MANUAL BEWERTUNG DES BIASRISIKOS IN INTERVENTIONSSTUDIEN

Version 2.0 vom 10.05.2021

Cochrane Deutschland

UNIVERSITÄTS KLINIKUM FREIBURG

AWMF IMWi - Institut für Medizinisches Wissensmanagement

äzq

Die initiale Version dieses Manuals mit dem Titel „Bewertung des Biasrisikos (Risiko systematischer Fehler) in klinischen Studien: ein Manual für die Leitlinienerstellung“ wurde im Rahmen des vom Bundesministerium für Gesundheit (BMG) geförderten Projekts „Acting on Knowledge“ (IIA5-2512MQS006) in Zusammenarbeit von Cochrane Deutschland mit dem Institut für Medizinisches Wissensmanagement der AWMF (AMWF-IMWi) und dem Ärztlichen Zentrum für Qualität in der Medizin (ÄZQ) erstellt. Die Gültigkeit des Manuals wurde auf zunächst drei Jahre (bis 03.05.2019) festgelegt. Die vorliegende erste Aktualisierung und Erweiterung dieses Manual wurde ohne externe Förderung erstellt.

Kommentare zu diesem Manual sind ausdrücklich erwünscht und können gerichtet werden an:
RoB@cochrane.de

KONTAKTE:

Cochrane Deutschland Stiftung
Berliner Allee 2
79110 Freiburg i. Br.
www.cochrane.de

Institut für Evidenz in der Medizin (für Cochrane Deutschland Stiftung)
Universitätsklinikum Freiburg
Breisacher Straße 86
79110 Freiburg i. Br.
www.uniklinik-freiburg.de/institut-fuer-evidenz-in-der-medizin.html

Arbeitsgemeinschaft der Wissenschaftlichen Medizinischen Fachgesellschaften -
Institut für Medizinisches Wissensmanagement (AWMF-IMWi)
Karl von Frisch Str. 1
Philipps Universität
35043 Marburg
www.awmf.org/leitlinien/awmf-imwi.html

Ärztliches Zentrum für Qualität in der Medizin
Tiergarten Tower
Straße des 17. Juni 106-108
10623 Berlin
www.aezq.de

Bitte wie folgt zitieren:

Cochrane Deutschland, Institut für Medizinische Biometrie und Statistik, Freiburg, Arbeitsgemeinschaft der Wissenschaftlichen Medizinischen Fachgesellschaften- Institut für Medizinisches Wissensmanagement, Ärztliches Zentrum für Qualität in der Medizin. *„Manual zur Bewertung des Biasrisikos in Interventionsstudien“*. 2. Auflage, 2021. Verfügbar bei: Cochrane Deutschland: **https://www.cochrane.de/de/literaturbewertung;** ; AWMF: **https://www.awmf.org/leitlinien/awmf-regelwerk/ll-entwicklung.html**; ÄZQ: **https://www.leitlinien.de/methodik**.
DOI: 10.6094/UNIFR/194900, **https://freidok.uni-freiburg.de/data/194900.**

Urheberrecht:

AUTOR*INNEN:

Braun C[1/2], Schmucker C[2], Nothacker M[3], Nitschke K[1], Schaefer C[5], Bollig C[1/2], Muche-Borowski C[3,5], Kopp I[3], Meerpohl JJ[1/2]

LAYOUT:

Conny Wegner[1]

[1] Cochrane Deutschland Stiftung
Berliner Allee 2
79110 Freiburg

[2] Institut für Evidenz in der Medizin (für Cochrane Deutschland Stiftung)
Universitätsklinikum Freiburg
Breisacher Straße 86
79110 Freiburg

[3] Arbeitsgemeinschaft der Wissenschaftlichen Medizinischen Fachgesellschaften - Institut für Medizinisches Wissensmanagement (AWMF-IMWi)
Karl-von-Frisch-Straße 1
Philipps Universität
35043 Marburg

[4]Ärztliches Zentrum für Qualität in der Medizin (ÄZQ)
Tiergarten Tower
Straße des 17. Juni 106-108
10623 Berlin

[5]Institut und Poliklinik für Allgemeinmedizin,
Zentrum für Psychosoziale Medizin
Universitätsklinikum Hamburg-Eppendorf
Martinistraße 52
20246 Hamburg

Wir bedanken uns bei Prof. Dr. Stefan Sauerland für die Durchsicht und Kommentierung der aktualisierten Version des Manuals. Wir bedanken uns bei Dr. Gerta Rücker für die Mitarbeit als Ko-Autorin und bei Prof. Dr. Gerd Antes, Nico Gagelmann, Dipl. Soz. Wiss. Thomas Langer, PD Dr. Petra Lynen, Dr. Jost Schnell und Prof Dr. Karl Werdan für die Durchsicht und Kommentierung der initialen Version des Manuals.

INHALTSVERZEICHNIS

GLOSSAR

Allocation sequence	Zuteilungsfolge in randomisierten kontrollierten Studien
Attrition Bias	Bias durch den Verlust von Teilnehmer*innen während der Studiendurchführung
Bias	Systematischer Fehler, Verzerrung
Biasrisiko	Auch RoB („risk of bias“); Risiko für das Vorliegen von Bias (da Bias nicht quantifizierbar ist)
Cluster-randomisierte Studie	Studie, in der nicht einzelne Personen, sondern „Einheiten“ („cluster“, z.B. Arztpraxen oder Kliniken) den Studiengruppen randomisiert zugeteilt werden.
Confounding (eines Ergebnisses)	Beeinflussung eines Studienergebnisses durch einen oder mehrere „Störfaktoren“ („confounder“). Ein Confounder ist ein Faktor, der nicht direkt Gegenstand der Untersuchung ist, der aber sowohl mit der (Studien-) Intervention (oder -Exposition) als auch mit dem beobachteten Ergebnis einer Studie assoziiert ist. Häufige Confounder sind z.B. Alter, Geschlecht oder Nikotinkonsum.
CONSORT (Statement)	**CON**solidated **S**tandards **O**f **R**eporting **T**rials; Das CONSORT Statement gibt Autor*innen Empfehlungen für die Erstellung von Publikationen oder Berichten von randomisierten kontrollierten Studien in Form einer Checkliste.
Cross-Over-Studie	Studiendesign, in dem die zu vergleichenden Interventionen in den Studiengruppen in zeitlicher Folge angewandt werden. Dabei erhält z.B. die eine Gruppe zunächst Therapie A, dann Therapie B, die andere Gruppe zuerst Therapie B und dann Therapie A.
Effektmaß	Maßzahl, um die Stärke bzw. Größe eines Effekts zu quantifizieren. Häufig verwendete Effektmaße sind für dichotome Endpunkte das relative Risiko (RR) oder Odds Ratio (OR), für kontinuierliche Endpunkte die Mittelwertdifferenz (MD), jeweils mit den zugehörigen Konfidenzintervallen.
Detection Bias	Auch Observer Bias; Bias durch systematische Unterschiede zwischen den Studiengruppen in der Art und Weise, wie die Ergebnisdaten erhoben bzw. die Ergebnisse verifiziert wurden.
Fall-Kontroll-Studie	Retrospektive Beobachtungsstudie, in der eine Gruppe von Personen mit einem Gesundheitsproblem („Fälle“) und eine Gruppe von Personen ohne das Gesundheitsproblem („Kontrollen“) bezogen auf das Vorhandensein von Expositionsfaktoren verglichen werden.
GRADE	**G**rading of **R**ecommendations, **A**ssessment, **D**evelopment and **E**valuation; Ansatz zur Bewertung der Vertrauenswürdigkeit der Evidenz eines Evidenzkörpers („Body of Evidence“).

Indirektheit	Einschränkungen der Übereinstimmung zwischen den Studien in einer Evidenzsynthese in Bezug auf eine oder mehrere PICO-Komponenten (Patient*innen oder Population, Interventionen, Vergleiche, Endpunkte) und der Fragestellung des Reviews. Aspekt der Bewertung der Vertrauenswürdigkeit eines „Evidenzkörpers“ mit dem GRADE-Ansatz.
Information Bias	Auch Measurement Bias; systematische Unterschiede in der Erhebung, Erinnerung, Dokumentation und Handhabung von Informationen in einer Studie, einschließlich des Umgangs mit fehlenden Daten. (Zur Verwendung des Begriffs im Kontext von nicht-randomisierten Studien, s. entsprechenden Manual-Abschnitt).
Inkonsistenz	Auch Heterogenität; Vorliegen widersprüchlicher Studienergebnisse (in Evidenzsynthesen mehrerer Studien zu einer Fragestellung). Aspekt der Bewertung der Vertrauenswürdigkeit eines „Evidenzkörpers“ mit dem GRADE-Ansatz.
Intention-to-Treat Analyse (ITT)	Spezifische Auswertungsmethodik, bei der alle Studienteilnehmenden entsprechend der Studiengruppe analysiert werden, der sie ursprünglich zugeteilt worden waren.
Kohortenstudie	Prospektiv oder retrospektiv angelegte vergleichende nicht-randomisierte Beobachtungsstudie, in der eine Gruppe von Personen (Kohorte) mit einer Intervention oder Exposition und eine Gruppe von Personen ohne die Intervention oder Exposition über einen definierten Zeitraum beobachtet werden, um Unterschiede im Auftreten des interessierenden Gesundheitsproblems zu ermitteln.
Minimierung	Randomisierungsverfahren, um auch bei kleinen Fallzahlen eine Gleichverteilung der Patient*innencharakteristika zu erreichen.
NOS	**N**ewcastle **O**ttawa **S**kala; Instrument für die Bewertung des Biasrisikos in nicht-randomisierten, nicht-vergleichenden Interventionsstudien.
NRSI	Abkürzung für „**N**on-**R**andomized **S**tudy of **I**ntervention“, nicht-randomisierte Interventionsstudie.
Performance Bias	Bias durch systematische Unterschiede in der Art, wie die Teilnehmer*innen der Gruppen in einer Studie behandelt oder betreut werden, z.B. wenn die Teilnehmer*innen einer Gruppe eine zusätzliche (oder längere) Behandlung erhalten, die die Teilnehmer*innen der anderen Gruppe nicht erhalten.
Per-Protocol-Analyse (PP)	Analysemethode, bei der nur die Teilnehmer*innen in die Analyse eingeschlossen werden, bei denen die Intervention(en) protokollgemäß durchgeführt wurden (vgl. Intention-to-treat-Analyse).

Präzision (eines Ergebnisses)	Genauigkeit einer Effekt- bzw. Ergebnisschätzung. Ein präzises Ergebnis wird durch einen möglichst geringen Zufallsfehler bedingt und ist durch ein enges Konfidenzintervall gekennzeichnet. U.a. Aspekt der Bewertung der Vertrauenswürdigkeit eines „Evidenzkörpers“ mit dem GRADE-Ansatz.
Publication Bias	Bias aufgrund der selektiven, bevorzugten Publikation von Studien mit „positiven“ und „signifikanten“ Ergebnissen, die eine größere Chance haben, publiziert zu werden, als Studien mit „negativen“ oder „nicht-signifikanten“ Ergebnissen. U.a. Aspekt der Bewertung der Vertrauenswürdigkeit eines „Evidenzkörpers“ mit dem GRADE-Ansatz.
Randomisierung	Zuteilung von Studienteilnehmer*innen zu den Studiengruppen durch ein Zufallsverfahren (z.B. durch computergenerierte Zufallszahlen)
RCT	**R**andomized **C**ontrolled **T**rial; randomisierte kontrollierte Studie
Reporting Bias	Bias durch Selektion eines berichteten Ergebnisses basierend auf der „Signifikanz“, Größe oder Richtung des Ergebnisses, der Endpunktmessung oder der Analyse
RevMan Web	Webbasierte (neue) Version der Cochrane Review Manager Software zur Erstellung von Cochrane Reviews und Metaanalysen
RoB 2	Name des überarbeiteten Cochrane Risk **o**f **B**ias Tools zur Bewertung des Biasrisikos in randomisierten kontrollierten Studien
ROBINS-I	**R**isk **O**f **B**ias **I**n **N**on-randomized **S**tudies - of **I**nterventions
ROB-ME	Abkürzung von „A tool for assessing **R**isk **O**f **B**ias due to **M**issing **E**vidence in a synthesis“; neu entwickeltes Cochrane Tool zur Bewertung des Biasrisikos durch fehlende Daten in Evidenzsynthesen.
robvis	Web-basierte Anwendung zur Visualisierung der Bewertungen des Biasrisikos in Evidenzsynthesen
Selection Bias	Bias durch systematische Unterschiede in Teilnehmer*innen-Charakteristika zwischen den Studiengruppen bei Studieneinschluss. (Zur Verwendung des Begriffs im Kontext von nicht-randomisierten Studien, s. entsprechenden Abschnitt des Manuals).
Validität, interne	Einschätzung, inwieweit einem ermittelten Studienergebnis in Abhängigkeit von methodischen Aspekten (der Durchführung, Auswertung und Berichterstattung der Studie) vertraut werden kann; erfolgt über die Bewertung des Biasrisikos.
Validität, externe	Generalisierbarkeit oder Übertragbarkeit eines Studienergebnisses in Abhängigkeit von der Fragestellung, den Ein- und Ausschlusskriterien und dem Setting der Studie.

TABELLEN- UND ABBILDUNGSVERZEICHNIS

1 EINLEITUNG

1.1 Warum ist die Bewertung des Biasrisikos in Studien wichtig?

Angehörige aller Gesundheitsberufe treffen täglich eine Vielzahl gesundheitsbezogener Entscheidungen. Diese Entscheidungen basieren überwiegend auf dem in Studium bzw. Ausbildung erlernten Wissen und der persönlichen Erfahrung. Es ist jedoch wichtig, dass bei Entscheidungen im Gesundheitswesen darüber hinaus die Wertvorstellungen und Lebensumstände der Patient*innen und die wissenschaftliche Evidenz, die zu Nutzen und Schaden einer Intervention vorliegt, berücksichtigt werden. Die **Evidenzbasierte Medizin (EbM)** hat zum Ziel, dass Behandlungsentscheidungen für die/den einzelne/n Patient*in auf der Basis der individuellen Erfahrung der Ärztin/des Arztes oder einer anderen Gesundheitsfachperson unter Berücksichtigung der besten verfügbaren Evidenz in Abwägung der Wertvorstellungen und Lebensumstände der/des Patient*in getroffen werden. Das Vorgehen in der EbM gliedert sich in fünf Schritte[1]: **(1)** Übersetzung eines klinischen Problems in eine durch wissenschaftliche Untersuchungen beantwortbare **Fragestellung** (entsprechend dem PICO-Format: P= Patient*innen, I= Intervention, C= Comparator bzw. Vergleich, O= Outcome bzw. Endpunkt), **(2) systematische Suche nach relevanter Evidenz** (Studien) in der medizinischen Literatur oder anderen Quellen[2], **(3) kritische Bewertung der Evidenz** (Bewertung der internen Validität/des Biasrisikos („risk of bias“, RoB) und der klinischen Relevanz der Ergebnisse), **(4) Anwendung der Evidenz auf die/den einzelnen Patient*in,** und **5) kritische Evaluation** des Vorgehens, ggf. mit Anpassungen der Vorgehensweise. Ohne ein ausreichendes Verständnis der methodischen Grundlagen von klinischen Studien, insbesondere im Hinblick auf eine verzerrungsfreie Auswahl und Bewertung der Evidenzbasis, besteht die Gefahr der Fehleinschätzung der vorhandenen Evidenz. In systematischen Übersichtsarbeiten und nachfolgend in Leitlinienempfehlungen kann es zu einer verzerrten Darstellung der Aussagesicherheit von Studienergebnissen kommen. Dies kann eine suboptimale Versorgung bis hin zu Behandlungsfehlern zur Konsequenz haben. Entsprechend wichtig ist es für wissenschaftlich tätige Personen und Gruppen wie Leitlinienentwickelnde- und -beratende oder Autor*innen von systematischen Reviews, über Kompetenzen in der kritischen Bewertung von Evidenz zu verfügen.

Im vorliegenden Manual wird in grundlegende Aspekte der **Bewertung des Biasrisikos in klinischen Studien zu Interventionen (Interventionsstudien)** eingeführt. Unter Interventionsstudien werden in diesem Manual alle klinischen Studien adressiert, deren Ziel die Untersuchung der Effekte (von Nutzen und Schaden) therapeutischer Interventionen ist, wobei der Schwerpunkt des Manuals auf (prospektiven) vergleichenden randomisierten kontrollierten

klinischen Studien (randomized controlled trials, RCTs) und nicht-randomisierten kontrollierten Interventionsstudien (non-randomized studies of interventions, NRSI) liegt. Entsprechend steht die Bewertung des Biasrisikos für diese beiden Studientypen im Vordergrund. Die Bewertung nicht-vergleichender Studien wird kurz adressiert. Die beiden wesentlichen in diesem Manual beschriebenen Bewertungsinstrumente sind **RoB 2**, für randomisierte kontrollierte Studien, und **ROBINS-I**, für nicht-randomisierte Interventionsstudien. Ergänzend wird die **Newcastle-Ottawa Scale** (NOS) für die Bewertung des Biasrisikos in nicht-randomisierten und nicht-vergleichenden Interventionsstudien vorgestellt.

1.2 Wesentliche Neuerungen gegenüber der initialen Version des Manuals

Die erste Version dieses Manuals richtete sich explizit primär an Leitlinienerstellende und-beratende. Da die Inhalte jedoch über diese Gruppe hinaus für eine breitere Zielgruppe relevant und von Interesse sind, haben wir den expliziten Fokus auf die ursprüngliche Zielgruppe herausgenommen. Das Manual richtet sich primär an alle wissenschaftlich tätigen Personen und Gruppen, die sich mit der kritischen Bewertung von Interventionsstudien befassen bzw. sich für diese interessieren.

Die Kapitel zur Bewertung des Biasrisikos in randomisierten und nicht-randomisierten Studien wurden komplett neu verfasst (Kapitel 5 und 6.1-6.3). Der Aufbau des Manuals wurde in Teilen entsprechend verändert. Alle anderen Kapitel wurden editiert und teilweise angepasst bzw. ergänzt. Das Inhaltsverzeichnis, das Glossar und die Literaturliste wurden ergänzt bzw. aktualisiert.

1.3 Ziele und Struktur des Manuals

Das vorliegende Manual richtet sich primär an wissenschaftlich tätige Personen und Gruppen wie Autor*innen von systematischen Reviews und Leitlinienerstellende und -beratende, aber auch an alle, die sich für die kritische Bewertung von Interventionsstudien interessieren. Es soll Nutzer*innen einen Überblick über die qualifizierte Evidenzbewertung vermitteln und in Verbindung mit praktischer Anleitung (z.B. im Rahmen eines Workshops oder Leitlinienseminars) zur eigenständigen Bewertung des Biasrisikos befähigen.

