillippo et al. [36] untersuchen den Einfluss von Effektmodifikation und **Non-Collapsibility** in populationsadjustierten Methoden für indirekte Vergleiche. Die Autoren geben einen Überblick über bedingte und marginale Schätzfunktionen, veranschaulichen die jeweiligen Eigenschaften bei vorhandener Effektmodifikation und diskutieren die Auswirkungen auf die Entscheidungsfindung. An Beispielen zeigen sie, dass Effektmodifikation zu widersprüchlichen Ergebnissen zwischen bedingten und marginalen Schätzungen führen kann. Derzeit ist die multilevel Network Meta-Regression (ML-NMR) die einzige Methode, mit der sowohl bedingte als auch marginale Schätzungen in jeder Zielpopulation für die Entscheidung erstellt werden können. Die Autoren schließen mit praktischen Empfehlungen für die Entscheidungsfindung bei Vorliegen einer Effektmodifikation, basierend auf pragmatischen Vergleichen sowohl der bedingten als auch der marginalen Schätzungen in der Zielpopulation für die Entscheidung.

Bei dem Verfahren Matching-adjusted indirect Comparison (MAIC) werden Probanden aus einer Studie mit individuellen Teilnehmerdaten (IPD) mittels potenzieller Confounder neu gewichtet, um sie an die zusammenfassenden Statistiken der Kovariaten in einer anderen Studie mit aggregierten Daten (AgD) anzupassen, um einen Vergleich der Interventionen durchzuführen. Wenn jedoch Ungleichgewichte bei den Effektmodifikatoren mit unterschiedlichem Ausmaß der Modifikation zwischen den Behandlungen bestehen, kann es zu widersprüchlichen Schlussfolgerungen kommen, wenn der MAIC mit den IPD und AgD durchgeführt wird, die zwischen den Studien ausgetauscht wurden („**MAIC-Paradoxon**"). In ihrem Artikel verwenden Jiang et al. [37] ein anschauliches Beispiel, um dieses Paradoxon zu veranschaulichen und betonen, wie wichtig es ist, die Zielpopulation in HTA-Einreichungen klar zu definieren. Darüber hinaus empfehlen die Autoren, den HTA-Behörden anonymisierte IPD zur Verfügung zu stellen, um weitere indirekte Vergleiche zu ermöglichen, die die Gesamtpopulation, die sowohl durch IPD- als auch durch AgD-Studien repräsentiert wird, sowie andere relevante Zielpopulationen besser widerspiegeln.

Nourredine et al. [38] weisen darauf hin, dass bei indirekten Vergleichen in unverbundenen Netzwerken, in denen ein externer Vergleichsarm unter Verwendung routinemäßig erhobener Daten konstruiert wird, sich vor allem auch die Endpunkterhebung von der auf der Seite der Prüfintervention durchgeführten Studie unterscheiden kann. Daher müsse für die Schätzung eines unverzerrten indirekten Effekts zunächst diese „**Missklassifikation**" des Endpunkts adressiert werden. Das Ziel der Autoren ist die simulations¬basierte Quantifizierung einer Verzerrung durch das Ignorieren einer Missklassifikation bei einem binären Endpunkt im Rahmen eines indirekten Vergleichs auf Basis von 2 Studien im unverbundenen Netzwerk. Zudem schlagen sie eine Likelihood-basierte Methode zur Korrektur dieser Verzerrung vor. Mit Simulationen zeigen sie, dass die vorgeschlagene Modellierung in verschiedenen Szenarien zu einer Verbesserung hinsichtlich Coverage und Mean Squared Error (MSE) führt. Allerdings wird – wie üblich – davon ausgegangen, dass im Modell alle relevanten Prädiktoren und Effektmodifikatoren korrekt enthalten sind, wovon in der Praxis allerdings nicht auszugehen ist.

3.6 Bayessche Ansätze

Yao et al. [39] vergleichen in einer Simulationsstudie verschiedene **bayessche Metaanalyse-Modelle** untereinander, aber auch mit dem von Kuss (Stat Med 2015; 34: 1097-1116; siehe Bericht "*Wesentliche neue EbM-relevante Methoden 2014/2015*") vorgeschlagenen Beta-Binomial-Modell für Datensituationen mit seltenen Ereignissen und Doppelnull-Studien. Insgesamt schließen sie, dass keines der betrachteten Modelle den anderen grundsätzlich überlegen ist, jedoch weisen sie darauf hin, dass bei seltenen Ereignissen Metaanalysen häufig zu wenig Power haben und schlagen vor, dass bei einer Metaanalyse mit seltenen Ereignissen in der Ergebnisdarstellung auch eine post hoc Power-Berechnung angegeben werden sollte.

Insgesamt wird bei seltenen Ereignissen das Betabinomialmodell sowie das bayessche Random-Effects-Modell zur Anwendung empfohlen.

Yao et al. [40] untersuchen in einer weiteren Simulationsstudie die Eignung 7 verschiedener A-priori-Verteilungen für den Heterogenitätsparameter τ in **bayesschen Metaanalysen bei wenigen Studien** und seltenen Ereignissen. Die Simulationsszenarien decken Situationen im Bereich zwischen geringer (τ=0,05) bis extremer Heterogenität (τ = 1) ab. Die Autoren schlussfolgern, dass die Wahl der A-priori-Verteilung einen starken Einfluss auf die gepoolte Schätzung des Gesamteffekts der Metaanalyse hat, dass aber keine der betrachteten A-priori-Verteilungen über alle Simulationsszenarien betrachtet als die beste identifiziert werden konnte. Insgesamt werden für Situationen, in denen die Heterogenität nicht sehr hoch ist, die bereits in der Literatur empfohlenen Priors mit Lognormal- und Halbnormal-Verteilung empfohlen (siehe Berichte "*Wesentliche neue EbM-relevante Methoden 2020/2021“ bis "2023/2024*“).

4 Spezielle Themenbereiche

4.1 Estimands

Lanius et al. [41] geben 5 konkrete Empfehlungen dazu, das Estimand-Framework bei der Berichterstattung von Behandlungseffekten in klinischen Studien systematisch anzuwenden. Unter anderem betrifft dies die Dokumentation von Intercurrent Events, die Verknüpfung von Ergebnissen mit spezifischen Estimands und die Offenlegung zugrundeliegender Annahmen. Anhand von zwei Fallbeispielen zeigen sie, wie diese Empfehlungen praktisch umgesetzt werden können. Fierenz und Zapf [42] stellen zunächst die wichtigsten Aspekte des Estimand-Konzepts dar und präsentieren das Ergebnis einer Literaturrecherche zu Erweiterungen des Estimands-Konzepts. Es werden verschiedene modifizierte Strategien identifiziert und vorgestellt. Mütze et al. [43] schlagen in ihrem Viewpoint-Artikel 4 **Prinzipien zur Definition** eines Estimands für eine klinische Studie vor. Diese Prinzipien beschreiben sie als Schlussfolgerungen aus den Festlegungen in der ICH E9(R1) und möchten damit eine Diskussion der grundlegenden Eigenschaften von Estimands in klinischen Studien anregen.