Die Grundlagen dieses Manuals bilden international anerkannte Standards. Das Manual ergänzt und vertieft u.a. das AWMF-Regelwerk zur Erstellung von Leitlinien, insbesondere das Kapitel „Kritische Bewertung der Evidenz".[3]

Das Manual ist in verschiedene Kapitel gegliedert; Kernthemen sind das Biasrisiko in Interventionsstudien (**Kapitel 2**), Verschiedene Biasformen und ihre Auswirkungen (**Kapitel 3**), Bewertungsintrumente (**Kapitel 4**) und die Bewertung des Biasrisikos in verschiedenen Arten von Interventionsstudien (**Kapitel 5**, **Kapitel 6**, **Kapitel 7**).

Das Manual soll weiter kontinuierlich aktualisiert werden. Kommentare sind daher jederzeit ausdrücklich erwünscht: **RoB@cochrane.de**

2 DAS BIASRISIKO IN INTERVENTIONSSTUDIEN (INTERNE VALIDITÄT)

Die Ermittlung des Biasrisikos, d.h. des Risikos für systematische Verzerrungen, ist ein wichtiger Aspekt, um einzuschätzen, inwieweit ein Studienergebnis möglicherweise systematisch vom „wahren“ Ergebnis abweicht, und damit, inwieweit es vertrauenswürdig ist. Zu weiteren Aspekten, die für die Bewertung eines Studienergebnisses wichtig sind, zählen u.a. die externe **Validität** und die **Präzision**, die im Folgenden gegen das Biasrisiko abgegrenzt werden. Die *Bewertung* dieser Aspekte ist nicht Gegenstand dieses Manuals. Begrifflich vom Biasrisiko abzugrenzen sind zudem die **Berichtsqualität** und die **Studienqualität**.

Hinweis: In diesem Manual wird die Bewertung des *Biasrisikos einzelner* Studien beschrieben. Die Bewertung der *Vertrauenswürdigkeit der Evidenz aus mehreren* Studien zu einer Fragestellung (d.h. „Evidenzkörpers“ bzw. „body of evidence“), ist Gegenstand des **GRADE-Ansatzes** (s. Abschnitt 2.5).

2.1 Biasrisiko *versus* externe Validität

Einen wesentlichen Aspekt einer jeden Studie stellt die Validität, d.h. die Gültigkeit der Studienergebnisse, dar.[4] Bei der Validität wird unterschieden zwischen der **internen und externen Validität** (Tabelle 1):

(i) Die **interne Validität** lässt eine Aussage darüber zu, inwieweit dem gemessenen Effekt vertraut werden kann. Sie hängt von der Durchführung, Auswertung und Berichterstattung der Studie ab und wird über die Bewertung des Biasrisikos ermittelt. Bias in klinischen Studien kann sowohl zu einer Über- als auch Unterschätzung der Wirksamkeit und/oder Risiken einer Maßnahme oder Exposition führen.

(ii) Die **externe Validität** hingegen bezeichnet die Generalisierbarkeit oder Übertragbarkeit der Studienergebnisse und hängt damit von der Fragestellung, den Ein- und Ausschlusskriterien und dem Setting der Studie ab. Sie gibt an, ob Studienergebnisse auf andere Personen (-gruppen), Situationen und/oder Zeitpunkte übertragen werden können.

Tab. 1: Wesentliche Aspekte der internen und externen Validität in randomisierten kontrollierten Studien (Quelle: modifiziert nach Jüni et al.[4])

Interne Validität: Ausmaß, in dem der in einer Studie ermittelte Effekt nicht durch systematische Fehler verzerrt wurde				
Selection Bias	**Performance Bias**	**Detection Bias**	**Attrition Bias**	**Reporting Bias**
Verzerrung durch Unterschiede in den Patient*innen-charakteristika zwischen den Studiengruppen bei Studieneinschluss	Verzerrung durch Unterschiede in der Behandlung; abgesehen von der untersuchten Intervention	Verzerrte Erfassung von Endpunkten	Verzerrung durch Unterschiede in der Anzahl und den Gründen für fehlende Daten zwischen den Studiengruppen	Verzerrung durch selektives Berichten von positiven Ergebnissen
Externe Validität: Ausmaß, in dem der in einer Studie ermittelte Effekt auf andere Personen (-gruppen), Situationen und/oder Zeitpunkte übertragen werden kann				
Patient*in		**Behandlungsplan**		**Setting**
Alter, Geschlecht, Schweregrad, (biopsycho-soziale) Risikofaktoren, Ko-Morbidität		Dosierung, Häufigkeit und Art der Verabreichung, Art des Präparats, Begleitbehandlungen		Versorgungsstufe (primär, sekundär, tertiär), Erfahrung und Spezialisierung des Leistungserbringers

Die externe Validität kann zum Problem werden, wenn die in einer Studie untersuchte oder abgebildete Forschungsfrage nicht zu der interessierenden eigenen Fragestellung passt. Unterschiede können sich hierbei in jeder PICO-Komponente (Patient*innen, Intervention, Vergleich, Endpunkte) ergeben. Relevante Unterschiede führen dazu, dass Studienergebnisse nur noch indirekt passen und in ihrer Aussagekraft für die eigene Fragestellung reduziert sind.[5] Ein besonders häufiges Problem in diesem Kontext stellen Surrogatendpunkte dar, die in klinischen Studien oft anstelle patient*innenrelevanter Endpunkte untersucht werden. Wenn sich in Surrogatendpunkten (Laborwerten, Bildbefunden, etc.) ein Interventionseffekt zeigt, lässt sich hieraus nicht oder nur unsicher ein Effekt im patient*innenrelevanten Endpunkt ableiten (es sei denn, die Surrogatendpunkte wurden diesbezüglich bereits validiert).[6] Diese Unsicherheit ist nicht als Biasrisiko zu werten, kann jedoch wie jede andere Einschränkung der Übertragbarkeit von Evidenz zum Beispiel eine Abschwächung einer Leitlinienempfehlung erforderlich machen.

2.2 Biasrisiko *versus* unzureichende Präzision der Ergebnisse

Eine unzureichende **Präzision** von Studienergebnissen ist insbesondere auf kleine Fallzahlen bzw. eine geringe Anzahl an Ereignissen (Events) und nicht auf systematische Fehler (RoB) zurückzuführen, und ist von **Bias** zu unterscheiden. Die Präzision des Effektschätzers wird dabei durch das Konfidenzintervall (KI) angegeben. Ein enges Konfidenzintervall kennzeichnet ein präzises Ergebnis, ein weites Konfidenzintervall ein unpräzises (bzw. weniger präzises) Ergebnis. In

einer Metaanalyse spiegelt sich die Präzision einer Studie im jeweiligen ‚Gewicht' der Studie wider. Studien mit präziseren Ergebnissen (d.h. Studien mit hohen Fall- und Ereigniszahlen) bekommen dabei ein größeres Gewicht als Studien mit nicht oder weniger präzisen Ergebnissen (d.h. Studien mit kleinen Fall- und/oder Eventzahlen).[7]

2.3 Biasrisiko *versus* Studienqualität

Grundsätzlich soll in einer klinischen Studie von der Planung über die Durchführung bis zur Auswertung und Publikation nach einem standardisierten Konzept vorgegangen werden, um eine hohe **Studienqualität** zu gewährleisten. Neben einem Votum einer Ethikkommission gehört dazu vor allem ein publiziertes Studienprotokoll, in dem alle wichtigen Methoden und Vorgehensweisen prospektiv beschrieben werden, sowie der Eintrag der Studie in einem Studienregister (z.B. im Deutschen Register Klinischer Studien (DRKS)[8], ClinicalTrials.gov[9]). Die wichtigsten Qualitätsstandards für (randomisierte) Studien stellen die Grundprinzipien der „Good Clinical Practice" (GCP)[10,11] und für epidemiologische Studien die Empfehlungen zur Sicherung „Guter Epidemiologischer Praxis" (GEP)[12] dar.

Im Gegensatz zur Studienqualität adressiert die **Bewertung des Biasrisikos** das Vertrauen in die im Rahmen einer Studie generierten Effektschätzer. Das Biasrisiko wird von der Qualität einer Studie zwar maßgeblich beeinflusst, die Bewertung des Verzerrungspotentials ist jedoch nicht gleichbedeutend mit einer Qualitätseinschätzung einer Studie. Demzufolge kann ein Biasrisiko in Studien auftreten, die methodisch adäquat durchgeführt wurden. Zum Beispiel ist es in der Chirurgie häufig nicht möglich, Studienteilnehmer*innen und/oder -personal im Hinblick auf die Intervention zu verblinden. Obwohl derartige Studien nach bestmöglichen Standards durchgeführt sein können, können ihre Ergebnisse, bedingt durch die fehlende Verblindung, ein hohes Biasrisiko haben. Andererseits führen nicht alle methodischen Mängel zwangsläufig zu verzerrten Ergebnissen. So wirkt sich z.B. das Fehlen eines Ethikvotums nicht zwangsläufig auf die interne Validität einer Studie aus, erschwert aber die Abschätzung bzw. die Aussage dazu.

2.4 Biasrisiko *versus* Berichtsqualität

Das **Biasrisiko** einer klinischen Studie muss klar von der **Berichtsqualität** abgegrenzt werden.[13] Die Berichtsqualität umfasst Aspekte wie die Vollständigkeit, Detailliertheit, Objektivität und Nachvollziehbarkeit eines Studienberichts (i.d.R. einer Publikation in einer wissenschaftlichen Fachzeitschrift). Studien mit einem niedrigen Biasrisiko können durchaus eine geringe Berichtsqualität aufweisen, wenn zum Beispiel Angaben über wichtige Details zu methodischen

Aspekten wie der Randomisierung oder Verblindung fehlen. Eine mangelhafte Berichtsqualität erschwert jedoch die Bewertung des Biasrisikos oder macht sie ggf. unmöglich, zumal meist unklar ist, ob Aspekte versehentlich oder bewusst nicht oder nur unzureichend berichtet wurden. Auf der anderen Seite können Studien mit einem hohem Biasrisiko, z.B. durch die fehlende Geheimhaltung der Behandlungsfolge (Allocation Concealment), fehlende Verblindung oder einen hohen Verlust an Patient*innen bei der Nachbeobachtung („Loss to follow-up“), eine hohe Berichtsqualität aufweisen, sofern diese Aspekte explizit beschrieben und aus ihnen möglicherweise resultierende Limitationen im Artikel diskutiert wurden. Eine valide Bewertung des Verzerrungspotentials einer Studie setzt eine ausreichende Berichtsqualität voraus.

Zur klareren Abgrenzung zwischen Mängeln in der Berichtsqualität und dem tatsächlichen Biasrisiko wurde von Herausgebenden wissenschaftlicher Zeitschriften, klinisch Forschenden, Epidemiolog*innen und Methodiker*innen zu Beginn der 1990er Jahre eine Initiative zur Verbesserung der Berichtsqualität von Publikationen zu randomisierten kontrollierten Studien ins Leben gerufen. Das Ergebnis war das CONSORT Statement[14] **(CONsolidated Standards Of Reporting Trials Statement)**, eine „Orientierungshilfe“, um die Berichterstattung von randomisierten kontrollierten Studien zu verbessern. Das CONSORT Statement umfasst 25 Aspekte, die in Publikationen zu randomisierten Studien enthalten sein sollten. Eine Überarbeitung des CONSORT Statements erfolgte zuletzt im Jahr 2010.[15] Neben wichtigen Aspekten zur Studienmethodik und Ergebnisdarstellung wird im CONSORT Statement ein Flussdiagramm gefordert, das die Anzahl der Patient*innen bzw. Teilnehmenden (einschließlich fehlender Daten) von Beginn bis Ende einer Studie abbildet. Wenige Jahre nach Veröffentlichung des CONSORT Statements verbesserte sich in drei hochrangigen Zeitschriften, die die obligatorische Verwendung des CONSORT Statements vorsehen (British Medical Journal, Journal of the American Medical Association, The Lancet), die Berichtsqualität von randomisierten kontrollierten Studien signifikant.[16] Die „Hauptversion“ des CONSORT Statement basiert auf dem „Standard-Design“ einer Studie mit zwei parallelen Studiengruppen. Um Varianten dieses Designs gerecht zu werden, wurden in den vergangenen Jahren verschiedene Versionen des CONSORT Statements, die sogenannten **CONSORT Extensions**, entwickelt, unter anderem für Cluster-randomisierte Studien, randomisierte Cross-Over-Studien und randomisierte Pilot- und Machbarkeitsstudien. Eine komplette Zusammenstellung aller verfügbaren CONSORT Extensions findet sich auf den **CONSORT-Internetseiten**.[14] Weitere Leitfäden für die Berichterstattung anderer Studientypen (als randomisierter kontrollierter Studien), sowohl für Studienautor*innen als auch für Herausgeber*innen von Zeitschriften und Gutachter*innen, sind auf **den Internetseiten des**

EQUATOR Netzwerks[17] verfügbar (siehe Abschnitt 8: Weiterführende Informationen und Praxishilfen).

✓ Werden in einer Publikation Standards der Berichterstattung nicht eingehalten, so erschwert dies die Bewertung einer Studie. Wichtig ist ein einheitliches Vorgehen bei der Bewertung, gerade, wenn diese durch zwei Bewertende erfolgt. So gilt es zu unterschieden, ob ein Aspekt, möglicherweise aufgrund einer schwachen Berichtsqualität, erfüllt aber nicht (oder nur unzureichend) berichtet ist, oder ob er, möglicherweise bei guter Berichtsqualität, tatsächlich nicht erfüllt ist und damit ein Biasrisiko vorliegt. Eine Option bei unklarer Berichterstattung stellt die Kontaktaufnahme mit den Studienautor*innen dar. Ggf. lassen sich dadurch fehlende, beziehungsweise unklare Studienaspekte, die für eine valide Bewertung des Biasrisikos ausschlaggebend sind, klären.

2.5 Biasrisiko im Kontext von GRADE (Bewertung der Vertrauenswürdigkeit der Evidenz)

GRADE (Grading of Recommendations, Assessment, Development and Evaluation) bietet einen strukturierten, transparenten Ansatz, um die Vertrauenswürdigkeit der zu einer Fragestellung verfügbaren Evidenz einzuschätzen.[18] Die GRADE-Bewertung bezieht sich dabei nicht auf einzelne Studien, sondern auf die Gesamtheit der identifizierten Evidenz für ein Ergebnis, d.h. die Evidenz wird jeweils bezogen auf ein spezifisches (in der Regel numerisches) Ergebnis bewertet. Für die Studien, die in einen solchen „Evidenzkörper“ („body of evidence“) eingehen, wird das Vorliegen einer Bewertung des Biasrisikos vorausgesetzt. Bei der Bewertung der Vertrauenswürdigkeit der gesamten Evidenz, wie sie in GRADE erfolgt, wird somit zunächst das (herkömmliche) Biasrisiko der Studie beurteilt. Zusätzlich werden weitere Aspekte beurteilt. Ein Evidenzkörper aus einer Anzahl adäquat geplanter und gut durchgeführter Studien kann beispielsweise mit einem geringen **Biasrisiko** behaftet sein. Trotzdem kann das Vertrauen in das Ergebnis durch andere Faktoren beeinträchtigt sein, wie z.B. eine **unzureichende Präzision** (ein weites Konfidenzintervall), **Inkonsistenz** (das Vorliegen widersprüchlicher Studienergebnisse), **Indirektheit** (Einschränkungen in der Übereinstimmung zwischen der interessierenden Population, Intervention, Vergleichsintervention oder den interessierenden Endpunkten und der in den eingeschlossenen Studien untersuchten Populationen, Interventionen, Vergleichsinterventionen oder Endpunkten) und **Publication Bias** (Bias durch das Nicht-Publizieren von Studien, z.B. solchen mit „negativen“ oder „nicht-signifikanten“ Ergebnissen). Die systematische Einschätzung all dieser Faktoren fließt in die Bewertung nach GRADE ein. Weitere Aspekte, die in GRADE, insbesondere im Kontext der Bewertung von nicht-randomisierten Interventionsstudien, berücksichtigt werden können, sind ein **großer Effekt**, eine **Dosis-Wirkungsbeziehung** und ein **plausibles Confounding**.

3 VERSCHIEDENE BIASFORMEN UND IHRE AUSWIRKUNG

3.1 Biasformen

In der Literatur wird eine Vielzahl von **Biasformen** beschrieben, die innerhalb klinischer Studien auftreten können. Die derzeit wahrscheinlich umfassendste Übersicht bietet der **Catalogue of Bias**.[19] Dieser geht auf eine Initiative von Wissenschaftler*innen der Oxford University, Großbritannien, aus dem Jahr 2017 zurück, die sich zum Ziel gesetzt haben, sämtliche Biasformen, die gesundheitsbezogene Evidenz beeinflussen können, in *einer* digitalen Ressource zusammenzuführen. Der Catalogue of Bias enthielt im September 2020 rund 60 verschiedene Biasformen. Er wird fortlaufend ergänzt.

In Tabelle 2 sind einige wesentliche Biasformen mit Relevanz für (randomisierte und nicht-randomisierte) Interventionsstudien aufgeführt.

Tab. 2: Wesentliche Biasformen in Interventionsstudien

Biasform	Merkmale
Selection Bias	Systematische Unterschiede in Teilnehmer*innen-Charakteristika zwischen den Studiengruppen.
Performance Bias	Systematische Unterschiede in der Versorgung der Teilnehmer*innen der Interventions- und Vergleichsgruppen, die zustande kommen können, wenn die Teilnehmer*innen oder das an der Verabreichung bzw. Durchführung der Intervention(en) beteiligte Studienpersonal Kenntnis davon haben, welche Teilnehmer*innen welcher Gruppe zugeteilt wurden.
Detection Bias (auch Observer Bias)	Systematische Unterschiede zwischen den Studiengruppen in der Art und Weise, wie die Ergebnisdaten erhoben bzw. die Ergebnisse verifiziert wurden.
Attrition Bias	Systematische Unterschiede zwischen den Studiengruppen in Bezug auf die Teilnehmer*innen, die aus der Studie ausgeschieden sind und denjenigen, die in der Studie (bis Studienende) verbleiben.
Information Bias	Systematische Unterschiede in der Erhebung, Erinnerung, Dokumentation und Handhabung von Informationen in einer Studie, einschließlich des Umgangs mit fehlenden Daten.
Reporting Bias	Beeinflussung der Berichterstattung von Forschungsergebnissen aufgrund der Art und Richtung der Ergebnisse.

Die in der Tabelle dargestellten klassischen Bezeichnungen der Biasformen finden sich in den neueren Bewertungsinstrumenten wie dem RoB 2 Tool und dem ROBINS-I Tool (bzw. den ihnen zugrunde liegenden Ressourcen) wieder. In den Benennungen der Biasdomänen dieser Instrumente sind sie jedoch nicht mehr zwingend enthalten, da diese sich teilweise stattdessen auf Aspekte der

Durchführung der zu bewertenden Studie beziehen. Beispiele sind „Bias durch den Randomisierungsprozess“ (RoB 2), „Bias durch Confounding“ (ROBINS-I) oder „Bias durch fehlende Daten“ (RoB 2 und ROBINS-I). Diese Aspekte lassen sich den klassischen Bezeichnungen bzw. Biasformen zuordnen.