In RCTs mit Endpunkten, die über die Zeit bis zum Ereignis operationalisiert sind, können interkurrente Ereignisse auf 2 Weisen auftreten: als semi-konkurrierende Ereignisse, die die Hazard-Funktion der interessierenden Ereignisse ändern oder als konkurrierende Ereignisse, die das Eintreten eines interessierenden Ereignisses verhindern. Auch wenn im ICH E9 (R1) Addendum 5 Strategien zum Umgang mit interkurrenten Ereignissen in RCTs vorgeschlagen wurden, sind diese nicht leicht anwendbar auf **Ereigniszeiten**, wenn auf eine kausale Interpretation abgezielt wird. Deng et al. [44] zeigen, wie in diesem Kontext kausale Interpretationen erfolgen können. Sie leiten mathematische Formeln der den 5 Strategien entsprechenden Estimands her und erläutern, welche Datenstruktur zur Identifizierung dieser Estimands erforderlich ist. Darüber hinaus führen die Autoren nichtparametrische Methoden zur Schätzung dieser Estimands, einschließlich asymptotischer Varianzschätzer und Hypothesentests, vor. Die Methoden werden an einem Datenbeispiel illustriert.

Lynggard et al. [45] diskutieren mögliche Umsetzungen des Estimand-Frameworks bei **Nichtunterlegenheitsstudien**. Sie stellen 7 Empfehlungen zusammen, die sich unter anderem mit der passenden Wahl des Estimands und der Nichtunterlegenheitsgrenze befassen. Die Autoren sprechen entgegen der EMA Guidance, gemäß der sowohl die Ergebnisse des Full-Analysis-Sets als auch des Per-Protocol-Sets (PPS) berücksichtigt werden sollen, eine Empfehlung gegen Per-Protocol-Analysen in diesem Zusammenhang aus.

4.2 Ereigniszeitanalysen

In einer systematischen Übersicht und einer Simulationsstudie untersuchen van der Ven et al. [46] 6 Methoden zur **Quantifizierung der Vollständigkeit des Follow-Up** in Kohortenstudien und klinischen Studien. Empfohlen werden die Simplified Person-Time-Methode und eine

modifizierte Version von Clark's Completeness Index (Xue et al., BMC-MRM 2017, 15: 155; siehe Bericht "*Wesentliche neue EbM-relevante Methoden 2017/2018*").

Bei der Planung einer klinischen Studie in der Onkologie wird üblicherweise von proportionalen Hazards und häufig sogar einer Exponentialverteilung für den Endpunkt Zeit bis zum Auftreten eines Ereignisses ausgegangen. Häufig wird neben dem Endpunkt Gesamtüberleben (OS) das progressionsfreie Überleben (PFS) als Endpunkt betrachtet. Erdmann, Beyersmann & Rufibach [47] verwenden ein **Survival-Multistate-Modell**, um diese beiden Endpunkte gemeinsam zu modellieren, und stellen fest, dass weder die Exponentialverteilung noch die Annahme proportionaler Hazards für beide Endpunkte gleichzeitig gelten können, was sich durch die Abhängigkeiten zwischen den beiden Endpunkten erklären lässt. Es wird gezeigt, wie die Planung klinischer Studien, insbesondere in der Onkologie, auf der Grundlage von Simulationen aus einem Multistate-Modell erfolgen kann. Ein R-Paket ist verfügbar.

Beyersmann, Schmoor & Schumacher [48] greifen die aktuelle Kritik am HR auf, dass das HR keine **kausale Interpretation** habe. Die Autoren beschreiben, dass Kontraste von Hazards sich für eine kausale Interpretation eignen können, aber eine solche Interpretation auf der Hazard-Skala zweifelhaft ist. Sie argumentieren, dass Hazard-Kontraste als Vergleich von Hazard-Funktionen und nicht als Vergleich von Hazards zu einem festen Zeitpunkt interpretiert werden sollten. Darüber hinaus empfehlen sie, die Hazard-Analysen routinemäßig in Wahrscheinlichkeiten zu übersetzen. Sie zeigen, dass das Problem der Konditionierung auf ein Ereignis nach der Randomisierung, das zu einem sogenannten „Collider-Bias" führt, aus funktionaler Sicht verschwindet. Sie veranschaulichen ihre Argumentation anhand von publizierten Beispielen aus der Nutzenbewertung. Der Artikel wird mit Verweis auf Fay & Lin [49], die zeigen, dass das HR zudem eine valide kausale Interpretation als Population-Level Estimand hat, unterstützend kommentiert [50].

Ereigniszeiten weisen bisweilen unterschiedliche Zensierungsmuster zwischen den Studienarmen auf. Dies kann ein Hinweis auf informative Zensierungen sein. Im Zusammenhang mit verzögert auftretenden Behandlungseffekten untersuchen Lin et al. [51] mithilfe einer Simulationsstudie in 12 unterschiedlichen Szenarien den Einfluss **informativer Zensierungen** auf die Verzerrung des HR. Dabei wird eine Zensierung von 15 % zu 12 Monaten in jedem der Studienarme angenommen. Im Falle einer positiven Korrelation zwischen Ereigniszeit und Zensierungszeit im Kontrollarm bei gleichzeitiger negativer Korrelation im Interventionsarm kommt es zu einer starken Überschätzung des HR. Haben Zensierungen in beiden Armen die gleiche Richtung und den gleichen Grad, so sind die Auswirkungen auf die Effektschätzung minimal.

Mehrotra & West [52] weisen darauf hin, dass Ereigniszeitanalysen ohne **Stratifizierung nach Risikofaktoren** zu verfälschten Werten des HRs führen. Die Autoren schlagen zur

Stratifizierung den 5-STAR-Ansatz vor, der anhand eines prä-spezifizierten Algorithmus die Patienten basierend auf Baseline-Kovariablen in geeignete Risikogruppen unterteilt, innerhalb derer ein Effektschätzer berechnet und anschließend über die Strata gemittelt wird.

Die **Restricted Mean Survival Time (RMST)** findet zunehmend Anwendung bei Ereigniszeitanalysen insbesondere, wenn die Proportional Hazards Annahme fraglich ist. Karrison, Hu & Dignam [53] schlagen neben der Differenz alternative Maßzahlen basierend auf der RMST als Effektmaß (RMST Ratio, Time Lost Ratio sowie die Average Survival Difference) vor. Die Autoren diskutieren Vor- und Nachteile der Maßzahlen im Vergleich mit dem klassischen Hazard Ratio anhand von zwei konkreten klinischen Datensätzen. Die Autoren kommen zu der Schlussfolgerung, dass bei plausibler Proportional-Hazard-Annahme das klassische Hazard Ratio Vorteile hat, bei nicht proportionalen Hazards die RMST-basierten Effektmaße eine gute Alternative darstellen.

Shi et al. [54] schlagen eine neue Gewichtung für das Prentice-Williams-Peterson (PWP)-Modell zur **Analyse rekurrenter Ereignisse** vor. Da im Ursprungsmodell zur Analyse des k-ten Ereignisses nur Patientinnen und Patienten herangezogen werden, die zuvor (k-1) Ereignisse erfahren haben, kann das zu Collider Bias führen, da sich die Populationen des Interventions- und Kontrollarms mit der Zeit immer mehr unterscheiden. Dieses Ungleichgewicht wird durch ein Gewichtungsverfahren ausgeglichen. In einer Simulationsstudie wird das neue Verfahren u. a. mit der Cox-Regression (unter Berücksichtigung des 1. Ereignisses), dem Andersen-Gill-Modell und dem Lin-Wie-Yang-Ying-Modell verglichen. Es besitzt unter allen Verfahren den niedrigsten Bias, hält das Signifikanzniveau von 5 % ein und besitzt die größte Power.

Shao, Ye und Zhang [55] schlagen einen **exakten Test** (auch für Äquivalenztestung) für das Cox-Modell mit wenigen Ereignissen vor. Durch Invertierung des exakten Tests wird auch ein exaktes Konfidenzintervall abgeleitet. Ein entsprechender Programmcode ist im R-Paket „ExactCox“ verfügbar.

Austin und Fine [56] erörtern die Anwendung der Gewichtung per inverser Behandlungswahrscheinlichkeit (IPTW) in Überlebenszeitanalysen mit **konkurrierenden Ereignissen** zur Schätzung von Behandlungseffekten in Beobachtungsstudien. Die Autoren untersuchen 3 Schätzer für zeitabhängige Risikodifferenzen (und NNTs) und relative Risiken: den gewichteten Aalen-Johansen-Schätzer, die Kombination der IPTW mit Gewichtung per inverser Zensierungswahrscheinlichkeit (IPTW-IPCW) und die Kombination der doppelt robusten, verstärkten IPTW mit IPCW (AIPTW-IPCW). In Simulationen unter der Verwendung der parametrischen Modellierung der Behandlungszuteilung per logistischer Regression lieferten alle 3 Methoden tendenziell unverzerrte Schätzungen, wobei IPTW-IPCW weniger präzise war. Der Vorteil des gewichteten Aalen-Johansen-Schätzers gegenüber AIPTW-IPCW ist, dass er leicht auf Gewichtungen angepasst werden kann, die nicht nur die Schätzung eines

ATE-Estimands zum Ziel haben (Gewichte für ATT, Matching Weights, Overlap-Weights, Entropie-Gewichte). Der Vorteil von AIPTW-IPCW ist dagegen die doppelte Robustheit.

4.3 Fehlende Werte

Der Artikel von Cro et al. [57] beschäftigt sich mit der Frage der Ersetzung fehlender Werte mit multipler Imputation und **Conditional Mean Imputation**. Die Autoren stellen die Vorteile der multiplen Imputation heraus, während sie von der Verwendung der Conditional Mean Imputation abraten. Wolbers et al. [58] nehmen hierauf Bezug und diskutieren die genannten Aspekte kritisch.

Wijesuriya et al. [59] haben eine Übersicht zur **multiplen Imputation für Längsschnittdaten** als Tutorium aufbereitet, worin mehrere Konfigurationen von Joint Modeling (JM) und Fully Conditional Specification (FCS) miteinander verglichen werden, wobei unter anderem zwischen Konfigurationen unterschieden wird, die nicht nur mit Korrelation zwischen den Beobachtungen einer Person, sondern auch mit Korrelation zwischen Personen in Clustern umgehen können. Die Autorengruppe verwendet Simulationen zur Demonstration und stellt in mehreren Tabellen die Charakteristika der Konfigurationen übersichtlich dar. Es wird auch auf die Notwendigkeit der Kompatibilität des Imputationsmodells mit dem späteren Analysemodell hingewiesen.

4.4 Subgruppenanalysen

Núñez & Belaunzarán-Zamundio [60] beschreiben potenzielle Verzerrungen bei Subgruppenauswertungen und Auswertungen sekundärer Endpunkte und illustrieren diese anhand publizierter Studienergebnisse. Prognostische Faktoren sollten nicht nur hinsichtlich ihrer Relevanz auf den primären Endpunkt hin erhoben werden, es müssen auch Faktoren, die für sekundäre Endpunkte oder für Subgruppen relevant sind, beachtet werden. In Subgruppenauswertungen ist es beispielsweise notwendig, Studienabbrüche für jede der Subgruppen separat zu berichten, da einzelne Subgruppenmerkmale die Wahrscheinlichkeit eines Studienabbruchs beeinflussen können. Die Autoren appellieren, für alle Analysen analoge Überlegungen wie für die Hauptanalysen zu treffen, um Verzerrungen zu vermeiden.

4.5 Cluster-randomisierte Studien

Marsden et al. [61] schlagen eine neue Klassifikation für Cluster-randomisierte Studien vor. Die Autoren unterscheiden sechs Typen der „Cluster-Mitgliedschaft“ (Closed Cohort, Non-Standard Closed Cohort, Cross-Sectional, New-Admission Continuous Recruitment, Open-Cohort Discrete-Recruitment und Open-Cohort Continuous-Recruitment) und definieren 6 Haupt- und 5 zusätzliche Merkmale zur Charakterisierung dieser Designs. Ziel ist es, dass Studien künftig klarer berichten, wie Individuen in Clustern ein- und austreten und wie deren Daten bzw. Endpunkte erhoben werden, um Verzerrungen zu minimieren und

Vergleichbarkeit zu ermöglichen. Weiterhin fordern die Autoren, dass CONSORT-Leitlinien erweitert werden, um diese Unterscheidungen besser abzubilden.

4.6 Studien zur diagnostischen Güte

Fierenz et al. [62] übertragen das **Estimand-Konzept** auf diagnostische Testgüte-Studien, bestehend aus den Attributen Population, zu identifizierender Zustand, Index-Test, Güte-Maß sowie Strategien zum Umgang mit sogenannten störenden Ereignissen (interfering Events). Störende Ereignisse treten vor oder während der Durchführung des diagnostischen Tests auf und können das Testergebnis beeinflussen oder zu einem fehlenden Testergebnis führen. Die Autoren stellen 6 Strategien zum Umgang mit störenden Ereignissen vor, die bei der Festlegung des Estimands gewählt werden können. Das Konzept wird an einem fiktiven Beispiel illustriert.

Die Receiver-Operating-Characteristic **(ROC)-Kurve** ist ein zentrales Werkzeug zur Bewertung der diagnostischen Genauigkeit und hilft Schwellenwerte für die Krankheitsklassifikation zu bestimmen. Ghosal [63] stellt verschiedene Modelle zur Schätzung von ROC-Kurven vor und erweitert bestehende Schätzverfahren durch die Integration von Kovariablen, wodurch eine Berechnung von Kovariablen-spezifischen Schwellenwerten möglich wird. Der Autor führt Simulationen hierzu durch und wendet die Methoden auf ein Datenbeispiel mit Biomarkern zur Diagnose von Alzheimer an. Hierbei zeigt sich, dass sowohl die Wahl der ROC-Schätzmethode als auch die Berücksichtigung von Kovariablen zu deutlichen Unterschieden in Sensitivität, Spezifität und optimalen Schwellenwerten führen.