Kapitel 5 und 6 enthalten ausführliche Beschreibungen der für Interventionsstudien relevanten Biasformen, getrennt für randomisierte kontrollierte Studien (Kapitel 5) und nicht-randomisierte Interventionsstudien (Kapitel 6). Da die beiden Bewertungsinstrumente RoB 2 (für randomisierte Studien) und ROBINS-I (für nicht-randomisierte Studien) die wesentlichen in diesem Manual vorgestellten Bewertungsinstrumente sind, wurden die Beschreibungen zu den Biasformen bewusst mit engem Bezug zu diesen Instrumenten und basierend auf den zu ihnen verfügbaren Ressourcen erstellt und orientieren sich an der in ihnen verwendeten Terminologie. Die Beschreibungen sind jedoch nicht auf die Anwendung dieser beiden Instrumente beschränkt. Einige der Beschreibungen zu den wichtigen Biasformen in den Kapiteln 5 und 6 gleichen bzw. ähneln sich. Etwaige Redundanzen bzw. Ähnlichkeiten wurden bewusst belassen, damit die beiden Kapitel – je nach interessierendem Studientyp – unabhängig voneinander verwendet werden können.

3.2 Auswirkung von Bias auf die Ergebnisse von Interventionsstudien

Eine Zusammenfassung von Daten aus sieben meta-epidemiologischen Studien zeigte eine **Überschätzung des Behandlungseffekts** bei inadäquater oder fehlender Randomisierung um im Durchschnitt 11% (95%-KI 4% bis 18%).[20] Durch eine fehlende oder inadäquate verdeckte Gruppenzuteilung wurden Behandlungseffekte um 7% (95%-KI 1% bis 13%), und im Fall einer fehlenden oder inadäquaten doppelten Verblindung um 13% (95%-KI 4% bis 21%) überschätzt.[20]

Eine meta-epidemiologische Untersuchung, in der der Effekt medizinischer Interventionen bezogen auf die Mortalität aus 16 Studien mit routinemäßig erhobenen Daten (RCD-Studien) und 36 nachfolgend durchgeführten RCTs zu derselben Fragestellung verglichen wurden, zeigte zusammengefasst eine systematische und substantielle Überschätzung des Effekts um durchschnittlich 1,34 bzw. 1,39 (relatives Odds Ratio, ROD; 95% KI: 1,06-1,69 bzw. 1,10-1,75).[21,22]

Das Ausmaß eines Bias aufgrund einer fehlenden oder inadäquaten Verblindung hängt wesentlich von den Untersuchungsparametern bzw. Endpunkten und der Intervention ab. Im Fall des Endpunktes Mortalität ist eine fehlende Verblindung mit einem deutlich geringeren Biasrisiko verbunden als bei einem subjektiven Endpunkt, der Interpretationsspielraum lässt. Bei chirurgischen Interventionen ist eine Verblindung von Chirurg*in und Patient*innen seltener

realisierbar oder, ggf. abhängig von der jeweiligen Prozedur, aus ethischen Gründen nicht zu rechtfertigen.[7,23] Bei bestimmten Endpunkten, wie zum Beispiel der intraoperativen Blutungsmenge, kann auch unabhängiges Studienpersonal zur Endpunkterhebung eingesetzt werden.[24] Das Ergebnis einer meta-epidemiologischen Studie war, dass eine inadäquate oder fehlende Verblindung bei subjektiven Endpunkten zu einer Überschätzung der Studienergebnisse von 25% (95%-KI 7% bis 39%) führt.[25]

Untersuchungen zum Ausmaß von Attrition Bias sind bis dato wenig aussagekräftig und generell schwierig. Fehlen nämlich in Publikationen die Daten von Teilnehmer*innen und Informationen darüber, warum Teilnehmer*innen eine Studie abgebrochen haben, oder sind die Angaben nicht nach Behandlungsgruppen differenziert berichtet, ist es nicht sicher möglich, das Ausmaß dieses Bias abzuschätzen.

Ob die industrielle beziehungsweise kommerzielle Finanzierung von Studien als eigenständiges Bias-Kriterium berücksichtigt werden sollte, ist kontrovers diskutiert worden[26,27] und ist nicht abschließend geklärt. Autor*innen eines Cochrane Reviews zu industrieller Forschungsförderung und Forschungsergebnissen konnten zeigen, dass eine Finanzierung bzw. finanzielle Unterstützung durch industrielle Förderer (Medikamente und Medizinprodukte) zu positiveren Ergebnissen und Schlussfolgerungen bezüglich der Wirksamkeit sowie schädlicher Wirkungen bzw. Nebenwirkungen einer Intervention führt.[28] Im Gegensatz hierzu kann argumentiert werden, dass in aller Regel nur große (und damit meist industrielle) Förderer über ausreichende Mittel verfügen, um robuste Studienprozesse zu ermöglichen bzw. gewährleisten.

Cochrane spricht sich explizit gegen den direkten Einschluss von Interessenkonflikten in die Bewertung des Biasrisikos aus.[29] Ein Grund hierfür ist, dass die Beschränkung der Betrachtung des potenziellen Einflusses von Interessenkonflikten in einer Studie auf *eine* Biasrisiko-Frage andere wichtige Aspekte außer Acht lässt, wie z.B. das Studiendesign und einen potenziellen Bias einer Meta-Analyse durch fehlende Daten. Diese sind Teil der Bewertung des Biasrisikos nach Cochrane-Methoden, ebenso wie das selektive Berichten (oder Nicht-Berichten) von Studienergebnissen.

4 BEWERTUNGSINSTRUMENTE

Zur standardisierten Bewertung des Biasrisikos steht eine Vielzahl von Instrumenten zur Verfügung, die sich im Wesentlichen drei Kategorien zuordnen lassen: **Skalen, Checklisten und domänenbasierte Instrumente (Komponentensysteme)**.[30]

Skalen

Anhand einer **Skala** werden verschiedene interne und externe Validitätsaspekte einer Studie mit Hilfe eines „Scores" (Punktezahl) numerisch bewertet. Für die Gesamtbewertung der Studie werden die einzelnen Punkte addiert. Bewertungsverfahren auf Skalenbasis werden im Allgemeinen wegen mangelnder Evidenz im Hinblick auf die Gewichtung der einzelnen Bewertungsaspekte durch die empirische Forschung nicht gestützt.[31,32] Zudem bewerten Skalen eher die Berichtsqualität als das Verzerrungspotential. Eine nicht im Detail dargestellte Methodik muss allerdings nicht unbedingt bedeuten, dass die Studie mangelhaft durchgeführt wurde (siehe Abschnitt 2.4: Biasrisiko versus Berichtsqualität). Einige Skalen, wie die bekannte Jadad Skala (Oxford Skala) für randomisierte kontrollierte Studien[33], berücksichtigen zudem nicht den wichtigen Biasrisiko-Aspekt der verdeckten Gruppenzuteilung („allocation concealment").[32] Es wurden außerdem substanzielle Unterschiede in den Ergebnissen von Metaanalysen berichtet, die auf die Verwendung unterschiedlicher Skalen bei randomisierten kontrollierten Studien zurückzuführen waren.

Zur Bewertung von nicht-randomisierten Studien (insbesondere von Fall-Kontrollstudien und Kohortenstudien) wird häufig die Newcastle Ottawa Skala (NOS) herangezogen.[34] Die Kritikpunkte dieser Skala gleichen denen an Skalen für randomisierte kontrollierte Studien (siehe oben).

Checklisten

Checklisten, wie zum Beispiel die des **S**cottish **I**ntercollegiate **G**uidelines **N**etworks (SIGN)[35], bewerten meist neben der internen auch die externe Validität einer Studie. Die externe Validität muss von der Biasrisiko-Bewertung unterschieden werden (siehe Abschnitt 2.1: Biasrisiko versus externe Validität einer Studie).[36]

Domänenbasierte Instrumente (Komponentensysteme)

In **domänenbasierten Bewertungsinstrumenten**, wie z.B. dem RoB 2 Tool für randomisierte kontrollierte Studien[37] oder dem ROBINS-I Tool für nicht-randomisierte kontrollierte Studien (**R**isk **o**f **B**ias **I**n **N**on-randomized **S**tudies of **I**nterventions[38]), wird das Biaspotential für verschiedene Domänen (Komponenten) getrennt dargestellt. Domänenbasierte Bewertungsinstrumente sind in der Regel deutlich umfangreicher als Skalen und Checklisten.

✓ Allen Instrumenten ist gemein, dass sie keine exakte Messung (Quantifizierung), sondern eine *Einschätzung* des Biasrisikos und damit immer ein Werturteil sind.

5 BEWERTUNG DES BIASRISIKOS IN RANDOMISIERTEN KONTROLLIERTEN STUDIEN

5.1 Definition: Randomisierte kontrollierte Studien

Randomisierte kontrollierte Studien gelten in der klinischen Forschung als Goldstandard für die Evaluation der Wirksamkeit von Interventionen. In der Arzneimittelentwicklung stellen sie die Grundlage für Zulassungsentscheidungen der Behörden dar.

5.2 Wichtige Biasformen in randomisierten kontrollierten Studien

Da in diesem Manual das RoB 2 Tool für die Bewertung des Biasrisikos in randomisierten Interventionsstudien vorgestellt wird, basiert die folgende Beschreibung der wesentlichen Biasformen in randomisierten Interventionsstudien auf den zum RoB 2 Tool verfügbaren Ressourcen[37,39,40] (s. hierzu auch Abschnitt 5.3.1), und orientiert sich an der im RoB 2 Tool verwendeten Terminologie. Das RoB 2 Tool ist das von Cochrane empfohlene Bewertungsinstrument für die Nutzung in Cochrane Reviews.[40]

5.2.1 Bias durch den Randomisierungsprozess

Eine erfolgreich durchgeführte **Randomisierung („randomization“)** verhindert den Einfluss bekannter und unbekannter prognostischer Faktoren (wie z.B. dem Alter oder dem Schweregrad einer Erkrankung) auf die Zuteilung zu den Interventionsgruppen. Dies bedeutet, dass die Gruppen vor Beginn der Intervention im Mittel eine vergleichbare Prognose haben. Wenn prognostisch relevante Faktoren die Zuteilung zu den Interventionsgruppen beeinflussen, liegt ein „Confounding“ vor und werden die beobachteten Wirkungen der Interventionen verzerrt. Confounding ist eine kritische potenzielle Ursache für Bias von Effektschätzungen in Beobachtungsstudien, da Behandlungsentscheidungen in der Routineversorgung häufig durch prognostische Faktoren beeinflusst werden.

Die randomisierte Zuteilung von Studienteilnehmer*innen zu Gruppen erfordert die **Generierung einer Zuteilungssequenz („allocation sequence generation“)**, anhand derer festgelegt ist, wie die Teilnehmer*innen den Interventionen zugeteilt werden, und die zwingend auf einem Zufallsverfahren basieren muss (z.B. via computergenerierter Zufallszahlen). Nach Generierung der Sequenz muss sichergestellt werden, dass die Teilnehmer*innen und das Studienpersonal die Zuteilungen erst nach Bestätigung der Studienteilnahme erfahren. Dieses Vorgehen wird als **„Verbergen der Zuteilungssequenz“ („allocation sequence concealment“)** bezeichnet. Wenn

die jeweils nächsten Gruppenzuteilungen (z.B. durch eine offene Liste) bekannt sind, ermöglicht dies eine selektive Zuteilung der Teilnehmer*innen (auf Grundlage prognostischer Faktoren) und bedeutet damit ein Biasrisiko. Aus diesem Grund ist das erfolgreiche Verbergen der Zuteilungssequenz ein wesentlicher Bestandteil der Randomisierung.

Das Verbergen der Zuteilungssequenz ist nicht mit der Verblindung während der Studiendurchführung zu verwechseln. Das Verbergen der Zuteilungssequenz soll Bias während des Prozesses der Zuteilung zu den Interventionsgruppen verhindern; es ist grundsätzlich immer möglich, unabhängig vom Studiendesign oder dem Setting der Studie.

Die Generierung der Zuteilungssequenz für die Randomisierung kann mit verschiedenen Methoden erfolgen. Die Randomisierung ohne Einschränkungen der Generierungssequenz wird als **einfache Randomisierung („simple randomisation“)** oder **uneingeschränkte Randomisierung („unrestricted randomisation“)** bezeichnet.

Soll ein bestimmtes Verhältnis in der Anzahl der Teilnehmer*innen zwischen den Interventionsgruppen erzielt werden, so wird zumeist eine **Blockrandomisierung („restricted randomization“, „block randomization“)** durchgeführt. Bei dieser werden die Teilnehmer*innen in Blöcken mit einer bestimmten Größe (z.B. jeweils 12 Teilnehmer*innen) im gewünschten Verhältnis (z.B. 1:1) randomisiert, wobei die Zuteilungen zu den Gruppen innerhalb jedes Blocks in zufälliger Reihenfolge erfolgen. Die Vorhersagbarkeit der Zuteilung bei konstanter Blockgröße kann durch Verwendung unterschiedlicher Blockgrößen und einer randomisierten Blockfolge („random permuted blocks“) verhindert werden.

Bei der **stratifizierten Randomisierung (auch geschichtete Randomisierung; „stratified randomisation“)** werden die Teilnehmer*innen in getrennten Untergruppen randomisiert, denen potenziell wichtige prognostische Faktoren zugrunde liegen (z.B. entsprechend Altersgruppen oder Studienzentren). Die stratifizierte Randomisierung wird häufig zusammen mit einer Block-Randomisierung durchgeführt. Ziel dieses Vorgehens ist es, sicherzustellen, dass sich die Interventions- und Vergleichsgruppen jenseits der Intervention in Bezug auf bestimmte prognostische Faktoren nicht unterscheiden.

Minimierung („minimization“) ist ein Ansatz zur Randomisierung, der die Konzepte der Stratifizierung und der Block-Randomisierung (eingeschränkten Randomisierung) vereint. Bei der Minimierung erfolgt die jeweils folgende Zuteilung zu einer Interventionsgruppe auf eine Weise, die ein bestmögliches Gleichgewicht zwischen den Interventions- und Vergleichsgruppen in Bezug auf

bestimmte prognostische Faktoren gewährleistet. Die Minimierung beinhaltet im Allgemeinen ein Element der zufälligen Zuteilung und sollte mit einer klaren Strategie für die verdeckte Zuteilung (s. weiter) einhergehen.

Klar abzugrenzen von der „echten“ Randomisierung ist die „Quasi“- oder „Pseudo“-Randomisierung. Zu dieser zählen z.B. Methoden, bei denen die Studienteilnehmer*innen in der Reihenfolge des Studieneinschlusses abwechselnd der einen oder der anderen Intervention zugeteilt werden (sog. alternierende Zuteilung) oder die Studienteilnehmer*innen entsprechend ihres Geburtsdatums (z.B. gerades oder ungerades Geburtsjahr) einer der beiden Gruppen zugeteilt werden. Quasi-randomisierte Studien zählen nicht als RCTs, weil hier die jeweilige Zuteilung der Teilnehmer*innen vorhersehbar ist und ggf. beeinflusst werden kann.

Der Erfolg der Randomisierung in Bezug auf die Generierung vergleichbarer Interventionsgruppen wird häufig anhand des Vergleichs der Werte wichtiger prognostischer Faktoren zwischen den Gruppen zum Studienbeginn („at baseline“) beurteilt. Von vorrangiger Bedeutung ist jedoch die Betrachtung von Situationen, in denen Unterschiede in Baseline-Charakteristika darauf hindeuten, dass der Randomisierungsprozess möglicherweise fehlgeschlagen ist. Es ist dabei wichtig zu beachten, dass Unterschiede, die auf den Zufall zurückzuführen sind, nicht als Evidenz für ein Biasrisiko zu betrachten sind. Zufällig zustande gekommene Unterschiede schränken die Aussagekraft der Ergebnisse ein, sind jedoch keine Quelle für einen systematischen Bias.

5.2.2 Bias durch Abweichungen von den vorgesehenen Interventionen

Die **vorgesehenen Interventionen („intended interventions“)** sind die im Studienprotokoll spezifizierten Interventionen. **Abweichungen von den vorgesehenen Interventionen („deviations from the intended interventions“)** können zu verzerrten Schätzungen des Interventionseffekts führen. Bias durch Abweichungen von den vorgesehenen Interventionen wird mitunter auch als **Performance Bias** bezeichnet. Zu möglichen Abweichungen von den vorgesehenen Interventionen zählen die Verabreichung von zusätzlichen, nicht mit dem Studienprotokoll in Einklang stehenden Interventionen, die unzureichende Implementierung der vorgesehenen Interventionen und ihre Nicht-Einhaltung durch Teilnehmer*innen. Bei der Bewertung des Biasrisikos gilt es zu ermitteln, welche Abweichungen von der vorgesehenen Intervention von den Studienautor*innen vorgesehen und möglicherweise beabsichtigt waren, und welche als Abweichungen von den vorgesehenen Interventionen zu betrachten sind.

Die Bewertung des Biasrisikos aufgrund von Abweichungen von den vorgesehenen Interventionen hängt auch vom interessierenden **Effekt bzw. Ergebnis der Intervention ab: dem Effekt der Zuteilung zur Intervention (Intention-to-treat-Effekt)** oder dem **Effekt der Einhaltung der Intervention (Per-Protocol-Effekt)** (s. auch 5.3.4). So können (und werden im RoB 2 Tool) die folgenden Abweichungen in Bezug auf den Effekt der Zuteilung zur Intervention adressiert werden: 1) Abweichungen, die nicht mit dem Protokoll übereinstimmen, 2) Abweichungen, die aufgrund des experimentellen (Studien-) Kontextes zustande kommen, und 3) Abweichungen, die einen Einfluss auf das Ergebnis haben. Abweichungen von der Intervention, die nicht aufgrund des experimentellen Kontextes zustande kommen, wie z.B. die Entscheidung von Patient*innen, die Einnahme der ihnen zugeteilten Medikation zu beenden, führen nicht zu einem Bias des Effektes der Zuteilung zur Intervention.

Die Bewertung des Effekts der Einhaltung der Intervention kann die folgenden Überlegungen beinhalten (die im RoB 2 Tool adressiert werden): 1) wie gut die Intervention implementiert wurde, 2) wie gut die Teilnehmer*innen die Intervention einhielten (ohne diese abzubrechen oder zu einer anderen Intervention zu wechseln), und 3) ob die Teilnehmer*innen neben den vorgesehenen Interventionen zusätzliche, nicht im Protokoll spezifizierte, Interventionen erhielten, und (wenn), ob diese zwischen den Gruppen ausgewogen waren. Bei Vorliegen entsprechender Abweichungen sollten Gutachter*innen betrachten, ob angemessene Methoden zur Adjustierung ihrer Effekte angewendet wurden.