In dem Artikel von Negeri [64] wird ein **bivariates Finite-Mixture-Random-Effects-Modell** vorgestellt, welches sowohl die Heterogenität innerhalb als auch zwischen den Studien berücksichtigt. Dieses Modell weist jeder Studie eine Wahrscheinlichkeit zu, ein Ausreißer zu sein und erlaubt so die Bewertung des Einflusses von Ausreißern auf die gepoolte Sensitivität, Spezifität und die Heterogenität zwischen den Studien. Anhand von Simulationen und einem Datenbeispiel wird dieses Modell untersucht. Hierbei zeigt sich, dass das Modell bei Vorhandensein von Ausreißern robuste und präzise Schätzungen liefert und ähnliche Ergebnisse wie die Standardmodelle liefert, wenn keine Ausreißer vorliegen.

Sun & Zhou [65] behandeln die Schätzung der **Genauigkeit diagnostischer Tests**, wenn kein perfekter Referenzstandard verfügbar ist. Die Autoren geben nicht nur erklärende Hilfestellungen, sondern stellen auch R-Code zur Verfügung. Dabei unterscheiden sie nach dem Skalenniveau des Testergebnisses, dem Vorliegen des Einflusses von Kovariablen, der Anzahl der durchgeführten Tests und ob bei Mehrfachdurchführung zwischen den Tests bedingte Unabhängigkeit besteht. Dabei konzentrieren sie sich auf existierende statistische Methoden, die den wahren Krankheitsstatus als latente Variable modellieren, nachdem sie die Vorteile davon gegenüber „discrepant Analysis“ und „composite Reference Standards“ in

der Einleitung herausgestellt haben. Es wird betont, dass die Wahl der Methode von der Datenstruktur abhängt und vereinfachende Annahmen zu Verzerrungen führen können. Eine Grafik gibt eine Übersicht über alle vorkommenden Methoden mit Referenzangaben.

4.7 Treatment Switching

Hu et al. [66] untersuchen Treatment Switching im Zusammenhang mit einer Teilgruppe als geheilt geltender Personen, da diese zwar weiterhin versterben können, aber nicht als Behandlungswechsler infrage kommen. Sie entwickeln ein Multistate-Transition-Modell in dem die Stadien Heilung, Krankheitsprogress, Treatment Switching und Tod berücksichtigt werden. Die Heilungswahrscheinlichkeit für alle Patientinnen und Patienten sowie das Risiko zu versterben in der Gruppe der Geheilten wird separat modelliert. In der Gruppe der Nicht-Geheilten werden die Risiken für Krankheitsprogression, Treatment Switching und Tod modelliert. In einer Simulationsstudie zeigt sich, dass das Verfahren einen geringeren Bias bezüglich des Behandlungseffekts auf das Gesamtüberleben aufweist als Auswertungen mittels Intention to treat (ITT), per Protocol (PP) und Rank-Preserving-Structural-Failure-Time-Modellen (RPSFTM).

4.8 Kausale Inferenz

Keele & Greeve [67] geben einen Überblick über **Adjustierungsmethoden in nicht randomisierten Studien**. Die wichtigsten Annahmen der verschiedenen parametrischen, semi- und nichtparametrischen Ansätze werden erklärt und Vor- und Nachteile herausgestellt. Es werden eine Reihe von praktischen Empfehlungen abgeleitet und R-Code zur Verfügung gestellt.

Eine zentrale Anforderung bei der Anwendung von **Propensity-Score (PS)**-Verfahren ist die Annahme der Positivität. Zur Sicherstellung dieser Annahme existieren für die Schätzung des durchschnittlichen Behandlungseffekts (Average Treatment Effect, ATE) – neben Trimming- und Truncation-Methoden – eine Reihe alternativer Ansätze. Zur Schätzung des durchschnittlichen Behandlungseffekts unter den Behandelten (Average Treatment Effect on the Treated, ATT) schlagen Liu et al. [68] einen neuen Ansatz mit Overlap-Gewichten (Overlap Weighted Average Treatment Effect on the Treated, OWATT) vor. Das neue Verfahren wird im Rahmen einer Simulationsstudie mit bestehenden Alternativen wie Trimming und Truncation verglichen und zeigt insgesamt die höchste Effizienz.

Shang, Shiu & Kong [69] schlagen eine **robuste PS-Schätzung** basierend auf einer Kalibrierung der Verlustfunktion durch Einführung eines Penalty-Faktors für Kovariablen-Imbalanzen vor. Die Autoren untersuchen den Ansatz in Verbindung mit klassischen logistischen Regressionsmodellen sowie neuronalen Netzwerken im Rahmen einer Simulationsstudie. Bei Fehlspezifikationen des PS-Modells liefern Modelle basierend auf neuronalen Netzwerken die stabilsten Ergebnisse.

Chen [70] betrachtet die Schätzung des ATE, wenn die binäre Behandlungsvariable fehlerhaft erhoben wurde und die Beziehung zwischen Behandlung und Confoundern nicht linear ist. Es wird eine Methode entwickelt, mit der fehlerhafte Behandlungserhebungen adressiert werden können. Nach der Korrektur der Behandlungen wird der Propensity Score unter Verwendung der **Random-Forest-Methode**, mit der auch eine nicht lineare Beziehung zu Confoundern berücksichtigt werden kann, geschätzt und ein ATE-Schätzer bereitgestellt. Neben der Darstellung asymptotischer Eigenschaften des Schätzers werden numerische Ergebnisse präsentiert, um zu zeigen, wie wichtig die beschriebene Korrektur ist.

Yucel Karakaya & Unal [71] schlagen eine neue Methode zur Bewertung der **Balanciertheit** in der Propensity-Score-Analyse nach multipler Imputation vor. Es wurde eine Simulationsstudie durchgeführt, um die Leistungsfähigkeit der Methoden zur Bewertung der Balanciertheit zu bewerten und mit bestehenden Methoden zu vergleichen. Die simulierten Szenarien variierten hinsichtlich des Vorhandenseins fehlender Daten in der Kontroll- oder Behandlungs- und Kontrollgruppe sowie hinsichtlich des Imputationsmodells mit / ohne den Endpunkt. Die Autorengruppe empfiehlt die Verwendung der neuen Methode, da diese vergleichbar oder besser abschneidet als die bisherigen Verfahren, aber keine komplexen Berechnungen erfordert.

Yucel Karakaya & Unal [72] schlagen außerdem eine neue Methode zur **Ersetzung von fehlenden Werten** bei relevanten Kovariablen im Rahmen von PS-Analysen vor. Das Verfahren basiert auf der Verwendung einer Indikator-Variable für fehlende Werte in Kombination mit multipler Imputation. Die Indikator-Variable kann im PS-Modell, im Outcome-Modell oder in beiden verwendet werden. Die Autoren führen eine Simulationsstudie dieser Varianten unter unterschiedlichen Szenarien durch. Die Autoren empfehlen als Standard die Variante mit Verwendung einer Indikator-Variable sowohl im PS- als auch Outcome-Modell.

Neben Methoden basierend auf Propensity Scores gibt es weitere statistische Verfahren, die speziell zur Schätzung eines kausalen Effektes mithilfe von nicht randomisierten Studien entwickelt wurden. Dazu zählt die **Targeted Maximum Likelihood Estimation (TMLE)**, mit deren Performance (Bias und Coverage) sich Smith et al. [73] beschäftigen, und zwar insbesondere in bestimmten problematischen Situationen wie kleine Stichproben. Sie zeigen durch eine Simulationsstudie, dass eine kreuzvalidierte TMLE eine bessere Performance erreicht als eine nicht kreuzvalidierte TMLE, insbesondere im Fall kleiner Stichproben.

Tang et al. [74] schlagen eine Methode für **Sensitivitätsanalysen** für nicht gemessenes Confounding in Beobachtungsstudien vor, die unter minimalen Annahmen (Positivität und Austauschbarkeit) Worst-Case-Grenzen für den durchschnittlichen Behandlungseffekt (ATE) liefert. Auch wenn die sich ergebenden Grenzen weit auseinanderliegen, sind sie doch ein guter Startpunkt zur Beurteilung potenzieller Auswirkung von restlichem Confounding, die man mit nur minimalen Annahmen erhält.

4.9 Adaptive Designs

In einer zweiteiligen Serie beschreiben Robertson et al. [75] zunächst, welche Möglichkeiten es gibt, Konfidenzintervalle zu berechnen, um im Anschluss darzulegen, welche dieser Berechnungen im Fall von adaptiven Designs geeignet oder ungeeignet sind. Im zweiten Teil [76] vergleichen sie anhand eines Beispiels die verschiedenen, im ersten Teil beschriebenen Verfahren und beschreiben in Form eines Guides, wie bei Studien mit adaptivem Design Konfidenzintervalle berechnet werden sollten und was bei der Studienplanung zusätzlich zu beachten ist.

4.10 Künstliche Intelligenz (KI)

Eine interdisziplinäre und internationale Expertengruppe, das FUTURE-AI Consortium, hat eine **konsensbasierte Richtlinie** entwickelt, anhand derer die Vertrauenswürdigkeit und ethische Bedenken bezüglich KI-Anwendungen im Gesundheitsbereich beurteilt werden können [77]. Die Richtlinie basiert auf den Prinzipien Fairness, Universalität, Nachvollziehbarkeit (Traceability), Anwendbarkeit (Usability), Robustheit und Erklärbarkeit. Zu jedem Prinzip gibt es eine Reihe von Handlungsempfehlungen, die sich auf den gesamten Lebenszyklus der KI-Anwendung (Planung, Entwicklung, Evaluation und Verwendung) beziehen.

In einem Scoping Review untersuchen Lieberum et al. [78] die Verwendung von Large Language Models (LLMs) zur **Durchführung systematischer Übersichten**. LLMs werden aktuell vor allem in den folgenden 3 Bereichen eingesetzt: Literaturrecherche, Auswahl relevanter Artikel und Datenextraktion. Die Qualität der Ergebnisse ist allerdings gemischt und häufig fehlt eine Validierung. Die Autoren schlussfolgern, dass die Bedeutung von LLMs steigt, diese aber aktuell noch nicht bedenkenlos verwendet werden sollten.

Eine wichtige, aber komplexe Aufgabe bei der Durchführung systematischer Reviews ist die Bewertung des Risikos für Bias (RoB) der einbezogenen Studien. Daher war es das Ziel der Studie von Eisele-Metzger et al. [79], das LLM Claude 2 für die **RoB-Bewertung** von 100 randomisierten kontrollierten Studien unter Verwendung des überarbeiteten Cochrane-Tools zur Bewertung des Risikos für Bias („RoB 2"; mit Bewertungen für fünf spezifische Domänen und einer Gesamtbewertung) zu testen. Die Autoren bewerteten die Übereinstimmung der RoB-Beurteilungen durch Claude 2 mit den in Cochrane-Reviews veröffentlichten menschlichen Beurteilungen. Die beobachtete Übereinstimmung zwischen Claude 2 und den Bewertungen der Cochrane-Autoren reichte von 41 % für die Gesamtbeurteilung bis zu 71 % für Bereich 4 („Ergebnisbewertung"). Insgesamt wurde eine mäßige Übereinstimmung festgestellt (Cohen's κ: 0,22 [95 %-KI: 0,06 – 0,38]). Sensitivitätsanalysen unter Verwendung alternativer Prompting-Techniken oder der neueren Version Claude 3 führten zu keinen wesentlichen Änderungen. Derzeit können RoB-Bewertungen durch Claude 2 nicht die Bewertungen durch Menschen ersetzen.

5 Sonstiges

Gutzeit et al. [80] beschreiben eine Methode für **Volume-Outcome-Analysen**. Das vorgeschlagene Modell ist ein GAMM (generalised additive mixed Model), welches die relevanten patientenspezifischen Risikofaktoren, die Leistungsmenge als Spline, gegebenenfalls weitere Merkmale des Leistungserbringers und einen Zufallseffekt für den Leistungserbringer enthält. Der Zufallseffekt dient dazu, Clustereffekte auf Ebene des Leistungserbringers zu berücksichtigen. Die Modellierung der Leistungsmenge als Spline ermöglicht, auch einen nicht linearen Effekt der Leistungsmenge auf das Behandlungsergebnis zu identifizieren, beispielsweise eine U-Form oder Hockey-Stick-Form. Von einer Kategorisierung der Leistungsmenge wird explizit abgeraten. Das Modell wird in Simulationsstudien evaluiert, wobei auch die Simulation der Daten ausführlich beschrieben wird. Zusätzlich wird das Modell auf reale Beispieldaten angewendet.

Carlin & Moreno-Betancur [81] kritisieren den routinehaften **Einsatz von Regressionsmodellen** ohne klare Zieldefinition. Sie fordern eine Neuausrichtung der Statistik-Praxis und -Lehre auf den Zweck der Analyse: deskriptiv, prädiktiv oder kausal. Die 6 in Statistics in Medicine veröffentlichten Kommentare [82-87] stimmen im Grundsatz zu, dass Regressionsmodelle zu oft unreflektiert eingesetzt werden und zu wenig Bewusstsein für Grenzen und Annahmen von Modellen vorhanden ist; Statistik-Lehre soll stärker zweckorientiert und kontextbezogen werden. Diskussionslinien beschäftigen sich mit der Vermeidung von Überinterpretation angewandter Modelle [83], betonen die Wichtigkeit von initialer Datenbeschreibung, Reporting Guidelines und weiterer vergleichender Methodenforschung [85], oder fordern tiefere Auseinandersetzung mit Modellannahmen und Unsicherheiten in der Lehre und bei der Anwendung [84]. Carlin & Moreno-Betancur [88] nehmen die diskutierten Aspekte in ihrer Antwort wieder auf.

6 Interessantes

Zum Abschluss vielleicht ganz interessant:

Showell et al. [89] haben in einer systematischen Übersichtsarbeit untersucht, wie hoch der Anteil klinischer Studien mit publizierten Ergebnissen ist, wie lange es dauert, bis die Ergebnisse publiziert werden und welche Faktoren dies beeinflussen. Nur etwas über die Hälfte der Studien wurden publiziert, im Median 2,1 Jahre nach Ende der klinischen Studie. Studien mit positiven Ergebnissen wurden nicht nur mit höherer Wahrscheinlichkeit, sondern auch schneller publiziert. Dasselbe trifft auf Studien mit höherer Fallzahl und Studien, die nicht durch die Industrie finanziert wurden, zu. Multizentrische Studien wurden häufiger mit höherer Wahrscheinlichkeit publiziert als monozentrische Studien, die Zeit bis zur Publikation unterschied sich jedoch nicht signifikant.