Bias aufgrund von Abweichungen von den vorgesehenen Interventionen kann mitunter durch die Implementierung von Maßnahmen verringert oder vermieden werden, mit denen sichergestellt wird, dass Teilnehmer*innen und das Studienpersonal, d.h. die Personen, die an der Verabreichung der Interventionen beteiligt sind, keine Kenntnis von den Interventionen haben, die die Teilnehmer*innen erhalten. Dies wird weithin als **Verblindung („blinding“)** bezeichnet. Eine erfolgreiche Verblindung kann verhindern, dass Kenntnisse über die Zuteilung der Interventionen zu Kontamination (die Abgabe einer der Interventionen an Teilnehmer*innen, die die andere erhalten sollten), Wechsel zu nicht-Protokoll-Interventionen oder die Nichteinhaltung der Intervention durch Teilnehmer*innen führen. Die Verblindung ist in manchen Fällen schwierig oder nicht durchführbar, wie z.B. in Studien, in denen eine operative mit einer nicht-operativen Intervention verglichen wird.

Die fehlende oder unzureichende Verblindung von Teilnehmer*innen oder dem an der Verabreichung der Interventionen beteiligten Studienpersonal kann Bias verursachen, wenn sie zu Abweichungen von den vorgesehenen Interventionen führt. So kann zum Beispiel eine niedrige

Erwartung an eine Verbesserung bei Teilnehmer*innen einer Vergleichsgruppe dazu führen, dass sie sich darum bemühen, die experimentelle Intervention zu erhalten. Derartige Abweichungen von den vorgesehenen Interventionen können zu einer Verzerrung sowohl des Effektes der Zuteilung zur Intervention als auch des Effektes der Einhaltung der Intervention führen. Andere Abweichungen, wie das Absetzen einer erhaltenen Intervention bei Teilnehmer*innen aufgrund von Nebenwirkungen, die spezifisch für die experimentelle Intervention sind, führen als solche hingegen nicht zu einem Biasrisiko.

Das Biasrisiko bezogen auf die Verblindung kann sich auch zwischen Endpunkten unterscheiden. So kann die Kenntnis der zugeteilten Intervention z.B. ein Verhalten beeinflussen (wie die Anzahl von Praxisbesuchen), während sie keinen bedeutsamen Einfluss auf andere Endpunkte wie z.B. die Mortalität haben wird.

5.2.3 Bias durch fehlende Ergebnisdaten

Fehlende Ergebnisdaten („missing outcome data“) können zu einer verzerrten Schätzung des Effekts der interessierenden Intervention führen (Attrition Bias). Für das Fehlen von Ergebnisdaten kann es verschiedene Gründe geben:

- Teilnehmer*innen brechen die Studie ab („dropout“) oder können nicht mehr ausfindig gemacht werden („loss to follow-up“)
- Teilnehmer*innen nehmen einen Untersuchungstermin nicht wahr, bei dem Ergebnisse erhoben werden sollten
- Teilnehmer*innen nehmen einen Untersuchungstermin wahr, liefern aber keine relevanten Daten
- Daten oder Aufzeichnungen gehen verloren oder sind aus anderen Gründen nicht verfügbar
- Teilnehmer*innen können das Ergebnis nicht mehr erleben (z.B., weil sie gestorben sind)

Bias durch fehlende Ergebnisdaten beinhaltet auch Bias durch den *Umgang* mit den fehlenden Daten (z.B. durch Imputation oder eine anderweitige Adjustierung für die fehlenden Daten).

Das **Intention-to-Treat-Prinzip** (ITT-Prinzip) des Erhebens und Einschließens aller Ergebnisdaten in die Ergebnis-Analyse ist in der Praxis häufig schwierig aufgrund unvollständiger Daten und kann dann nur unter Zuhilfenahme von Annahmen zu den fehlenden Daten angewandt werden. Wenn Teilnehmer*innen mit fehlenden Ergebnisdaten aus einer als „ITT“ bezeichneten Analyse ausgeschlossen werden, ist diese Analyse mit einem Biasrisiko behaftet (derartige Analysen werden mitunter als „modifizierte ITT-Analysen“ bezeichnet).

Analysen, aus denen Teilnehmer*innen mit fehlenden Ergebnisdaten ausgeschlossen wurden, sind Beispiele für **Complete-Case-Analysen**, d.h. Analysen, die auf Teilnehmer*innen beschränkt sind, bei denen für die untersuchten Variablen keine Werte fehlten. Um zu verstehen, wann fehlende Ergebnisdaten in solchen Analysen zu Bias führen, müssen folgende Aspekte betrachtet werden: a) der wahre Wert („true value“) des Ergebnisses der Teilnehmer*innen mit fehlenden Ergebnisdaten; dies ist der Wert des Ergebnisses, der erhoben werden sollte, aber nicht erhoben wurde; und b) der dem Fehlen der Ergebnisdaten zugrunde liegende Mechanismus („missingness mechanism“); dies ist der Prozess, der zum Fehlen der Ergebnisdaten geführt hat.

Ob fehlende Ergebnisdaten in Complete-Case-Analysen zu Bias führen, hängt davon ab, ob der dem Fehlen der Ergebnisdaten zugrunde liegende Mechanismus in einem Zusammenhang mit dem wahren Wert des Ergebnisses steht. Entsprechend kann betrachtet werden, ob die erhobenen (nicht fehlenden) Ergebnisse systematisch von den fehlenden Ergebnissen (den wahren Werten der Teilnehmer*innen mit fehlenden Daten) abweichen. Ein Beispiel könnte eine Studie zur Behandlung von Depressionen sein, in der die Ergebnisse einer kognitiven Verhaltenstherapie mit denen der Regelversorgung verglichen werden. Wenn die Wahrscheinlichkeit, zur Nachuntersuchung zu erscheinen, bei denjenigen Teilnehmer*innen geringer ist, die an schwerwiegenderen Depressionen leiden, hängt das Fehlen depressionsbezogener Ergebniswerte von ihrem wahren Wert ab, was impliziert, dass in diesem Fall die erhobenen depressionsbezogenen Ergebnisse systematisch von den wahren Werten der fehlenden depressionsbezogenen Werte abweichen werden.

Detaillierte Erläuterungen zu spezifischen Situationen, in denen eine Complete-Case-Analyse von Bias betroffen ist, finden sich u.a. im **full guidance document**[39] zum RoB 2 Tool.

Die Frage, wann der Umfang fehlender Ergebnisdaten gering genug ist, um Bias auszuschließen, kann nicht pauschal beantwortet werden, da es keinen allgemeingültigen Grenzwert für ein „gering genug“ in Bezug auf den Anteil fehlender Daten gibt. Der potentielle Einfluss fehlender Daten auf die Schätzung eines Interventionseffektes hängt neben dem Anteil der Teilnehmer*innen mit fehlenden Daten auch von der Art des Ergebnisses und (bei dichotomen Ergebnissen) dem Risiko für das Ergebnisereignis ab.

Es ist nicht möglich *direkt* zu ermitteln, ob die Wahrscheinlichkeit des Fehlens von Ergebnisdaten von ihrem wahren Wert abhängt. Die Beurteilung des Biasrisikos hängt von den Gegebenheiten der jeweiligen Studie ab. Sicherheit, dass kein Bias durch fehlende Ergebnisdaten vorliegt, besteht nur, wenn 1) das Ergebnis bei allen Teilnehmer*innen erhoben wurde, 2) der Anteil fehlender Ergebnisdaten so gering ist, dass ein daraus resultierender Bias zu gering wäre, um bedeutsam zu

sein, oder 3) Sensitivitätsanalysen bestätigen, dass plausible Werte für die fehlenden Ergebnisdaten keinen bedeutsamen Unterschied im geschätzten Interventionseffekt machen. Indirekte Evidenz dafür, dass fehlende Ergebnisdaten wahrscheinlich zu Bias führen, kann das Resultat der Betrachtung 1) der Unterschiede zwischen den Anteilen fehlender Ergebnisdaten in den Interventions- und Vergleichsgruppen und 2) der Gründe für das Fehlen von Ergebnisdaten sein.

Einige Teilnehmer*innen können aus anderen Gründen als fehlenden Endpunktdaten von einer Analyse ausgeschlossen werden. Insbesondere eine naive Per-Protocol-Analyse ist auf die Teilnehmer*innen beschränkt, die ausschließlich die beabsichtigte Intervention erhalten haben. Ein möglicher Bias durch solche Analysen oder durch andere Ausschlüsse von Teilnehmer*innen, für die Endpunktdaten verfügbar sind, wird im RoB 2 Tool im Kontext von „Bias durch Abweichungen von den vorgesehenen Interventionen“ adressiert.

5.2.4 Bias durch die Ergebnismessung

Fehler bei der Messung bzw. Erhebung von Ergebnissen („bias in measurement of the outcome“) können Schätzungen von Interventionseffekten verzerren. Solche Fehler werden häufig als **Messfehler („measurement error“;** bei kontinuierlich erhobenen Ergebnissen), **Fehlklassifikation („misclassification“;** bei dichotom/kategorisch erhobenen Ergebnissen) oder **Unter-/Übererfüllung („under-ascertainment/over-ascertainment“;** bei Ereignissen) bezeichnet. Fehler bei der Erhebung von Ergebnissen können in Bezug auf die Zuteilung zur Intervention differentiell oder nicht-differentiell sein.

- **Differentielle Fehler** hängen mit der Zuteilung zur Intervention zusammen. Sie unterscheiden sich systematisch zwischen den Interventionsgruppen. Sie sind weniger wahrscheinlich, wenn die das Ergebnis ermittelnden Personen bezüglich der Zuteilung zur Intervention verblindet sind.
- **Nicht-differentielle Fehler** sind unabhängig von der Zuweisung zur Intervention.

Die Bewertung des Biasrisikos durch die Ergebnismessung kann folgende Überlegungen beinhalten, die sich primär auf differentielle Fehler beziehen (und im RoB 2 Tool adressiert werden): 1) ob die Methode der Erhebung des Ergebnisses angemessen ist, 2) ob sich die Erhebung bzw. Ermittlung des Ergebnisses zwischen den Interventionsgruppen unterscheidet oder unterscheiden könnte, 3) wer die das Ergebnis ermittelnde Person ist, 4) ob die das Ergebnis ermittelnde Person in Bezug auf die Zuteilung zur Intervention verblindet ist, und 5) ob die Ermittlung des Ergebnisses wahrscheinlich durch Kenntnis der erhaltenen Intervention beeinflusst ist.

5.2.5 Bias durch Selektion des berichteten Ergebnisses

Bias durch Selektion des berichteten Ergebnisses („bias in selection of the reported result“) bezieht sich auf Verzerrungen, die dadurch entstehen, dass Studienergebnisse aufgrund ihrer Richtung, Größe oder statistischen Signifikanz gezielt ausgewählt und berichtet wurden. Dabei ist zwischen den folgenden Aspekten zu unterscheiden:

- Die Endpunktdomäne: Dies ist ein Status oder interessierender Endpunkt, unabhängig davon, wie er gemessen wurde (z.B. der Schweregrad einer Erkrankung)
- Eine spezifische Endpunkt- bzw. Ergebnismessung (z.B. die Messung der Domäne Schweregrad der Erkrankung mit einem geeigneten Messinstrument 12 Wochen nach Interventionsstart)
- Die Ergebnisanalyse: Dies ist ein spezifisches Ergebnis, das durch das Analysieren einer oder mehrerer Ergebnismessungen gewonnen wurde (z.B. die Mittelwertdifferenz der Veränderung in der Messung des Schweregrads der Erkrankung zwischen Baseline und Woche 12 zwischen der Interventions- und Vergleichsgruppe)

Bias durch Selektion des berichteten Ergebnisses kommt üblicherweise durch den Wunsch nach Ergebnissen zustande, die persönliche bzw. eigennützige Interessen stützen oder ausreichend „beachtenswert“ sind, um publikationswürdig zu sein. So könnte es zum Beispiel sein, dass die Autor*innen einer Studie zum Vergleich einer Intervention mit einer Vergleichsintervention eine vorgefasste Meinung oder ein persönliches Interesse daran haben, zu zeigen, dass die experimentelle Intervention nützlich und sicher ist. Aus diesem Grund könnten sie geneigt sein, Ergebnisse zu berichten, die statistisch signifikant und zugunsten der experimentellen Intervention sind. Umgekehrt könnte es sein, dass Studienautor*innen selektiv Schätzungen schädlicher Effekte berichten, die statistisch signifikant und nachteilig für die experimentelle Intervention sind, wenn sie glauben, dass die Publikation des Vorliegens schädlicher Effekte die Chancen einer Publikation der Studie in einem hochrangigen Journal erhöhen.

Die Bewertung des Biasrisikos durch Selektion des berichteten Ergebnisses kann folgende Überlegungen beinhalten (die im RoB 2 Tool adressiert werden): 1) Ob die Studie in Übereinstimmung mit einem vorab spezifizierten Plan analysiert wurde, der abgeschlossen wurde, bevor nicht-verblindete Ergebnisdaten für die Analyse verfügbar waren, 2) das selektive Berichten einer bestimmten Ergebnismessung (basierend auf den Ergebnissen) aus mehreren Ergebnismessungen innerhalb einer Ergebnisdomäne, und 3) das selektive Berichten einer bestimmten Analyse (basierend auf den Ergebnissen) aus mehreren Analysen zur Schätzung des

Interventionseffektes. Jegliche Form der selektiven Berichterstattung führt zu Bias, wenn die Selektion auf der Richtung, Größe oder statistischen Signifikanz des geschätzten Effektes basiert.

Um Bias durch selektives Berichten von Ergebnissen bewerten zu können, ist es notwendig, zu prüfen, ob die aus einer Studie berichteten Ergebnisse der präspezifizierten Planung entsprechen. Dies erfordert in der Regel einen Abgleich mit dem zugehörigen Studienregistereintrag, dem Studienprotokoll oder einer früheren Publikation zur Studie bzw. zu ihrem Design.

Ergänzung: ROB-ME Tool zur Bewertung des Biasrisikos aufgrund der Nicht-Verfügbarkeit eines Studienergebnisses für den Einschluss in eine Synthese

Im Oktober 2020 wurde von den Entwickler*innen des RoB 2 Tools eine erste vorläufige Version eines separaten neuen „Bewertungsinstruments für die Bewertung des Biasrisikos aufgrund von fehlender Evidenz in einer Synthese“ publiziert, des ROB-ME Tools[41] („A tool for assessing Risk Of Bias due to Missing Evidence in a synthesis“). In Abgrenzung zur Bewertung des Biasrisikos durch Selektion des berichteten Ergebnisses wurde ROB-ME dazu entwickelt, das Biasrisiko durch Nicht-Verfügbarkeit eines Studienergebnisses für den Einschluss in eine Synthese (paarweise Meta-Analyse) in einer Evidenzsynthese zu bewerten.[42] Beide Bewertungen – Bias durch Selektion des berichteten Ergebnisses und Bias aufgrund des Nicht-Berichtens durch „fehlende Evidenz in einer Synthese“ – erfordern die Betrachtung derselben Informationsquellen (z.B. Studienprotokolle, Registrierungseinträge); daher wird empfohlen, ROB-ME parallel zu ROB 2 (oder ROBINS-I) anzuwenden.[42] Zu der vorläufigen Version sind auf den ROB-ME-Seiten[41] verschiedene Ressourcen verfügbar.

5.3. Die Bewertung des Biasrisikos in randomisierten Interventionsstudien mit dem RoB 2 Tool

5.3.1 Hintergrund

Das **Cochrane Tool zur Bewertung des Biasrisikos von randomisierten Studien** - „The Cochrane Collaboration’s tool for assessing risk of bias in randomised trials“ - (kurz „RoB Tool“) wurde in den Jahren 2005 bis 2007 von einer aus Methodiker*innen, Editor*innen und Review-Autor*innen bestehenden Arbeitsgruppe entwickelt, und wurde 2008 erstmalig publiziert[7]. Seither ist es weit verbreitet in Cochrane Reviews und anderen systematischen Reviews angewendet worden; in Cochrane Reviews ist seine Verwendung zur Bewertung des Biasrisikos von randomisierten Studien verpflichtend. Die Erfahrungen mit dem RoB Tool innerhalb des Jahrzehnts nach seiner initialen Publikation haben viele Stärken, aber auch einige Schwächen, gezeigt und zur Identifizierung

potentieller Verbesserungen geführt. So ergab ein 2016 publizierter **Review über publizierte Kommentare und die Nutzung des Tools43**, dass als wesentliche Stärken des Tools sein Ansatz (Bewertung der Studiendurchführung, nicht der Berichterstattung), seine Entwicklungsbasis (breite Konsultation, eine Basis aus empirischer und theoretischer Evidenz) sowie die transparente Vorgehensweise betrachtet werden. Die im Rahmen dieses Reviews berichtete Untersuchung einer Stichprobe von jeweils 100 (2014 publizierten) Cochrane Reviews und anderen („non-Cochrane“) systematischen Reviews ergab, dass das Tool in allen Cochrane Reviews (100/100) verwendet wurde und das bevorzugte Tool in den non-Cochrane Reviews (31/100) war. Es zeigte sich jedoch, dass das Tool sowohl in den Cochrane Reviews als auch in den non-Cochrane Reviews häufig nicht den Empfehlungen entsprechend implementiert wurde. Als weitere Schwächen wurden unter anderem die Auswahl der Haupt-Bias-Domänen, Schwierigkeiten in Bezug auf die Implementierung (eine moderate Inter-Rater-Übereinstimmung) und die Terminologie identifiziert.

Das RoB Tool wurde folglich in einem mehrjährigen Prozess überarbeitet, der denselben Prinzipien folgte, die der Entwicklung des ursprünglichen (originalen) RoB Tools und des ROBINS-I Tools (zur Bewertung des Biasrisikos in nicht-randomisierte Interventionsstudien, s. Kapitel 6) zugrunde lag. Die **neue Version** wurde 2019 unter dem Namen **RoB 2** publiziert.[37] Entsprechend wird in diesem aktualisierten Manual das RoB 2 Tool (kurz „RoB 2“) vorgestellt. Autor*innen von Cochrane Reviews sind angehalten, in neuen Reviews das RoB 2 Tool zu verwenden. Ende Oktober 2020 wurde der erste Cochrane Review publiziert, in dem das RoB 2 Tool angewandt wurde.[44] Für die Aktualisierung von (Cochrane) Reviews, die vor 2019 publiziert wurden und in denen entsprechend das initiale RoB Tool verwendet wurde, bietet das Cochrane Handbook (Kapitel 8 der Online-Version) **Hilfestellung für Entscheidungen hinsichtlich einer Änderung der Methodik bei Aktualisierung des Reviews**.[40]

Die (externe) Validierung dieses neuen Bewertungsinstruments steht noch aus.