7 Literatur

1. Hopewell S, Chan AW, Collins GS et al. CONSORT 2025 statement: Updated guideline for reporting randomised trials. BMJ 2025; 389: e081123. https://doi.org/10.1136/bmj-2024-081123.

2. Chan AW, Boutron I, Hopewell S et al. SPIRIT 2025 statement: Updated guideline for protocols of randomised trials. BMJ 2025; 389: e081477. https://doi.org/10.1136/bmj-2024-081477.

3. Hopewell S, Chan AW, Collins GS et al. CONSORT 2025 explanation and elaboration: Updated guideline for reporting randomised trials. BMJ 2025; 389: e081124. https://doi.org/10.1136/bmj-2024-081124.

4. Hrobjartsson A, Boutron I, Hopewell S et al. SPIRIT 2025 explanation and elaboration: Updated guideline for protocols of randomised trials. BMJ 2025; 389: e081660. https://doi.org/10.1136/bmj-2024-081660.

5. McKenzie JE, Taljaard M, Hemming K et al. Reporting of cluster randomised crossover trials: Extension of the CONSORT 2010 statement with explanation and elaboration. BMJ 2025; 388: e080472. https://doi.org/10.1136/bmj-2024-080472.

6. Kahan BC, Juszczak E, Beller E et al. Guidance for protocol content and reporting of factorial randomised trials: Explanation and elaboration of the CONSORT 2010 and SPIRIT 2013 extensions. BMJ 2025; 388: e080785. https://doi.org/10.1136/bmj-2024-080785.

7. Akl EA, Khabsa J, Iannizzi C et al. Extension of the PRISMA 2020 statement for living systematic reviews (PRISMA-LSR): Checklist and explanation. BMJ 2024; 387: e079183. https://doi.org/10.1136/bmj-2024-079183.

8. Cashin AG, Hansford HJ, Hernan MA et al. Transparent reporting of observational studies emulating a target trial: The TARGET Statement. BMJ 2025; 390: e087179. https://doi.org/10.1136/bmj-2025-087179.

9. Thomassen D, Roychoudhury S, Amdal CD et al. The role of the estimand framework in the analysis of patient-reported outcomes in single-arm trials: a case study in oncology. BMC Med Res Methodol 2024; 24(1): 290. https://doi.org/10.1186/s12874-024-02408-x.

10. Thomassen D, Roychoudhury S, Amdal CD et al. Handling missing values in patient-reported outcome data in the presence of intercurrent events. BMC Med Res Methodol 2025; 25(1): 56. https://doi.org/10.1186/s12874-025-02510-8.

11. Whiting P, Wolff R, Savovic J et al. Introducing the LATITUDES network: A library of assessment tools and training to improve transparency, utility and dissemination in evidence synthesis. J Clin Epidemiol 2024; 174: 111486. https://doi.org/10.1016/j.jclinepi.2024.111486.

12. Klugar M, Lotfi T, Darzi AJ et al. GRADE guidance 39: Using GRADE-ADOLOPMENT to adopt, adapt or create contextualized recommendations from source guidelines and evidence syntheses. J Clin Epidemiol 2024; 174: 111494. https://doi.org/10.1016/j.jclinepi.2024.111494.

13. Murad MH, Morgan RL, Falck-Ytter Y et al. Simultaneous evaluation of the imprecision and inconsistency domains of GRADE can be performed using prediction intervals. J Clin Epidemiol 2024; 175: 111543. https://doi.org/10.1016/j.jclinepi.2024.111543.

14. Guyatt G, Hultcrantz M, Agoritsas T et al. Why Core GRADE is needed: Introduction to a new series in The BMJ. BMJ 2025; 389: e081902. https://doi.org/10.1136/bmj-2024-081902.

15. Guyatt G, Agoritsas T, Brignardello-Petersen R et al. Core GRADE 1: Overview of the Core GRADE approach. BMJ 2025; 389: e081903. https://doi.org/10.1136/bmj-2024-081903.

16. Guyatt G, Zeng L, Brignardello-Petersen R et al. Core GRADE 2: Choosing the target of certainty rating and assessing imprecision. BMJ 2025; 389: e081904. https://doi.org/10.1136/bmj-2024-081904.

17. Guyatt G, Schandelmaier S, Brignardello-Petersen R et al. Core GRADE 3: Rating certainty of evidence-assessing inconsistency. BMJ 2025; 389: e081905. https://doi.org/10.1136/bmj-2024-081905.

18. Guyatt G, Iorio A, De Beer H et al. Core GRADE 5: Rating certainty of evidence-assessing indirectness. BMJ 2025; 389: e083865. https://doi.org/10.1136/bmj-2024-083865.

19. Guyatt G, Wang Y, Eachempati P et al. Core GRADE 4: Rating certainty of evidence-risk of bias, publication bias, and reasons for rating up certainty. BMJ 2025; 389: e083864. https://doi.org/10.1136/bmj-2024-083864.

20. Guyatt G, Yao L, Murad MH et al. Core GRADE 6: Presenting the evidence in summary of findings tables. BMJ 2025; 389: e083866. https://doi.org/10.1136/bmj-2024-083866.

21. Guyatt G, Vandvik PO, Iorio A et al. Core GRADE 7: Principles for moving from evidence to recommendations and decisions. BMJ 2025; 389: e083867. https://doi.org/10.1136/bmj-2024-083867.

22. Lunny C, Higgins JPT, White IR et al. Risk of Bias in Network Meta-Analysis (RoB NMA) tool. BMJ 2025; 388: e079839. https://doi.org/10.1136/bmj-2024-079839.

23. Moons KGM, Damen JAA, Kaul T et al. PROBAST+AI: An updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods. BMJ 2025; 388: e082505. https://doi.org/10.1136/bmj-2024-082505.

24. Wang Y, Keitz S, Briel M et al. Development of ROBUST-RCT: Risk Of Bias instrument for Use in SysTematic reviews-for Randomised Controlled Trials. BMJ 2025; 388: e081199. https://doi.org/10.1136/bmj-2024-081199.

25. Viana J, Machado V, Proenca L et al. Comparative assessment of Cochrane's ROB and ROB2 in dentistry trials: A meta-research study. Syst Rev 2025; 14(1): 154. https://doi.org/10.1186/s13643-025-02901-4.

26. Held L, Hofmann F, Pawel S. A comparison of combined p-value functions for meta-analysis. Res Synth Methods 2025; 16(5): 758-785. 758. https://doi.org/10.1017/rsm.2025.26.

27. Abdulmajeed J, Chivese T, Doi SAR. Overcoming challenges in prevalence meta-analysis: The case for the Freeman-Tukey transform. BMC Med Res Methodol 2025; 25(1): 89. https://doi.org/10.1186/s12874-025-02527-z.

28. Emprechtinger R, Schwarzer G, Schreder G et al. Assessing the effectiveness of metaHelper: a randomized controlled trial of a web application for statistical transformations in meta-analysis. J Clin Epidemiol 2025; 179: 111640. https://doi.org/10.1016/j.jclinepi.2024.111640.

29. Nikoloulopoulos AK. Joint meta-analysis of two diagnostic tests accounting for within and between studies dependence. Stat Methods Med Res 2024; 33(10): 1800-1817. https://doi.org/10.1177/09622802241269645.