Die in diesem Manual dargestellten Informationen zum RoB 2 Tool basieren im Wesentlichen (und wo nicht explizit/ anders ausgewiesen) auf den folgenden zu dem Tool verfügbaren Ressourcen. Die vollständigen bibliographischen Angaben finden sich in der Literaturliste am Ende des Manuals:

- Sterne et al. 2019: „**RoB 2: A revised tool for assessing risk of bias in randomised trials**“[37]
- Higgins et al. 2019: „**Revised Cochrane risk-of-bias tool for randomized trials (RoB 2): full guidance document**“ (Version 22. August 2019)[39]
- Higgins et al. 2019: „**Assessing risk of bias in a randomized trial**“ - **Chapter 8, Cochrane Handbook for Systematic Reviews of Interventions**[40]

Das RoB 2 Tool und die zu ihm verfügbaren Ressourcen einschließlich Hinweise auf relevante Publikationen in Fachzeitschriften sind auf einer **eigenen Website**[45] verfügbar. Dort findet sich auch ein „**Cribsheet**“46, eine Vorlage des Tools, die zusammenfassende Erläuterungen zu allen Domänen enthält, sowie eine leere Vorlage des Tools, ein **RoB 2 template**[47], für eigene Bewertungen. Zusätzlich ist das Tool als **Exceltabelle**[48] verfügbar.

Weitere Informationen und Ressourcen zum RoB 2 Tool finden sich auf den Webseiten von Cochrane Methods.[49]

Was unterscheidet das RoB 2 Tool vom ursprünglichen RoB Tool?

In Abbildung 1 sind die wesentlichen Aspekte dargestellt, in denen sich das RoB 2 Tools vom ursprünglichen RoB Tool unterscheidet.[40(p225)]

- Die Bewertung des Biasrisikos erfolgt in RoB 2 bezogen auf ein einzelnes Ergebnis statt auf eine Studie oder einen Endpunkt.
- Die Namen der Bias-Domänen beschreiben eindeutiger, auf welche Aspekte sie abzielen. Dies soll das Potential für Verwirrung durch Begriffe, die in unterschiedlicher Weise verwendet werden oder nicht geläufig sind (z.B. ‚Selection Bias‘ und ‚Performance Bias‘), verringern.
- Es wurden „Signalfragen“ („signalling questions“) eingeführt, zusammen mit Algorithmen, die Autor*innen bzw. Gutachter*innen dabei helfen sollen, zu einer Bewertung des Biasrisikos für die einzelnen Domänen zu gelangen.
- Es wurde eine Differenzierung zwischen der Betrachtung des Effektes der Zuteilung zur Intervention und des Effektes der Einhaltung der Intervention eingeführt; dies hat Auswirkungen auf die Bewertung des „Bias aufgrund von Abweichungen von den vorgesehenen Interventionen“.
- Die Bewertung des Biasrisikos aufgrund des Ausschlusses von Teilnehmer*innen von der Analyse (z.B. als Teil einer naiven ‚per protocol‘-Analyse) erfolgt jetzt in der Domäne „Bias aufgrund von Abweichungen von der vorgesehenen Intervention“ anstelle der Domäne „Bias aufgrund von fehlenden Ergebnisdaten“.
- Das Konzept der selektiven Berichterstattung eines Ergebnisses wird von dem des selektiven Nicht-Berichtens eines Ergebnisses unterschieden, wobei das letztere Konzept aus dem Tool entfernt wurde, sodass es - methodisch angemessener - auf Ebene der Synthese adressiert werden kann.
- Die Option, neue Domänen hinzuzufügen, ist entfernt worden.
- Es wurde ein expliziter Prozess zum Erlangen eines Urteils über das Gesamt-Biasrisiko des Ergebnisses eingeführt.

Abb. 1: Unterschiede zwischen dem ursprünglichen RoB Tool und RoB 2[40 (S.225; übersetzt)]

5.3.2 Welche Studienarten können mit RoB 2 bewertet werden?

Das RoB 2 Tool wurde für die **Bewertung des Biasrisikos von randomisierten klinischen Studien** entwickelt. Grundsätzlich kann das Tool sowohl für individuell randomisierte Studien mit parallelen Gruppen als auch für Cluster-randomisierte Studien mit parallelen Gruppen oder individuell randomisierte Cross-Over-Studien verwendet werden. Spezifische Varianten, die den Unterschied

zwischen diesen Studienarten abbilden, sind in Kapitel 23 des Cochrane Handbook als "**Varianten des RoB 2 Tools für cluster-randomisierte Studien und Cross-Over-Studien**"[50] beschrieben. Auf der RoB 2-Website sind diese beiden Varianten - vorläufig als Testversionen - verfügbar. Der in diesem Manual (und in den zugrundeliegenden aktuell zu RoB 2 verfügbaren Publikationen) dargestellte Leitfaden für die Anwendung des RoB 2 Tools bezieht sich primär auf individuell randomisierte kontrollierte Studien mit parallelen Gruppen.

5.3.3 Wie ist RoB 2 aufgebaut?

RoB 2 ist ein **domänenbasiertes Tool**. Es ist in die folgenden **fünf Domänen** unterteilt, die sich auf unterschiedliche Bias-Aspekte beziehen. Die Domänen umfassen alle wesentlichen Biasformen, die nach derzeitigem Wissensstand bzw. Verständnis einen Einfluss auf die Ergebnisse randomisierter Studien haben können (s. auch Abschnitt 5.2):

1) Bias durch den Randomisierungsprozess

2) Bias durch Abweichungen von den vorgesehenen Interventionen

3) Bias durch fehlende Ergebnisdaten

4) Bias durch die Ergebnismessung

5) Bias durch Selektion des berichteten Ergebnisses

Zu jeder Domäne wurden **Signalfragen („signalling questions")** formuliert, deren Beantwortung die Bewertung des Biasrisikos für die Domäne erleichtern sollen. Als weitere Hilfestellung wurden **Algorithmen** erstellt, die die Antworten auf die Signalfragen mit Vorschlägen für die daraus resultierende Bewertung des Biasrisikos verknüpfen. Aus den Bewertungen der einzelnen Domänen wird eine **Gesamtbewertung des Biasrisikos** für die Studie, bezogen auf das interessierende Ergebnis, abgeleitet. Die Bewertung erfolgt „ergebnisbezogen", das heißt, bezogen auf ein spezifisches individuelles Ergebnis, und damit nicht bezogen auf einen Endpunkt, für den es mehrere Ergebnisse geben kann, oder auf eine ganze Studie. Entsprechend erfordert die Bewertung des Biasrisikos für mehrere Ergebnisse die Durchführung mehrerer Bewertungen. Sollen mehrere Ergebnisse bewertet werden, deren Bewertung als identisch betrachtet werden kann, kann die Bewertung auch für eine Ergebnisgruppe (ggf. einen Endpunkt) erfolgen.

5.3.4 Wie wird die Bewertung mit RoB 2 durchgeführt?

Vorabspezifizierung

Der eigentlichen Bewertung des Biasrisikos geht eine konkretisierende Vorauswahl in Bezug auf die zu bewertende Studie voran, die entsprechend dokumentiert werden. Abbildung 2 zeigt einen exemplarischen Ausschnitt aus dem originalen RoB 2 Tool[47] zu diesem Abschnitt.

Studiendesign

- ✞ Individuell randomisierte Parallelgruppen-Studie
- ☌ Cluster- randomisierte Parallelgruppen-Studie
- ☌ Individuell randomisierte Cross-Over- (oder anders gematchte) Studie

Für diese Bewertung werden die zu vergleichenden Interventionen definiert als

Experimentell: Vergleich:

Spezifizieren Sie den Endpunkt für die Bewertung des Biasrisikos

Spezifizieren Sie das zu bewertende numerische Ergebnis. Falls mehrere alternative Analysen dargestellt sind, spezifizieren Sie das numerische Ergebnis (z.B. RR = 1.52 (95% CI 0.83; 2.77) und oder einen Verweis (z.B. auf eine Tabelle, eine Abbildung oder einen Absatz), das das zu bewertende Ergebnis eindeutig definiert.

Ist das Ziel des Review-Teams für dieses Ergebnis…?

- ☌ die Bewertung des Effekts der *Zuteilung zur Intervention* ('Intention-to-treat'-Effekt)
- ☌ die Bewertung des Effekts der *Einhaltung der Intervention* ('Per-protocol'-Effekt)

Wenn das Ziel die Bewertung des Effekts der Einhaltung der Intervention ist, wählen Sie die Abweichungen von der vorgesehenen Intervention aus, die betrachtet werden sollen (es ist mindestens eine Option anzukreuzen):

- ☌ Das Vorkommen von im Protokoll nicht vorgesehenen Interventionen (Nicht-Protokoll-Interventionen)
- ☌ Fehler bei der Implementierung der Intervention, die das Ergebnis beeinflusst haben könnten
- ☌ Die Nicht-Einhaltung der zugeteilten Intervention durch Studienteilnehmer

Abb. 2: Vorabspezifizierung - Ausschnitt aus dem RoB 2 Tool[47] (übersetzt und modifiziert)

Zu den zu spezifizierenden Aspekten zählen:

- Die **Festlegung des Studiendesigns** (individuell randomisierte Studie mit parallelen Gruppen, Cluster-randomisierte Studie mit parallelen Gruppen oder individuell randomisierte Cross-Over- oder anderweitig „gematchte“ Studie)
- Der **interessierende Vergleich** (Intervention und Kontrolle)
- Der zu **bewertende Endpunkt** und das zu **bewertende Ergebnis**
- Das zu **bewertende Ziel der Studie** bezogen auf den interessierenden Effekt
- Die **verwendeten (Daten-) Quellen** (z.B. publizierter Bericht, Studienprotokoll)

Das zu bewertende Ergebnis spezifizieren

Vor Beginn der Bewertung des Biasrisikos ist festzulegen, welches spezifische (numerische) Ergebnis der jeweiligen Studie bewertet werden soll. Wichtig ist, dass dieses Ergebnis nicht auf Grundlage des wahrscheinlichen bzw. erwarteten Ergebnisses der Bewertung des Biasrisikos ausgewählt wird. Da Studien häufig viele Ergebnisse liefern, ist die Bewertung aller Ergebnisse aller eingeschlossenen Studien in einem systematischen Review nicht immer durchführbar. Es kann daher ggf. angemessen sein, die Anzahl der in einem Review zu bewertenden Ergebnisse durch Beschränkung auf die wichtigsten (wie im Falle von Cochrane Reviews die in der „Summary of Findings“ Tabelle dargestellten) Endpunkte in einem durchführbaren Rahmen zu halten.

Den zu bewertenden Effekt spezifizieren

Die Bewertung der RoB 2-Domäne „Bias aufgrund von Abweichungen von den vorgesehenen Interventionen“ unterscheidet sich abhängig vom interessierenden Effekt. Entsprechend ist vor der Bewertung festzulegen, welcher Effekt bewertet werden soll:

- Der **Effekt der Zuteilung zu den Interventionen zu Beginn der** Studie („at baseline“), unabhängig davon, ob die Interventionen wie vorgesehen erfolgt sind („Intention-to-treat“-Effekt), oder
- Der **Effekt der Einhaltung der Interventionen gemäß ihrer Spezifizierung im Studienprotokoll („Per Protocol“-Effekt)**

Wenn Teilnehmer*innen einer Studie die für sie vorgesehene Intervention nicht erhalten oder nach Studienbeginn von der ihnen zugeteilten Intervention abweichen, führt dies zu unterschiedlichen Effekten, die beide von Interesse sein können. Der **Effekt der Zuteilung zur Intervention** ist zum Beispiel der am besten geeignete, wenn es um die Beantwortung einer gesundheitspolitischen Fragestellung dazu geht, ob eine Intervention in einem bestimmten Gesundheitssystem bzw. Kontext empfohlen werden soll - zum Beispiel, ob ein bestimmtes Screening-Programm veranlasst werden oder ein neues blutdruckregulierendes Medikament grundsätzlich verordnungsfähig sein soll. Der Effekt der Zuteilung zur Intervention sollte durch eine Intention-to-Treat-Analyse (ITT-Analyse) geschätzt werden, bei der alle randomisierten Teilnehmer*innen in die Analyse eingeschlossen werden und alle Teilnehmer*innen in den Gruppen analysiert werden, zu denen sie randomisiert wurden - unabhängig davon, welche Intervention sie tatsächlich erhalten haben. Durch eine ITT-Analyse werden die Vorteile der Randomisierung bewahrt, das heißt, dass sich die Interventionsgruppen im Mittel zu Studienbeginn hinsichtlich bekannter oder nicht bekannter prognostischer Faktoren nicht (relevant) unterscheiden.

Der **Effekt der Einhaltung der Intervention** gemäß ihrer Spezifizierung im Studienprotokoll hingegen wäre am besten geeignet, wenn es um eine Versorgungsentscheidung bei einer/einem einzelnen Patient*in geht – zum Beispiel, ob diese/r sich dem Screening unterziehen oder das neue Medikament einnehmen soll. Zwei in randomisierten Studien weit verbreitete Ansätze für die Schätzung des Per-Protocol-Effekts können zu schwerwiegendem Bias führen: As-Treated-Analysen, in denen die Teilnehmer*innen entsprechend der Intervention analysiert werden, die sie tatsächlich erhalten haben, auch wenn ihre randomisierte Zuteilung zu einer anderen Intervention erfolgte, und naive Per-Protocol-Analysen, die sich auf die Teilnehmer*innen beschränken, die die ihnen zugeteilten Interventionen eingehalten haben. Beide Analyseansätze sind problematisch, weil prognostische Faktoren möglicherweise einen Einfluss darauf hatten, ob die Teilnehmer*innen die ihnen zugeteilte Intervention eingehalten haben. Für angemessene Analyseansätze zur Schätzung des Effekts der Einhaltung der Intervention wird im Cochrane Handbook[40] auf Literatur von Hernán and Robins 2017[51] verwiesen.

Bias-Domänen in RoB 2

Tabelle **3** zeigt eine Übersicht über **die in RoB 2 enthaltenen Bias-Domänen** und die Aspekte, die sie adressieren. [40 Tab. 8.2a]. Es sind alle Domänen zu bewerten, und es sollten keine weiteren Domänen hinzugefügt werden.

Tab. 3: Übersicht über die Bias-Domänen in RoB 2[40] (S. 209, Tabelle 8.2.a, übersetzt und modifiziert)

Bias-Domäne	In der Domäne adressierte Aspekte*
Bias durch den Randomisierungsprozess („Bias arising from the randomization process“)	• War die Zuteilungssequenz randomisiert? • Wurde die Zuteilungssequenz angemessen verdeckt durchgeführt? • Gibt es Unterschiede zwischen den Gruppen bei Studienbeginn („baseline differences“), die auf ein Problem mit dem Randomisierungsprozess hinweisen?
Bias durch Abweichungen von den vorgesehenen Interventionen („Bias due to deviations from intended interventions”)	• Wussten die Teilnehmer*innen während der Studie, welcher Intervention sie zugeteilt waren? • Wussten die Personen, die die Teilnehmer*innen betreuten und die Interventionen verabreichten, welcher Intervention die Teilnehmer*innen zugeteilt waren?
	Wenn der interessierende Effekt der Effekt der Zuteilung zur Intervention ist: • (sofern relevant) Gab es Abweichungen von den vorgesehenen Interventionen aufgrund des experimentellen Kontextes (das heißt, die nicht die übliche Praxis widerspiegeln)? Wenn ja, waren sie unausgewogen zwischen den Gruppen und ist es wahrscheinlich, dass sie das Ergebnis beeinflusst haben? • Wurde eine angemessene Analyse zur Schätzung des Effekts der Zuteilung zur Intervention angewendet? Wenn nicht, gibt es ein Potential für einen substantiellen Einfluss des Vorgehens auf das Ergebnis?

	Wenn der interessierende Effekt der Effekt der Einhaltung der Intervention ist: • (sofern relevant) Waren wichtige Nicht-Protokoll-Interventionen (nicht im Protokoll spezifizierte Interventionen) ausgewogen zwischen den Gruppen? • (sofern relevant) Könnten Mängel bei der Implementierung der Intervention das Ergebnis beeinflusst haben? • (sofern relevant) Haben die Teilnehmer*innen das ihnen zugeteilte Interventions-Regime eingehalten? • (sofern relevant) Wurde eine angemessene Analyse zur Schätzung des Effekts der Einhaltung der Intervention durchgeführt?
Bias durch fehlende Ergebnisdaten („Bias due to missing outcome data")	• Waren für das interessierende Ergebnis Daten für alle, bzw. fast alle, randomisierten Teilnehmer*innen verfügbar? • (sofern relevant) Gibt es Evidenz dafür, dass das Ergebnis nicht durch fehlende Ergebnisdaten verzerrt wurde? • (sofern relevant) Steht das Fehlen von Ergebnisdaten wahrscheinlich in einem Zusammenhang mit dem wahren Wert der fehlenden Ergebnisdaten (z.B., wenn sich die Anteile fehlender Ergebnisdaten oder die Gründe für das Fehlen zwischen den Gruppen unterscheiden
Bias durch die Ergebnismessung („Bias in measurement of the outcome")	• War die Methode der Ergebnismessung unangemessen? • Könnte sich die Messung bzw. Ermittlung des Ergebnisses zwischen den Gruppen unterschieden haben? • Wussten die das Ergebnis erhebenden Personen, welche Intervention die Teilnehmer*innen erhalten hatten? • (sofern relevant) Ist es wahrscheinlich, dass die Erhebung des Ergebnisses durch das Wissen, welche Intervention die Teilnehmer*innen erhalten hatten, beeinflusst wurde?
Bias durch Selektion des berichteten Ergebnisses („Bias in selection of the reported result")	• Wurde die Studie in Übereinstimmung mit einem vorab spezifizierten Plan analysiert, der finalisiert wurde, bevor nicht verblindete Ergebnisdaten für die Analyse verfügbar waren? • Ist es wahrscheinlich, dass das zu bewertende numerische Ergebnis, auf Grundlage der Ergebnisse, aus mehreren Ergebnismessungen innerhalb der Endpunktdomäne ausgewählt wurde? • Ist es wahrscheinlich, dass das zu bewertende numerische Ergebnis, auf Grundlage der Ergebnisse, aus mehreren Analysen der Daten ausgewählt wurde?

*Pragmatische Teil-Übersetzung; die vollständigen, präzisen Formulierungen der Signalfragen und Hilfestellungen für ihre Beantwortung finden sich im originalen vollständigen RoB 2 Tool[47] (bzw. dem ‚Crib Sheet' [46]).

Abbildung 3 zeigt exemplarisch Domäne 1 aus dem ROB 2 Tool[47]. Jede RoB 2-Domäne umfasst die folgenden Elemente:

- Mehrere **Signalfragen**;
- Ein **Urteil über das Biasrisiko für die Domäne**, mit Hilfestellung durch einen **Algorithmus**, der die Antworten auf die Signalfragen mit einem Vorschlag für die Bewertung verknüpft;
- **Freie Textfelder** zum Eintragen von Begründungen für die Antworten auf die Signalfragen und die Bewertungen; und
- Die **optionale Möglichkeit einer Vorhersage** (und Erklärung) **der wahrscheinlichen bzw. erwarteten Richtung des Bias**.

Bewertung des Biasrisikos

Grün unterstrichene Antworten sind potenzielle Hinweise auf ein geringes Biasrisiko und Antworten in Rot sind potenzielle Hinweise auf ein Biasrisko. Bei Fragen, die sich nur auf Verweise zu weiteren Fragen beziehen, wird keine Formatierung verwendet.

Domäne 1: Biasrisiko durch den Randomisierungsprozess

Signalfragen	Kommentare	Antwortoptionen*
1.1 War die Zuteilungssequenz zufällig?		Y/ PY / PN / N / NI
1.2 War die Zuteilungssequenz verborgen, bis die Teilnehmer in die Studie eingeschlossen und den Interventionen zugeteilt waren?		Y/ PJ / PN / N / NI
1.3 Weisen Baseline-Unterschiede zwischen den Interventionsgrippen auf ein Problem mit dem Randomisierungsprozess hin?		Y/ PY / PN / N / NI
Beurteilung des Biasrisikos		Niedrig/hoch/einige Bedenken
Optional: Was ist die prognostizierte Richtung des Bias durch den Randomisierungsprozess?		NA/begünstigt experimentelle Intervention/begünstigt Vergleich/in Richtung Null/weg von der Null

Abb. 3: Domäne 1, RoB 2 Tool [47] (übersetzt und adaptiert)

*Antwortoptionen in englischer Sprache belassen: Y = Yes (Ja), PY = Probably Yes (wahrscheinlich ja), PN = Probably No (wahrscheinlich nein), N = No (nein); NA = Not applicable (nicht anwendbar bzw. nicht relevant)

Signalfragen

Die **Signalfragen** bieten einen strukturierten Ansatz für die Berücksichtigung relevanter Informationen für die Bewertung des Biasrisikos.

Die Antwortoptionen sind:

- **Ja** („Yes“ → „Y“)
- **Wahrscheinlich ja** („Probably yes“ → „PY“)
- **Wahrscheinlich nein** („Probably no“ → „PN“)
- **Nein** („No“ → „N“)
- **Keine Informationen** („No information“ → „NI“)

Um die Signalfragen so einfach und klar wie möglich zu halten, sind diese so formuliert, dass, abhängig von der natürlichsten Art, die jeweilige Frage zu stellen, ein „Ja“ entweder auf ein niedriges oder ein hohes Biasrisiko hinweist. Die farbliche Gestaltung der Antwortoptionen bietet hier eine zusätzliche Hilfestellung: **grün unterstrichene Antworten** (z.B. Y/PY) kennzeichnen ein potenziell niedriges Biasrisiko, **rot geschriebene Antworten** ein potentielles Biasrisiko (z.B. PN/N). Die Antworten „Ja“ und „Wahrscheinlich ja“ wirken sich in gleicher Weise auf das Biasrisiko aus; gleiches gilt für die Antworten „Nein“ und „Wahrscheinlich nein“. Die definitiven Antworten „Ja“ und „Nein“ implizieren typischerweise die Verfügbarkeit von robuster Evidenz in Bezug auf die jeweilige

Signalfrage, während die „Wahrscheinlich"-Antworten typischerweise implizieren, dass diese (in Abwesenheit von robuster Evidenz) auf einer Einschätzung der Gutachter*innen basieren. Die **Antwort „Keine Informationen" („no information")** sollte nur dann verwendet werden, wenn (1) unzureichende Informationen für ein „Ja", „Wahrscheinlich ja", „Nein" oder „Wahrscheinlich nein" verfügbar sind und 2) die Antwort „Wahrscheinlich ja" oder „Wahrscheinlich nein" im Kontext der Studie bei Nichtverfügbarkeit dieser Informationen unangemessen ist. Wenn zum Beispiel im Kontext einer großen klinischen Studie, die von einem erfahrenen Studienzentrum organisiert und durchgeführt wurde, in einem Bericht in einem Journal mit einer strikt durchgesetzten Wortzahlbegrenzung spezifische Informationen zu den Randomisierungsmethoden fehlen, kann es bei der Beantwortung der Signalfrage zur verdeckten Zuteilungssequenz angemessen sein, die Antwort „Wahrscheinlich Ja" anstelle der Antwort „Keine Informationen" zu wählen.

Die Bedeutung der Antwort „Keine Informationen" auf eine Signalfrage ist, in Abhängigkeit vom Zweck der Frage, unterschiedlich: Wenn die Frage bezweckt, Evidenz für ein Problem zu identifizieren, entspricht die Antwort „Keine Information" der Nichtverfügbarkeit von Evidenz für das Problem. Wenn sich die Frage auf einen Aspekt bezieht, bei dem erwartet wird, dass er berichtet ist (zum Beispiel, ob Teilnehmer*innen vorzeitig aus der Studie ausgeschieden sind), führt das Fehlen der entsprechenden Informationen zu Bedenken, dass hier ein Problem vorliegen könnte.

Für Signalfragen, deren Relevanz von der Beantwortung einer vorausgehenden Frage abhängt, gibt es die Antwortoption „**Nicht relevant" (**oder **„Nicht zutreffend", „not applicable")**

Die Signalfragen sollten grundsätzlich unabhängig voneinander beantwortet werden, ***Urteil über das Biasrisiko für eine Domäne***

Auf die Beantwortung der Signalfragen folgt die Bewertung des Biasrisikos für die jeweilige Domäne, für die es die folgenden drei Optionen gibt:

- **Niedriges Biasrisiko** → **„low risk"**
- **Einige Bedenken** → **„some concerns"**
- **Hohes Biasrisiko** → **„high risk"**

Die **Verknüpfung der Signalfragen mit den Algorithmen** soll die Biasbewertung erleichtern. Abbildung 4 zeigt einen exemplarischen Algorithmus aus dem **RoB 2 template**[47] Die Algorithmen umfassen dabei die Verknüpfung jeglicher möglicher Kombinationen von Antworten auf die Signalfragen mit dem Urteil „niedriges Biasrisiko", „einige Bedenken" oder „hohes Biasrisiko". Dem Begriff „Urteil" (Englisch „judgement") kommt im Rahmen der Bewertung des Biasrisikos eine

wichtige Bedeutung zu: die Algorithmen bieten Vorschläge für die Einschätzung, die von den bewertenden Personen verifiziert und ggf. verändert werden sollten. Für die finale Einschätzung sollte das „Biasrisiko“ im Sinne eines Risikos für „gewichtigen Bias“ („material bias“) interpretiert werden. Dies bedeutet, dass Bedenken nur dann Berücksichtigung finden sollten, wenn es wahrscheinlich ist, dass sie einen Einfluss auf die Formulierung verlässlicher Schlussfolgerungen haben.

Die **freien Textfelder** bieten Raum für den Eintrag von unterstützenden Informationen und Begründungen für Antworten und Einschätzungen.

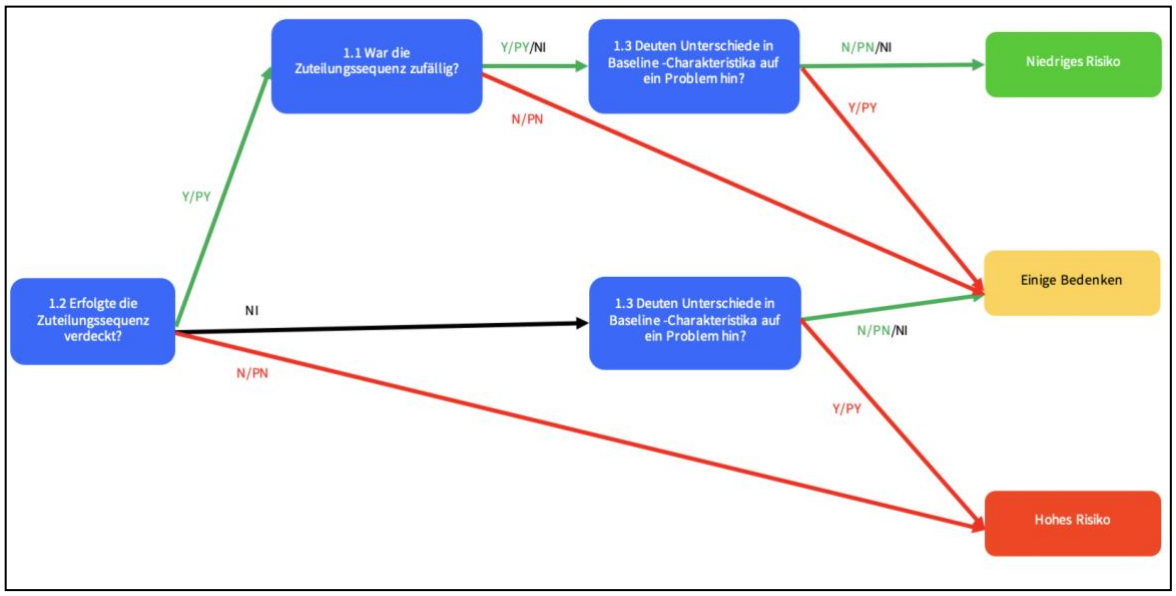

Abb. 4: Algorithmus (Vorschlag) für Domäne 1 des RoB 2 Tools[47] **(übersetzt und adaptiert)**

Einschätzung der prognostizierten Richtung des Bias

Am Ende jeder RoB 2-Domäne und für die Gesamtbewertung gibt es die Möglichkeit, eine **Aussage über die erwartete Richtung des Bias** zu treffen. Die Kenntnis der Größe und Richtung eines identifizierten Bias ist für die Beurteilung eines Studienergebnisses wünschenswert, stellt jedoch eine weitaus größere Herausforderung als die Bewertung des Biasrisikos dar. Bei einigen Domänen ist der Bias am einfachsten als **„in Richtung Null“ („towards the null“;** das beobachtete Ergebnis liegt näher an der Null, d.h. an „keinem Effekt“, als das wahre Ergebnis) oder **„weg von der Null“ („away from the null“;** das beobachtete Ergebnis liegt weiter entfernt von der Null, d.h. von „keinem Effekt“, als der wahre Wert) einzuschätzen. So kann zum Beispiel ein hoher Anteil von Teilnehmer*innen, die von der ihnen zugeteilten Intervention zur Vergleichsintervention wechseln, den beobachteten Unterschied für das Ergebnis zwischen den Gruppen verringern und damit zu

einem geschätzten Effekt der Einhaltung der Intervention führen, der „in Richtung Null“ verzerrt ist. Bei anderen Domänen ist der Bias wahrscheinlich als Vergrößerung oder Verringerung des geschätzten Effektes zugunsten der experimentellen Intervention oder des Vergleichs zu betrachten. Zwei Beispiele für Bias „weg von der Null“ sind die Manipulation des Randomisierungsprozesses und die selektive Berichterstattung von Ergebnissen. Angaben der Biasrichtung sollten nachvollziehbar begründet sein. Gibt es keine klare Begründung für die Beurteilung der wahrscheinlichen Richtung des Bias, sollte nicht versucht werden, eine solche auf der Basis einer Vermutung vorzunehmen.

Ein Gesamturteil über das Biasrisiko fällen

Die Bewertungen des Biasrisikos der fünf Domänen stellen die Grundlage für das **Gesamturteil über das Biasrisiko** des begutachteten Studienergebnisses dar. Die Antwort-Optionen für das Gesamturteil entsprechen dabei denen für die einzelnen Domänen („niedriges Biasrisiko“, „einige Bedenken“, „hohes Biasrisiko“). Wenn ein Ergebnis in einer einzelnen Biasdomäne als mit einem bestimmten Biasrisiko-Level behaftet bewertet wird, bedeutet dies, dass das Ergebnis insgesamt mit einem Gesamt-Biasrisiko behaftet ist, das mindestens dieser Stufe entspricht. Entsprechend sollte die Bewertung „Hohes Biasrisiko“ in einer beliebigen Domäne vergleichbare Auswirkungen auf die Gesamtbewertung haben. Für die praktische Umsetzung bedeutet das, dass die bewertenden Personen, wenn die Antworten auf die Signalfragen in einer Domäne zur (durch den Algorithmus) vorgeschlagenen Bewertung „hohes Biasrisiko“ führen, überlegen sollten, ob die ermittelten Probleme so schwerwiegend sind, dass sie dieses Gesamturteil rechtfertigen. Ist dies nicht der Fall, wäre ein angemessenes Vorgehen, sich mit einer entsprechenden Begründung über den Bewertungsvorschlag hinwegzusetzen.

Die **grundlegenden Kriterien für die Einschätzung des Gesamt-Biasrisikos** für ein spezifisches Ergebnis sind in Tabelle 4 dargestellt.[40(p212) Tab. 8.2b]

Tab. 4: Einschätzung des Gesamt-Biasrisikos[40] (S. 212, Tab. 8.2; übersetzt und modifiziert)

Gesamt-Urteil zum Biasrisiko	Kriterien
Niedriges Biasrisiko	Das Biasrisiko der Studie wurde in Bezug auf das Ergebnis in allen Domänen als niedrig eingeschätzt.
Einige Bedenken	Die Studie wurde hinsichtlich des Biasrisikos in Bezug auf das Ergebnis in mindestens einer Domäne als „mit einigen Bedenken“ behaftet eingeschätzt, das Biasrisiko wurde jedoch in keiner Domäne als „hoch“ eingeschätzt.
Hohes Biasrisiko	Das Biasrisiko der Studie wurde in Bezug auf das Ergebnis in mindestens einer Domäne als hoch eingeschätzt; oder Die Studie wurde hinsichtlich des Biasrisikos in Bezug auf das Ergebnis in mehreren Domänen als „mit einigen Bedenken“ behaftet eingeschätzt, in einer Weise, die das Vertrauen in das Ergebnis substantiell verringert.

Zur **Darstellung der Bewertungsergebnisse** mehrerer Studien, zum Beispiel im Rahmen eines systematischen Reviews, wird die textliche und/oder tabellarische Darstellung häufig durch Abbildungen ergänzt. In Cochrane Reviews wurden bislang typischerweise zwei (in Cochranes **Software RevMan**[52] generierbare) Varianten für die Darstellung der Bewertungsergebnisse mit dem Cochrane RoB Tool verwendet: ein **„risk of bias graph“**, der den Anteil der Studien mit den einzelnen Urteilen („niedriges Biasrisiko“, „einige Bedenken“, „hohes Biasrisiko“) illustriert, und ein **„risk of bias summary“,** das die Bewertungen aller Domänen und für alle Studien zeigt. Abbildung 5 zeigt einen exemplarischen „risk of bias graph“, Abbildung 6 ein „risk of bias summary“. Beide Abbildungen entstammen dem Cochrane Review „Second-generation antidepressants for preventing seasonal affective disorder in adults“ von **Gartlehner et al. 2019**.[53]

Eine graphische Darstellung der Ergebnisse der Bewertung des Biasrisikos mit dem RoB 2 Tool kann in Cochranes neuer Software RevMan Web[54] (Zugang auf Cochrane-Autor*innen und –Editor*innen beschränkt) generiert werden. Zudem steht neuerdings mit robvis[55] eine web-basierte Anwendung zur Visualisierung der Bewertungen des Biasrisikos in systematischen Reviews zur Verfügung. Abb. 7 zeigt die graphische Darstellung eines RoB 2-Bewertungsergebnisses aus dem Cochrane Review „Physical activity interventions for people with congenital heart disease” von Williams et al. 2020.[44] . Alternativ können die RoB 2-Bewertungsergebnisse auch in die Metaanalyse-Graphiken (Forest Plots) integriert werden. Ein Beispiel hierfür, aus dem Review von Williams et al. 2020, ist in Abbildung 8 dargestellt.

Figure 2. Risk of bias graph: review authors' judgements about each risk of bias item presented as percentages across all included studies.

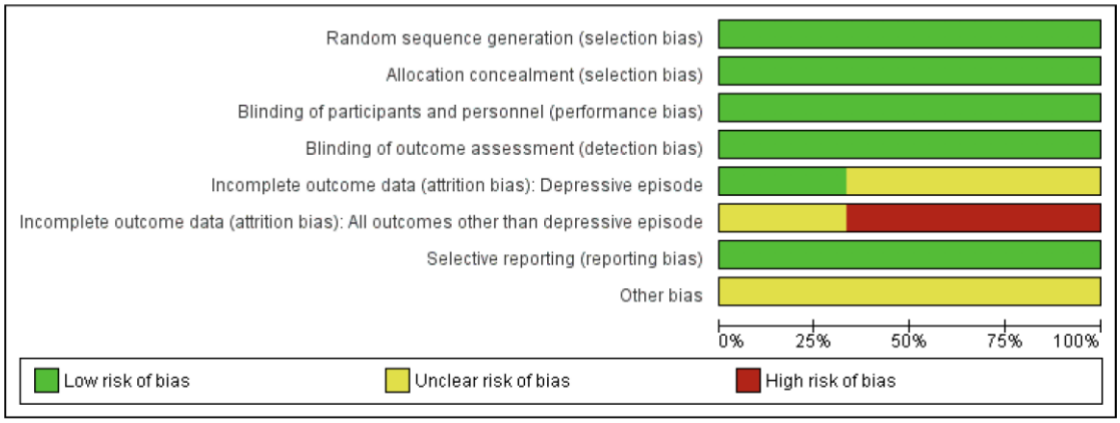

Abb. 5: Beispiel Cochrane „risk of bias graph“ zu einer Cochrane RoB Bewertung[53] (S.13)

Figure 3. Risk of bias summary: review authors' judgements about each risk of bias item for each included study.

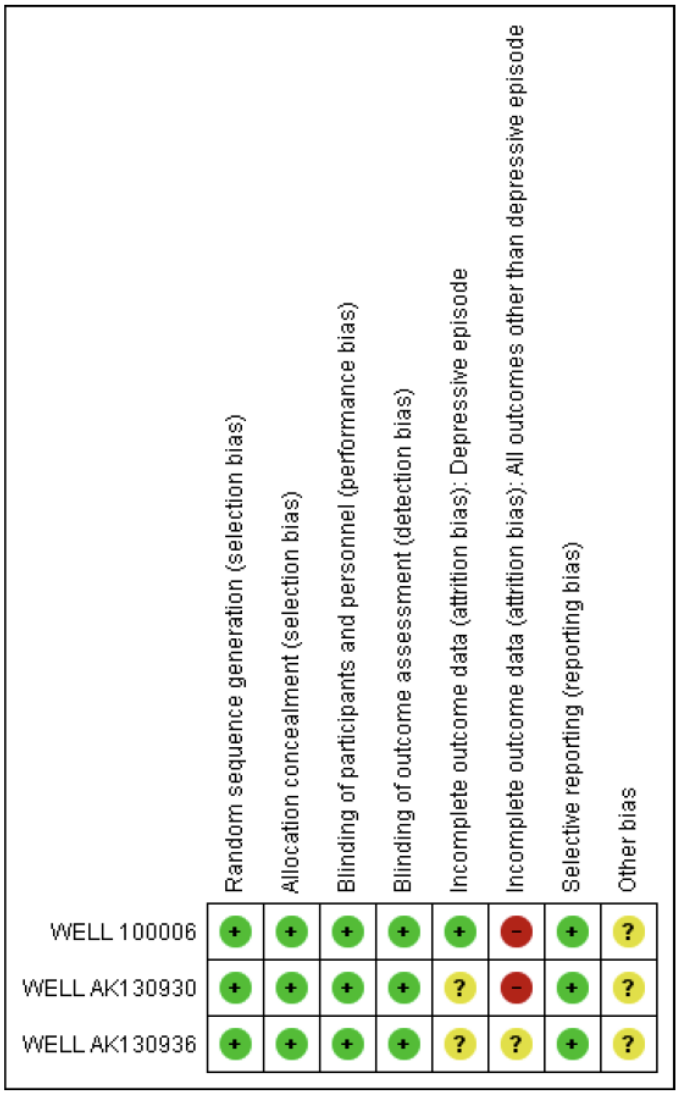

Abb. 6: Beispiel „risk of bias summary” zu einer Cochrane RoB Bewertung[53] (S.14)

Risk of bias for analysis 1.3 Physical activity (device-worn)

	Bias					
Study	**Randomisation process**	**Deviations from intended interventions**	**Missing outcome data**	**Measurement of the outcome**	**Selection of the reported results**	**Overall**
Duppen 2015	+	+	+	~	~	~
Klausen 2016	+	+	~	+	+	~
Morrison 2013	~	+	~	~	~	~
Opotowsky 2018	+	+	+	~	~	~

Abb. 7: Beispiel „risk of bias summary" zu einer RoB 2-Bewertung[44 (S.65)]

Figure 4. Physical activity promotion and exercise training interventions versus no activity (usual care) in people with congenital heart disease. Outcome: Physical activity (device-worn).