30. Zapf A, Frömke C, Hardt J et al. Meta-analysis of diagnostic accuracy studies with multiple thresholds: Comparison of approaches in a simulation study. Biom J 2024; 66(7): e202300101. https://doi.org/10.1002/bimj.202300101.

31. Mohammadi M. Publication bias in prevalence studies should not be ignored. Syst Rev 2025; 14(1): 85. https://doi.org/10.1186/s13643-025-02845-9.

32. Xing X, Xu C, Al Amer FM et al. Methods for assessing inverse publication bias of adverse events. Contemp Clin Trials 2024; 145: 107646. https://doi.org/10.1016/j.cct.2024.107646.

33. Tierney JF, Burdett S, Fisher DJ. Practical methods for incorporating summary time-to-event data into meta-analysis: Updated guidance. Syst Rev 2025; 14(1): 84. https://doi.org/10.1186/s13643-025-02752-z.

34. Gianola S, Guida S, Ravot G et al. Gaps in completeness of reporting and methodological quality: A metaresearch study of 139 network meta-analyses published in January 2023 using PRISMA-NMA and AMSTAR-2. J Clin Epidemiol 2025; 183: 111783. https://doi.org/10.1016/j.jclinepi.2025.111783.

35. Qin L, Zhao S, Guo W et al. A comparison of two models for detecting inconsistency in network meta-analysis. Res Synth Methods 2024; 15(6): 851-871. https://doi.org/10.1002/jrsm.1734.

36. Phillippo DM, Remiro-Azócar A, Heath A et al. Effect modification and non-collapsibility together may lead to conflicting treatment decisions: A review of marginal and conditional estimands and recommendations for decision-making. Res Synth Methods 2025; 16(2): 323-349. 323. https://doi.org/10.1017/rsm.2025.2.

37. Jiang Z, Liu J, Alemayehu D et al. A critical assessment of matching-adjusted indirect comparisons in relation to target populations. Res Synth Methods 2025; 16(3): 569-574. 569. https://doi.org/10.1017/rsm.2025.10.

38. Nourredine M, Gavoille A, Lepage C et al. Accounting for misclassification of binary outcomes in external control arm studies for unanchored indirect comparisons: Simulations and applied example. Stat Med 2025; 44(20-22): e70236. https://doi.org/10.1002/sim.70236.

39. Yao M, Jia Y, Mei F et al. Comparing various Bayesian random-effects models for pooling randomized controlled trials with rare events. Pharm Stat 2024; 23(6): 837-853. https://doi.org/10.1002/pst.2392.

40. Yao M, Mei F, Zou K et al. Comparison of prior distributions for the heterogeneity parameter in a rare events meta-analysis of a few studies. Pharm Stat 2025; 24(2): e2448. https://doi.org/10.1002/pst.2448.

41. Lanius V, Glocker B, Losch C et al. Realizing the benefits of the estimand framework when reporting and communicating clinical trial results-some recommendations. Trials 2025; 26(1): 241. https://doi.org/10.1186/s13063-025-08915-6.

42. Fierenz A, Zapf A. Current developments of the estimand concept. Pharm Stat 2024; 23(6): 864-869. https://doi.org/10.1002/pst.2395.

43. Mütze T, Bell J, Englert S et al. Principles for defining estimands in clinical trials – A proposal. Pharm Stat 2025; 24(1): e2432. https://doi.org/10.1002/pst.2432.

44. Deng Y, Han S, Zhou XH. Inference for cumulative incidences and treatment rffects in randomized controlled trials with time-to-event outcomes under ICH E9 (R1). Stat Med 2025; 44(10-12): e70091. https://doi.org/10.1002/sim.70091.

45. Lynggaard H, Keene ON, Mütze T et al. Applying the estimand framework to non-inferiority trials. Pharm Stat 2024; 23(6): 1156-1165. https://doi.org/10.1002/pst.2433.

46. van der Ven C, Ikram MA, van Rooij FJA et al. Calculating follow-up completeness: A comparison of multiple methods under different simulated scenarios and a use case. J Clin Epidemiol 2025; 182: 111757. https://doi.org/10.1016/j.jclinepi.2025.111757.

47. Erdmann A, Beyersmann J, Rufibach K. Oncology clinical trial design planning based on a multistate model that jointly models progression-free and overall survival endpoints. Biom J 2025; 67(1): e70017. https://doi.org/10.1002/bimj.70017.

48. Beyersmann J, Schmoor C, Schumacher M. Hazards constitute key quantities for analyzing, interpreting and understanding time-to-event data. Biom J 2025; 67(3): e70057. https://doi.org/10.1002/bimj.70057.

49. Fay MP, Li F. Causal interpretation of the hazard ratio in randomized clinical trials. Clin Trials 2024; 21(5): 623-635. https://doi.org/10.1177/17407745241243308.

50. Bender R, Beckmann L. Hazards, causality, and practical relevance of collider effects – Comment on Beyersmann et al. "Hazards constitute key quantities for analyzing, interpreting and understanding time-to-event data". Biom J 2025; 67(5): e70071. https://doi.org/10.1002/bimj.70071.

51. Lin J, Zhao Y, Chen XG et al. Impact of informative censoring on estimation and testing in randomized trials with delayed treatment effects. Contemp Clin Trials 2025; 152: 107860. https://doi.org/10.1016/j.cct.2025.107860.

52. Mehrotra DV, West RM. Is inadequate risk stratification diluting hazard ratio estimates in randomized clinical trials? Clin Trials 2024; 21(5): 571-575. https://doi.org/10.1177/17407745231222448.

53. Karrison T, Hu C, Dignam J. Scaling and interpreting treatment effects in clinical trials using restricted mean survival time. Clin Trials 2025; 22(1): 3-10. https://doi.org/10.1177/17407745241254995.

54. Shi C, Wei JW, Zhan ZS et al. A new method for dealing with collider bias in the PWP model for recurrent events in randomized controlled trials. BMC Med Res Methodol 2025; 25(1): 142. https://doi.org/10.1186/s12874-025-02596-0.

55. Shao Y, Ye Z, Zhang Z. Exact test and exact confidence interval for the Cox model. Stat Med 2024; 43(23): 4499-4518. 4499. https://doi.org/10.1002/sim.10189.

56. Austin PC, Fine JP. Inverse probability of treatment weighting using the propensity score with competing risks in survival analysis. Stat Med 2025; 44(5): e70009. https://doi.org/10.1002/sim.70009.

57. Cro S, Morris TP, Roger JH et al. Comments on 'standard and reference-based conditional mean imputation': Regulators and trial statisticians be aware! Pharm Stat 2024; 23(5): 598-603. https://doi.org/10.1002/pst.2373.

58. Wolbers M, Noci A, Delmar P et al. Rejoinder to the letter: "Standard and reference-based conditional mean imputation: Regulators and trial statisticians be aware!". Pharm Stat 2024; 23(5): 604-610. https://doi.org/10.1002/pst.2374.

59. Wijesuriya R, Moreno-Betancur M, Carlin JB et al. Multiple imputation for longitudinal data: A tutorial. Stat Med 2025; 44(3-4): e10274. https://doi.org/10.1002/sim.10274.