Study or Subgroup	Physical activity Mean	SD	Total	No activity (usual care) Mean	SD	Total	Weight	Std. Mean Difference IV, Random, 95% CI	Risk of Bias A B C D E F
Duppen 2015 (1)	12.7	8.1	28	11.8	6.2	18	23.7%	0.12 [-0.47 , 0.71]	+ + + ? ? ?
Klausen 2016	40.3	21.8	81	41.3	22.9	77	30.2%	-0.04 [-0.36 , 0.27]	+ + ? + + ?
Morrison 2013	57.2	32.2	62	29.2	27.3	39	27.8%	0.91 [0.49 , 1.33]	? + ? ? ? ?
Opotowsky 2018	103	56	11	75	25	12	18.3%	0.63 [-0.21 , 1.47]	+ + + ? ? ?
Total (95% CI)			**182**			**146**	**100.0%**	**0.38 [-0.15 , 0.92]**	

Heterogeneity: Tau² = 0.22; Chi² = 13.81, df = 3 (P = 0.003); I² = 78%
Test for overall effect: Z = 1.42 (P = 0.16)
Test for subgroup differences: Not applicable

Std. Mean Difference IV, Random, 95% CI

-2 -1 0 1 2

Favours no activity (usual care) Favours physical activity

Footnotes
(1) Measure of activity: time spent in moderate-to-very-vigorous activity as a percentage. All other studies report minutes of MVPA per day.

Risk of bias legend
(A) Bias arising from the randomization process
(B) Bias due to deviations from intended interventions: Physical activity (device-worn)
(C) Bias due to missing outcome data: Physical activity (device-worn)
(D) Bias in measurement of the outcome: Physical activity (device-worn)
(E) Bias in selection of the reported result: Physical activity (device-worn)
(F) Overall bias: Physical activity (device-worn)

Abb. 8: Beispiel Metaanalyse-Ergebnisse mit RoB 2-Biasbewertungen[44 (S.13)]

Schlussfolgerungen formulieren

Das Gesamturteil über das **Biasrisiko** für ein bestimmtes Studienergebnis sollte *mit diesem zusammen* dokumentiert und in den Analysen, der **Diskussion** und den **Schlussfolgerungen** eines Berichtes (z.B. zu einem systematischen Review oder einer Leitlinie) angemessen berücksichtigt werden. Hilfestellung hierfür geben u.a. die verfügbaren Berichtsqualitäts-Instrumente[56].

6 BEWERTUNG DES BIASRISIKOS IN NICHT-RANDOMISIERTEN VERGLEICHENDEN INTERVENTIONSSTUDIEN

6.1 Definition: Nicht-randomisierte vergleichende Interventionsstudien

Wie im vorangegangenen Kapitel dargestellt, sind randomisierte kontrollierte Studien das am besten geeignete Studiendesign zur Schätzung der Effekte von Interventionen. Die erfolgreiche Randomisierung einer ausreichend hohen Anzahl von Teilnehmer*innen führt im Allgemeinen zu Interventions- und Vergleichsgruppen, in denen die Verteilung von (bekannten sowie unbekannten) prognostischen Faktoren vergleichbar ist, sodass Unterschiede in den Ergebnissen der Studiengruppen kausal auf die Intervention zurückgeführt werden können.

Nicht immer lässt sich jedoch eine klinische Frage mit randomisierten kontrollierten Studien beantworten. Manchmal ist eine Randomisierung schwierig (ggf. nicht akzeptiert) oder es liegen aus anderen Gründen (z.B. durch fehlende Forschungsförderung) keine randomisierten kontrollierten Studien vor. Manchmal sind die verfügbaren randomisierten kontrollierten Studien klein und methodisch mangelhaft. In Fällen wie diesen kann die Betrachtung von (idealerweise großen, methodisch hochwertigen) **nicht-randomisierten Interventionsstudien („non-randomized studies of interventions“, NRSI)** sinnvoll und notwendig sein.

NRSI sind quantitative Studien, in denen die Effektivität (Nutzen oder Schaden) einer Intervention untersucht wurde, in denen jedoch keine Randomisierung für die Zuteilung der Teilnehmer*innen zu den Interventions- bzw. Vergleichsgruppen vorgenommen wurde. Hierzu zählen Studien, in denen die Zuteilung im Rahmen gewöhnlicher Behandlungsentscheidungen erfolgt und die häufig als „beobachtend“ bzw. „Beobachtungsstudien“ bezeichnet werden. Es gibt eine Reihe unterschiedlicher Arten von NRSI, darunter **Kohortenstudien**, **Fall-Kontroll-Studien**, **kontrollierte Vorher-Nachher-Studien**, **„unterbrochene Zeitserien“** („interrupted time series“) und kontrollierte Studien, in denen zur Zuteilung der Teilnehmer*innen zu den Interventions- bzw. Vergleichsgruppen Methoden angewandt werden, die nicht einer vollwertigen Randomisierung entsprechen (und die mitunter als **„quasi-randomisierte Studien“** bezeichnet werden).

Von den nicht-randomisierten vergleichenden Studien sind **nicht-vergleichende Studien** zu unterscheiden, die in Kapitel 7 adressiert werden.

6.2 Wichtige Biasformen in nicht-randomisierten vergleichenden Interventionsstudien

Das Potential für Bias ist in NRSI grundsätzlich größer als in RCTs. Ein wesentlicher Aspekt dabei ist Bias durch **Confounding**. NRSI können zudem von anderen Biasformen betroffen sein, die in der epidemiologischen Literatur als **Selection Bias** und **Information Bias** bezeichnet werden (s. weiter). In einigen Aspekten gleichen diese den gleichnamigen Biasformen für RCTs, jedoch gibt es auch einige, teilweise in erster Linie die Terminologie betreffende, Unterschiede. Auch der **Reporting Bias** ist eine relevante Biasform in NRSI.

Da in diesem Manual das ROBINS-I Tool für die Bewertung des Biasrisikos in NRSI vorgestellt wird, basiert die folgende Zusammenstellung der wesentlichen Biasformen in NRSI auf den im **detailed guidance document**[57] zum ROBINS-I Tool dargestellten Biasformen und den weiteren zu ROBINS-I verfügbaren Ressourcen[38,58] (s. hierzu Abschnitt 6.3.1) und orientiert sich an der in ROBINS-I verwendeten Terminologie.

6.2.1 Bias durch Confounding

Aufgrund des Fehlens einer randomisierten Zuteilung der Teilnehmer*innen zu der/den Interventions- und Vergleichsgruppe(n) unterscheiden sich die Teilnehmer*innen in NRSI in der Regel zwischen den Studien-Gruppen in ihren Merkmalen bzw. in verschiedenen Faktoren (wie beispielsweise Alter, Schweregrad der Erkrankung oder Vorerkrankungen). Daher ist die Bewertung des **Biasrisikos durch Confounding** ein wesentlicher Bestandteil der Bewertung des Biasrisikos in NRSI.

Confounding (von Lat. „confundere“ = sich vermischen) des Effekts einer Intervention tritt auf, wenn ein oder mehrere prognostische(r) Faktor(en) (Faktoren, die das interessierende Ergebnis vorhersagen) auch vorhersagen, welche Intervention ein/e individuelle/r Teilnehmer*in erhält. Solche Faktoren werden als „Confounding-Domäne“ (oder vereinfacht als **„Confounder“**) bezeichnet. Liegt ein Confounding vor, entspricht das berichtete Studienergebnis nicht mehr dem kausalen Effekt der Intervention, da es auch durch den bzw. die Confounder (-Domäne) zustandegekommen sein könnte. Wenn beispielsweise die Teilnehmer*innen einer Studiengruppe, die eine Intervention A erhalten, jünger sind als die Teilnehmer*innen der Gruppe, die eine Intervention B erhalten, ist es schwierig zu ermitteln, inwieweit die Ergebnisse der Gruppe A auf die Intervention oder auf das geringere Alter zurückzuführen sind. Das Alter der Teilnehmer*innen würde in diesem Fall einen Confounder darstellen.

Beim Confounding wird zwischen **Baseline Confounding** und **zeitabhängigem Confounding („time-varying confounding“)** unterschieden. Baseline Confounding tritt auf, wenn ein vor dem Start der Intervention(en) („pre-intervention“) erhobener prognostischer Faktor (ggf. mehrere Faktoren) vorhersagt, welche Intervention ein/e Teilnehmer*in initial durchführt bzw. erhält. Zeitabhängiges Confounding tritt auf, wenn sich die durchgeführte Intervention im Laufe der Zeit ändern kann, beispielsweise, wenn Teilnehmer*innen zwischen den zu vergleichenden Interventionen wechseln, und wenn ein (ggf. mehrere) nach dem Start der Intervention („post-baseline“) erhobener prognostischer Faktor (ggf. mehrere Faktoren) einen Einfluss auf die nach dem Start durchgeführte Intervention hat. Zeitabhängiges Confounding muss in Studien berücksichtigt werden, in denen die Nachbeobachtungszeit für die Teilnehmer*innen entsprechend der Zeit aufgeteilt wurde, die diese in verschiedenen Interventionsgruppen verbrachten.

In Studien werden üblicherweise bestimmte (häufig als Confounder oder „Störfaktoren“ bezeichnete) Variablen erhoben, mit denen für eine Confounding-Domäne ganz oder teilweise kontrolliert werden soll. So können beispielsweise der Body-Mass-Index und das Hüft-Taillen-Verhältnis („hip-waist ratio“) verwendet werden, um für die Confounding-Domäne Adipositas zu kontrollieren. Confounding-Domänen, die für eine bestimmte Intervention relevant sind, können sich zwischen verschiedenen Studien-Settings unterscheiden.

Das Risiko von Confounding kann prinzipiell durch das Design einer Studie minimiert werden, beispielsweise durch Einschränkung des Einschlusses auf Personen, bei denen die Baseline-Confounder denselben Wert haben. Ansonsten kann, was häufiger vorkommt, mit Confounding durch geeignete statistische Verfahren umgegangen werden, mit denen für den bzw. die relevanten Confounder adjustiert („kontrolliert“) wird. Das Adjustieren für Faktoren, die keine Confounder sind, insbesondere aber das Adjustieren für Variablen, die durch die Intervention beeinflusst worden sein könnten („post-intervention variables“), kann zu Bias führen.

In der Praxis kann Confounding nicht vollständig ausgeräumt werden. **Residuales Confounding („residual confounding“)** liegt vor, wenn eine Confounding-Domäne fehlerhaft gemessen wurde oder die Beziehung zwischen der Confounding-Domäne und dem Ergebnis oder der Exposition (abhängig vom verwendeten Analyse-Ansatz) unvollkommen modelliert wurde. Beispielsweise (s. **Kapitel 25 Cochrane Handbook**[58]) wäre in einer NRSI, in der zwei blutdrucksenkende Medikamente miteinander verglichen werden, residuelles Confounding zu erwarten, wenn der initiale („pre-intervention“) Blutdruck drei Monate vor Beginn der Intervention gemessen worden wäre, der Blutdruck jedoch, den Kliniker dazu verwenden, um zum Zeitpunkt der Intervention zwischen den

Medikamenten zu entscheiden, im Datensatz nicht verfügbar wäre. **Nicht gemessenes Confounding („unmeasured confounding“)** liegt vor, wenn eine Confounder-Domäne gar nicht gemessen wurde oder wenn für sie in der Analyse nicht kontrolliert wurde. In dem genannten Beispiel wäre dies der Fall, wenn keine initialen Blutdruckmessungen verfügbar wären oder wenn in der Analyse nicht für den initialen Blutdruck kontrolliert worden wäre, obwohl er gemessen wurde. Das Vorliegen eines nicht gemessenen Confounding kann in der Regel nicht ausgeschlossen werden, weil selten Sicherheit darüber besteht, dass alle relevanten Confounder-Domänen bekannt sind.

6.2.2 Selection Bias

Selection Bias kann in NRSI auftreten, wenn die Selektion der Teilnehmer*innen oder ihrer Nachbeobachtungszeit für den Einschluss in die Studie in einem Zusammenhang sowohl mit der Intervention als auch mit dem Ergebnis steht. So waren beispielsweise Studien zur Folsäure-Supplementation während der Schwangerschaft zur Prävention von Neuralrohrdefekten bei Kindern verzerrt, weil in sie Mütter und Kinder nur dann eingeschlossen wurden, wenn die Kinder lebend geboren worden waren[59] (s. **Kapitel 25 Cochrane Handook**[58]). Der Bias trat deshalb auf, weil eine Lebendgeburt (anstelle einer Totgeburt oder therapeutischen Abtreibung, für die in den Studien keine Ergebnisdaten verfügbar waren) in einem Zusammenhang sowohl mit der Intervention als auch mit dem Ergebnis stand (eine Folsäure-Supplementation erhöht die Wahrscheinlichkeit einer Lebendgeburt, während das Vorliegen von Neuralrohrdefekten die Wahrscheinlichkeit einer Lebendgeburt verringert)[59,60] (s. **Kapitel 25 Cochrane Handbook**[58]).

Selection Bias kann auch auftreten, wenn ein Teil der Nachbeobachtungszeit („follow-up“) von der Analyse ausgeschlossen wird. Dies ist zum Beispiel der Fall, wenn bei Teilnehmer*innen in einer Interventionsgruppe der Beobachtungszeitpunkt (gleichzusetzen mit Studienbeginn) nicht mit dem Behandlungsbeginn zusammenfällt. Dies kann der Fall sein, wenn Patient*innen in eine Studie eingeschlossen werden, die die zu untersuchende Intervention oder medizinische Maßnahme bereits einnehmen beziehungsweise erhalten („prevalent users“ anstelle von „new users“). In solchen Fällen liegt der Behandlungs- vor dem Studienbeginn und Ereignisse, die vor Studienbeginn aufgetreten sind, werden nicht erfasst. Ein systematischer Ausschluss der initialen Beobachtungszeit kann zu einer Über- oder Unterschätzung des Effektes (Nutzens oder Schadens) einer Intervention führen. Diese Form des Selection Bias wird auch als Inception Bias oder Lead-time Bias (Vorlaufzeitbias) bezeichnet.

Selection Bias kann zudem aufgrund von fehlenden Daten, beispielsweise durch Studienabbrüche (fehlende Daten von Teilnehmer*innen für einen oder mehrere Nachbeobachtungs-Zeitpunkte),

versäumte Untersuchungstermine, eine unvollständige Datenerhebung oder den Ausschluss von Teilnehmer*innen aus der Analyse durch die Studienleiter auftreten. Fehlende Daten in NRSI können Baseline-Charakteristika (einschließlich bereits erhaltener Interventionen oder Baseline-Confounder), vorab spezifizierte Ko-Interventionen, Ergebnismessungen, andere in den Analysen verwendeten Variablen oder Kombinationen aus diesen betreffen.

6.2.3 Information Bias

Information Bias (auch als **Measurement Bias** bezeichnet) kann auftreten, wenn der Interventionsstatus (welche Intervention(en) die Teilnehmer*innen erhalten haben) fehlklassifiziert, d.h. falsch zugeordnet, wird, oder wenn Ergebnisse fehlklassifiziert oder fehlerhaft gemessen werden. Hierbei wird zwischen differentiellen und nicht-differentiellen Fehlern unterschieden. Differentielle Fehler stellen in der Regel das größere Problem dar.

Die **nicht-differentielle Fehlklassifikation („non-differential misclassification“)** steht in keinem Zusammenhang mit dem Ergebnis. Beispielsweise könnte in einem Vergleich der Installation von Rauchwarnmeldern mit keiner Installation von Rauchwarnmeldern der Interventionsstatus unvollständig dokumentiert sein und einige Personen, die einen Rauchwarnmelder installierten, fälschlich der „Kein Rauchwarnmelder“-Gruppe zugeordnet werden. Wenn eine solche Fehlklassifikation in keinem Zusammenhang mit dem nachfolgenden Ergebnis steht, d.h. beispielsweise das Auftreten feuerbedingter Verletzungen in keinem Zusammenhang mit den Gründen für die fehlende Identifizierung der Installation von Rauchwarnmeldern steht, ist sie nicht-differentiell und wird in der Regel den geschätzten Effekt der Intervention in Richtung Null (d.h. in Richtung keines Effektes) verzerren.

Eine **differentielle Fehlklassifikation („differential misclassification“)** des Interventionsstatus liegt vor, wenn die Fehlklassifikation in einem Zusammenhang mit dem nachfolgenden Ergebnis oder dem Risiko für das Ergebnis steht. Eine differentielle Fehlklassifikation (oder ein Messfehler) eines Ergebnisses liegt vor, wenn die Fehlklassifikation in einem Zusammenhang mit dem Interventionsstatus steht. In randomisierten Interventionsstudien ist die Fehlklassifikation des Interventionsstatus selten ein Problem, weil in ihnen die Interventionen von den Wissenschaftler*innen aktiv zugeteilt werden und ihre genaue Dokumentation ein wesentliches Merkmal der Studie ist. In Beobachtungsstudien ist die Ermittlung von Informationen über die zugeteilten bzw. die durchgeführten oder erhaltenen Interventionen jedoch häufig schwieriger. Ein bekanntes Beispiel für eine differentielle Fehlklassifikation in einem Fall, bei dem die Kenntnis eines nachfolgenden Ergebnisses die Klassifikation von Interventionen beeinflussen kann, ist der Recall

Bias in Fall-Kontroll-Studien: die „Fälle“ erinnern sich möglicherweise eher als die „Kontrollen“ an potenziell relevante Ereignisse, oder geben eher die Exposition gegenüber Risikofaktoren an, die sie für verantwortlich für ihre Erkrankung halten.