60. Nunez I, Belaunzaran-Zamudio PF. Preventable sources of bias in subgroup analyses and secondary outcomes of randomized trials. Contemp Clin Trials 2024; 145: 107641. https://doi.org/10.1016/j.cct.2024.107641.

61. Marsden LE, Surr CA, Griffiths AW et al. Different types of cluster membership in parallel-group cluster-randomised trials, where the clusters are institutions: A classification system to aid identification, with six proposed designs. Trials 2025; 26(1): 380. https://doi.org/10.1186/s13063-025-09066-4.

62. Fierenz A, Akacha M, Benda N et al. The estimand framework in diagnostic accuracy studies. Stat Med 2025; 44(20-22): e70248. https://doi.org/10.1002/sim.70248.

63. Ghosal S. Impact of methodological assumptions and covariates on the cutoff estimation in ROC analysis. Biom J 2025; 67(3). https://doi.org/10.1002/bimj.70053.

64. Negeri ZF. A bivariate finite mixture random effects model for identifying and accommodating outliers in diagnostic test accuracy meta-analyses. Biom J 2025; 67(3). https://doi.org/10.1002/bimj.70062.

65. Sun A, Zhou XH. Estimation of diagnostic test accuracy without gold standards. Stat Med 2025; 44(3-4): e10315. https://doi.org/10.1002/sim.10315.

66. Hu H, Wang L, Wu K et al. A multistate transition model for survival estimation in randomized trials with treatment switching and a cured subgroup. BMC Med Res Methodol 2025; 25(1): 196. https://doi.org/10.1186/s12874-025-02623-0.

67. Keele L, Grieve R. So many choices: A guide to selecting among methods to adjust for observed confounders. Stat Med 2025; 44(5): e10336. https://doi.org/10.1002/sim.10336.

68. Liu Y, Li H, Zhou Y et al. Average treatment effect on the treated, under lack of positivity. Stat Methods Med Res 2024; 33(10): 1689-1717. https://doi.org/10.1177/09622802241269646.

69. Shang Y, Chiu YH, Kong L. Robust propensity score estimation via loss function calibration. Stat Methods Med Res 2025; 34(3): 457-472. https://doi.org/10.1177/09622802241308709.

70. Chen LP. Nonparametric estimation for propensity scores with misclassified treatments. Stat Med 2025; 44(1-2): e10306. https://doi.org/10.1002/sim.10306.

71. Yucel Karakaya SP, Unal I. Balance diagnostics in propensity score analysis following multiple imputation: A new method. Pharm Stat 2024; 23(5): 763-777. https://doi.org/10.1002/pst.2389.

72. Yucel Karakaya SP, Unal I. Incorporation of missing indicator with multiple imputation in propensity score analysis with partially observed covariates: A simulation study. Stat Methods Med Res 2025; 34(7): 1293-1302. https://doi.org/10.1177/09622802251338365.

73. Smith MJ, Phillips RV, Maringe C et al. Performance of cross-validated targeted maximum likelihood estimation. Stat Med 2025; 44(15-17): e70185. https://doi.org/10.1002/sim.70185.

74. Tang C, Zhou Y, Huang A et al. A simple sensitivity analysis method for unmeasured confounders via linear programming with estimating equation constraints. Stat Med 2025; 44(3-4): e10288. https://doi.org/10.1002/sim.10288.

75. Robertson DS, Burnett T, Choodari-Oskooei B et al. Confidence intervals for adaptive trial designs I: A methodological review. Stat Med 2025; 44(18-19): e70174. https://doi.org/10.1002/sim.70174.

76. Robertson DS, Burnett T, Choodari-Oskooei B et al. Confidence intervals for adaptive trial designs II: Case study and practical guidance. Stat Med 2025; 44(18-19): e70202. https://doi.org/10.1002/sim.70202.

77. Lekadir K, Frangi AF, Porras AR et al. FUTURE-AI: International consensus guideline for trustworthy and deployable artificial intelligence in healthcare. BMJ 2025; 388: e081554. https://doi.org/10.1136/bmj-2024-081554.

78. Lieberum JL, Toews M, Metzendorf MI et al. Large language models for conducting systematic reviews: On the rise, but not yet ready for use-a scoping review. J Clin Epidemiol 2025; 181: 111746. https://doi.org/10.1016/j.jclinepi.2025.111746.

79. Eisele-Metzger A, Lieberum JL, Toews M et al. Exploring the potential of Claude 2 for risk of bias assessment: Using a large language model to assess randomized controlled trials with RoB 2. Res Synth Methods 2025; 16(3): 491-508. 491. https://doi.org/10.1017/rsm.2025.12.

80. Gutzeit M, Rauh J, Kahler M et al. Modelling volume-outcome relationships in health care. Stat Med 2025; 44(6): e10339. https://doi.org/10.1002/sim.10339.

81. Carlin JB, Moreno-Betancur M. On the uses and abuses of regression models: A call for reform of statistical practice and teaching. Stat Med 2025; 44(13-14): e10244. https://doi.org/10.1002/sim.10244.

82. Shmueli G. To explain, to predict, or to describe: Figuring out the study goal [Commentary on "On the uses and abuses of regression models" by Carlin and Moreno-Betancur]. Stat Med 2025; 44(13-14): e10307. https://doi.org/10.1002/sim.10307.

83. Nold M, Heinze G. Commentary: Teaching statistics as minor subject-handing on fire, not worshipping ashes. Stat Med 2025; 44(13-14): e10284. https://doi.org/10.1002/sim.10284.

84. Greenland S. Some ways to make regression modeling more helpful than misleading. Stat Med 2025; 44(13-14): e10313. https://doi.org/10.1002/sim.10313.

85. Sauerbrei W, Ambrogi F, de Bin R et al. Commentary: Regression models-efforts are required to improve statistical practice and teaching. Stat Med 2025; 44(13-14): e10341. https://doi.org/10.1002/sim.10341.

86. Vansteelandt S, Steen J. Discussion of "On the uses and abuses of regression models: A call for reform of statistical practice and teaching". Stat Med 2025; 44(13-14): e10312. https://doi.org/10.1002/sim.10312.

87. Platt RW. Regression-A Means, Not an End. Stat Med 2025; 44(5): e70000. https://doi.org/10.1002/sim.70000.

88. Carlin JB, Moreno-Betancur M. Rejoinder to commentaries on: On the uses and abuses of regression models: A call for reform of statistical practice and teaching. Stat Med 2025; 44(13-14): e70065. https://doi.org/10.1002/sim.70065.

89. Showell MG, Cole S, Clarke MJ et al. Time to publication for results of clinical trials. The Cochrane Database of Systematic Reviews 2024; 11(11): MR000011. https://doi.org/10.1002/14651858.MR000011.pub3.

Anhang A Gescreente Zeitschriften

1	Biometrical Journal
2	Biometrics
3	BMC Medical Research Methodology
4	British Medical Journal
5	Clinical Trials
6	Cochrane Database of Methodology Reviews
7	Contemporary Clinical Trials
8	International Journal of Epidemiology
9	Journal of Clinical Epidemiology
10	Pharmaceutical Statistics
11	Research Synthesis Methods
12	Statistical Methods in Medical Research
13	Statistics in Medicine
14	Systematic Reviews
15	Trials