Zur Vermeidung einer differentiellen Fehlklassifikation des Interventionsstatus ist es wichtig, dass die Interventionen, wenn möglich, ohne Kenntnis der nachfolgenden Ergebnisse definiert und klassifiziert werden.

Bias aufgrund der Ergebnismessung wird häufig als **Detection Bias** bezeichnet. Beispiele für das Auftreten eines solchen Bias sind Situationen, in denen die das Ergebnis messenden Personen den Interventionsstatus der Teilnehmer*innen kennen (insbesondere, wenn die Ergebnismessung subjektiv ist), oder wenn Ergebnisse in unterschiedlichen Interventionsgruppen auf unterschiedliche Weise gemessen werden.

6.2.4 Reporting Bias

Die möglichen **Bedenken hinsichtlich des selektiven Berichtens** der Ergebnisse von NRSI sind dieselben wie bei RCTs (s. auch 5.2.5). So tritt Bias durch Selektion eines berichteten Ergebnisses („bias in selection of the reported result“) auf, wenn die Selektion auf dem p-Wert oder der Größe oder Richtung des geschätzten Effektes der Intervention basiert. Bias durch Selektion der Endpunktmessung tritt auf, wenn ein Ergebnis für einen bestimmten Endpunkt aus mehreren Messungen ausgewählt wurde, beispielsweise, wenn eine Messung mehrmals oder mit unterschiedlichen Verfahren durchgeführt wurde. Bias durch Selektion der Analyse tritt auf, wenn das berichtete Ergebnis aus Interventionseffekten ausgewählt wurde, die auf mehrere Arten ermittelt wurden, zum Beispiel durch Analysen von Veränderungswerten und für Nach-Interventions-Werte, die für Baseline-Werte adjustiert wurden. Außerdem kann Bias durch selektives Berichten einer Teilnehmer*innen-Subgruppe innerhalb einer größeren Studie auftreten, wenn die Auswahl auf einem „interessanteren Ergebnis“ basiert.

Ergänzung: ROB-ME Tool zur Bewertung des Biasrisikos aufgrund der Nicht-Verfügbarkeit eines Studienergebnisses für den Einschluss in eine Synthese

Das im Kontext der Bewertung des Biasrisikos von randomisierten kontrollierten Studien (Abschnitt 5.2.5) erwähnte neue „Bewertungsinstrument für die Bewertung des Biasrisikos aufgrund von fehlender Evidenz in einer Synthese“, das ROB-ME Tool[41] („A tool for assessing Risk Of Bias due to Missing Evidence in a synthesis“) kann auch für die Bewertung des Biasrisikos in nicht-randomisierten Interventionsstudien angewendet werden (s. Abschnitt 5.2.5).

6.3 Das ROBINS-I Tool

6.3.1 Hintergrund

Das **ROBINS-I** Tool („Risk Of Bias In Non-randomized Studies - of Interventions“, im Folgenden kurz „ROBINS-I“ genannt) ist ein Instrument zur Bewertung des Biasrisikos der Ergebnisse von NRSI, in denen die Effekte von zwei oder mehr Interventionen verglichen werden.[38] Das Tool wurde von Mitgliedern der **Cochrane Bias Methods Group**[61] und der **Cochrane Non-Randomized Studies for Interventions Methods Group**[62] entwickelt. Es wurde initial 2016 publiziert. ROBINS-I ist eine Weiterentwicklung des **ACROBAT-NRSI**[63] Tools („Cochrane Risk Of Bias Assessment Tool for Non-Randomized Studies of Interventions“) und hat dieses ersetzt. Von der weiteren Verwendung von ACROBAT-NRSI wird explizit abgeraten[63]. ROBINS-I wird von Cochrane zur Bewertung des Biasrisikos von NRSI in Cochrane Reviews empfohlen[58]. Die (externe) Validierung dieses neuen Bewertungsinstruments steht noch aus.

Die in diesem Manual dargestellten Informationen zu ROBINS-I basieren auf den folgenden zu dem Tool verfügbaren Ressourcen (die vollständigen bibliographischen Angaben finden sich in der Literaturliste am Ende des Manuals):

- Sterne et al. 2016: „**ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions**“[38]
- Sterne et al. (Hrsg.) „**Risk Of Bias In Non-randomized Studies of Interventions (ROBINS-I): detailed guidance**“[57] (letzte Aktualisierung: 20. Oktober 2016)
- Sterne et al. 2019: „**Assessing risk of bias in a non-randomized study“ - Chapter 25, Cochrane Handbook for Systematic Reviews of Interventions**[58]

Das ROBINS-I Tool und die zu ihm verfügbaren Ressourcen einschließlich relevanter Publikationen in Fachzeitschriften sind auf einer **eigenen Webpage**[64] verfügbar. Dort findet sich auch **eine Version des Tools mit Erklärungen**[65] zu den einzelnen Signalfragen und Bewertungen, die zusammenfassende Erläuterungen zu allen Domänen enthält, sowie eine leere Vorlage des Tools, ein **ROBINS-I template**[66], für eigene Bewertungen.

6.3.2 Welche nicht-randomisierten Studientypen können mit ROBINS-I bewertet werden?

Grundsätzlich wurde ROBINS-I für **verschiedene Arten von NRSI** entwickelt, d.h. für jegliche quantitative Studien, in denen die Effektivität (Nutzen oder Schaden) einer Intervention untersucht wurde, in denen jedoch keine Randomisierung für die Zuteilung der Teilnehmer*innen zu den

Interventions- bzw. Vergleichsgruppen vorgenommen wurde, das heißt für Kohortenstudien, Fall-Kontroll-Studien, kontrollierte Vorher-Nachher-Studien, „unterbrochene Zeitserien“ (Englisch: „interrupted time series“) und kontrollierte Studien, in denen für die Zuteilung zu den Vergleichsgruppen Methoden angewandt werden, die nicht einer vollwertigen Randomisierung entsprechen (und mitunter als „quasi-randomisierte Studien“ bezeichnet werden).

Der in diesem Manual (und in den ihm zugrundeliegenden aktuell zu ROBINS-I verfügbaren Publikationen) dargestellte Leitfaden für die Anwendung von ROBINS-I bezieht sich primär auf NRSI mit Kohorten-Designs, quasi-randomisierte Studien und andere kontrollierte Studien mit parallelen Interventionen. ROBINS-I ist in großen Teilen auch für andere Arten von NRS, wie Fall-Kontroll-Studien, Querschnittsstudien, unterbrochene Zeitreihen und kontrollierte Vorher-Nachher-Studien relevant. Jedoch weisen die ROBINS-I-Autor*innen in ihrem **detailed guidance document**[67] zu ROBINS-I darauf hin, dass es Überlegungen gibt, für diese Studiendesigns einige der in dem Tool enthaltenen (Signal-) Fragen zu verändern bzw. anzupassen. In **Kapitel 25 der aktuellen Version des Cochrane Handbook**[58] finden sich für Interessierte Informationen zur Bewertung des Biasrisikos mit ROBINS-I in Kohortenstudien („Follow-up“ Studien), nicht-kontrollierten Vorher-Nachher-Studien und kontrollierten Vorher-Nachher-Studien. Informationen zu den aktuellsten Varianten, Ergänzungen oder Änderungen von ROBINS-I sind auf der **ROBINS-I-Website**[64] verfügbar.

Mit ROBINS-E **(Risk Of Bias In Non-randomized Studies - of Exposures)** ist eine erste Variante von ROBINS-I in Arbeit, die sich mit der Bewertung des Biasrisikos in nicht-randomisierten Studien zu Expositionen befasst. ROBINS-E wird voraussichtlich in naher Zukunft zur Verfügung stehen, ein Publikationsdatum war zum Zeitpunkt des Redaktionsschlusses für die vorliegende Version des Manuals jedoch noch nicht bekannt.

6.3.3 Wie ist ROBINS-I aufgebaut?

ROBINS-I ist, wie RoB 2, ein **domänenbasiertes Tool** und weist in Aufbau und Gestaltung viele Gemeinsamkeiten mit RoB 2 auf. ROBINS-I ist in **sieben Domänen** unterteilt, die sich auf unterschiedliche Bias-Aspekte beziehen. Die Domänen decken alle wesentlichen Biasformen ab, die nach aktuellem Verständnis die Ergebnisse von NRSI beeinflussen können (s. auch 6.2). Dies sind:

1) Bias durch Confounding

2) Bias durch die Selektion der Teilnehmer*innen für den Einschluss in die Studie

3) Bias durch die Klassifikation der Interventionen

4) Bias durch Abweichungen von den vorgesehenen Interventionen

5) Bias durch fehlende Daten

6) Bias durch die Ergebnismessungen

7) Bias durch die Selektion des berichteten Ergebnisses

Zu jeder Domäne gibt es mehrere **Signalfragen**, deren Beantwortung die Bewertung des Biasrisikos für die Domäne erleichtern sollen. Aus den Bewertungen der einzelnen Domänen wird eine **Gesamtbewertung des Biasrisikos** für die interessierende Studie in Bezug auf das interessierende Ergebnis abgeleitet. Die Bewertung mit ROBINS-I erfolgt, wie bei ROB 2, ergebnisbezogen, d.h. sie wird für jeweils ein vorab spezifiziertes (numerisches) Ergebnis durchgeführt. Entsprechend erfordert die Bewertung des Biasrisikos für mehrere Ergebnisse die Durchführung mehrerer Bewertungen. Sollen mehrere Ergebnisse bewertet werden, deren Bewertung als identisch betrachtet wird, kann die Bewertung auch, wie bei RoB 2, für Ergebnisgruppen erfolgen.

6.3.4 Wie wird die Bewertung mit ROBINS-I durchgeführt?

Für eigene Bewertungen mit ROBINS-I ist ein **ROBINS-I template** verfügbar. Vor bzw. begleitend zu eigenen Bewertungen sollten die zuvor genannten Ressourcen[38,57,58] (s. Abschnitt 6.3.1) verwendet werden, da diese wichtige Hilfestellungen für das Verständnis und die Anwendung des Tools bieten und hierüber der Sicherstellung der angemessenen Anwendung dienen.

Die Bewertung des Biasrisikos für NRSI ist allgemein komplexer als die Bewertung für RCTs und erfordert eine entsprechend hohe methodische und inhaltliche Expertise.[68] Daher wird empfohlen, in den Bewertungsprozess Expert*innen einzubeziehen, d.h. Methodiker*innen mit Erfahrung in den relevanten Studien-Designs (bzw. relevanten Aspekten der Studiendesigns) und Gesundheitsfachpersonen, die über Kenntnisse der prognostischen Faktoren verfügen und Behandlungsentscheidungen für die jeweilige Zielgruppe (Teilnehmer*innengruppe bzw. Population) treffen.

Bewertungsprozess in drei Phasen

Der Bewertungsprozess lässt sich in drei grundsätzliche Phasen unterteilen:

- Phase 1: Protokollstadium
- Phase 2: Bewertung des Biasrisikos
- Phase 3: Schlussfolgerungen (abschließende Gesamt-Beurteilung)

Phase 1: Protokollstadium

Entsprechend dem Vorgehen bei der Bewertung eines RCT mit RoB 2 sollten vor Beginn der Bewertung des Biasrisikos einer NRSI einige **Vorabspezifizierungen** stehen, die dabei helfen sollen, wichtige potenzielle Probleme der einzuschließenden NRSI zu identifizieren bzw. zu bedenken. Abbildung 9 zeigt einen exemplarischen Ausschnitt aus dem originalen **ROBINS-I template**[66] zu diesem Abschnitt.

In Phase 1 der ROBINS-I-Bewertung sind die folgenden Aspekte zu spezifizieren:

- Die **Fragestellung des Reviews** (Teilnehmer*innen, experimentelle Intervention, Vergleich, Endpunkte)
- **Relevante Confounder-Domänen**, die für alle oder die meisten Studien relevant sind
- **Relevante Ko-Interventionen**, die sich zwischen den Gruppen unterscheiden und die einen Einfluss auf die Ergebnisse haben könnten

ROBINS-I Tool (Phase I): Protokollstadium

Spezifizieren Sie die Fragestellung des Reviews

Teilnehmer*innen	
Experimentelle Intervention	
Vergleich	
Endpunkte	

Benennen Sie die Confounder-Domänen, die für alle oder die meisten Studien relevant sind

Benennen Sie Ko-Interventionen auf, die sich zwischen den Interventionsgruppen unterscheiden und die Ergebnisse beeinflussen könnten

Abb. 9: Protokollstadium: Auszug aus dem ROBINS-I template[66] (übersetzt und adaptiert)

Relevante Confounder-Domänen

Relevante **Confounder-Domänen** sind diejenigen prognostischen Faktoren, die vorhersagen, ob ein/e Teilnehmer*in die eine oder die andere interessierende Intervention erhält. Ihre Identifizierung erfolgt in der Regel auf Basis von Fachkenntnissen von Experten aus dem Review-Team und initialen Literaturübersichten. Sie kann ergänzt werden durch Diskussionen mit Gesundheitsfachpersonen, die an Interventionsentscheidungen für die jeweilige Zielgruppe beteiligt sind.

Relevante Ko-Interventionen

Relevante **Ko-Interventionen** sind diejenigen Interventionen (oder Expositionen), die Teilnehmer*innen nach oder mit Beginn der interessierenden Intervention erhalten, die in einem Zusammenhang mit der jeweiligen Intervention stehen und prognostisch für das interessierende Ergebnis sind. Ihre Identifizierung erfolgt in der Regel, wie bei den Confoundern, durch Fachkenntnisse, Literaturübersichten und den Einbezug von Gesundheitsfachpersonen.

Phase 2: Bewertung des Biasrisikos

Vor der eigentlichen Bewertung des Biasrisikos sind in ROBINS-I in dieser Phase einige weitere Aspekte zu spezifizieren:

- Ein **„Ziel-RCT“** (Design, Teilnehmer*innen, experimentelle Intervention, Vergleich)
- Das der Bewertung zugrunde gelegte **Ziel der Studie**
- Der zu bewertende **Endpunkt**
- Das interessierende **numerische Ergebnis** (Effektschätzer)
- Eine vorläufige Beurteilung der **Confounder und Ko-Interventionen**

Spezifizierung eines „Ziel-RCTs“

Der Schwerpunkt von ROBINS-I liegt, genau wie bei RoB 2, auf der Bewertung der internen Validität (des Biasrisikos) der zu bewertenden Studie. Für beide Studienarten, NRSI und RCT, kann Bias definiert werden als die Tendenz der Studienergebnisse, systematisch von den Ergebnissen abzuweichen, die von einer randomisierten Studie zu erwarten wären, die mit derselben Teilnehmer*innengruppe durchgeführt wurde und frei von methodischen Schwächen, d.h. von Bias, ist. Eine solche Studie wäre typischerweise ein großer RCT, der alle wesentlichen Aspekte mit Relevanz für ein Bias-Risiko erfüllt und in dem die Effekte der Intervention für alle Endpunkte, die in der Studie gemessen wurden, berichtet sind.

Für die ROBINS-I-Bewertung wird jede NRSI als Versuch betrachtet, einen hypothetischen RCT zu imitieren (nachzuahmen), der als **„Ziel-RCT“ („target trial“)** bezeichnet wird. Der Ziel-RCT dient als „Referenzstandard“ und soll die Bewertung erleichtern. Es handelt sich dabei um den hypothetischen pragmatischen RCT, mit dem die gesundheitsbezogenen Effekte derselben Intervention mit derselben Teilnehmer*innengruppe ermittelt werden und der frei von Bias ist. Das Biasrisiko der NRSI wird entsprechend in Bezug auf den Effekt ermittelt, der in diesem RCT zu erwarten wäre. Der Ziel-RCT muss dabei weder realisierbar noch ethisch vertretbar sein: so könnte er beispielsweise in einem Vergleich von zwei Teilnehmer*innengruppen bestehen, von denen eine

dem „regelmäßigen Konsum alkoholischer Getränke“, die andere dem „Nicht-Konsum alkoholischer Getränke“ zugeteilt wird.

Für die ROBINS-I-Bewertung sind die folgenden Angaben zum Ziel-RCT zu spezifizieren, die in dem RCT untersucht würden, den die zu bewertende NRS zu imitieren versucht:

- Das **Design**
- Die **Teilnehmer*innen**
- Die **experimentelle Intervention**
- Der **Vergleich**

Da die in den NRSI verfügbaren Informationen zu den Interventionen in aller Regel für die Definition des Ziel-RCT nicht ausreichen, ist hierfür häufig eine entsprechende Fachkenntnis erforderlich.

Hinweis zur Terminologie: Analog zur Terminologie des Ziel-RCT wird im **detailed guidance document**[57] (und folglich in diesem Manual) der Begriff „Interventions“-Gruppe (bzw. „experimentelle Intervention“) anstelle von „Behandlungs“- oder „Expositions“-Gruppe, wie es sonst im Zusammenhang mit Beobachtungsstudien üblich ist, verwendet, auch wenn in solchen Studien faktisch keine Intervention implementiert wurde.

Studienziel – Spezifizierung des interessierenden Effektes

Das Biasrisiko wird in ROBINS-I bezogen auf ein **vorab definiertes Ziel,** bzw. auf einen von zwei interessierenden Effekten bewertet. Der interessierende Effekt einer beliebigen Forschungsfrage kann entweder der **Effekt der Zuteilung zur Intervention** zu Beginn der Studie („at baseline“), unabhängig von dem Ausmaß, in dem die Interventionen in der Nachbeobachtungszeit durchgeführt bzw. verabreicht wurden, sein, oder der **Effekt des Startens und Einhaltens der Intervention** wie im Protokoll spezifiziert. Die Entscheidung für den einen oder den anderen Effekt ist von den Gutachter*innen zu treffen und sollte nicht auf den Entscheidungen der Studien-Autor*innen zu den Analysen der NRSI basieren. Die Analysen der NRSI können jedoch einem der beiden interessierenden Ziele mehr entsprechen und entsprechend in Bezug auf das andere Ziel verzerrt sein.

Im Kontext von RCTs kann der Effekt der Zuteilung zur Intervention mit einer Intention-to-treat-(ITT)-Analyse geschätzt werden, bei der die Teilnehmer*innen entsprechend der Interventionsgruppe, der sie randomisiert zugeteilt wurden, analysiert werden. Alternativ kann das Interesse sein, den Effekt des Startens und Einhaltens der Intervention wie im Studienprotokoll spezifiziert (und wie auch als „per protocol“ Effekt bezeichnet) zu schätzen. Für NRSI lassen sich

Analogien zu diesen Effekten definieren. So kommt beispielsweise der Effekt des Startens (ggf. auch des Verordnens) der experimentellen Intervention gegenüber dem Starten (ggf. auch dem Verordnen) der Vergleichsintervention(en) dem ITT-Effekt in einer klinischen Studie nahe, in der die Teilnehmer*innen, die einer Intervention zugeteilt werden, diese immer beginnen. Ein kleiner Unterschied zum ITT-Effekt in RCTs ist hierbei, dass einige Teilnehmer*innen, die einer Intervention randomisiert zugeteilt wurden, diese unter Umständen niemals beginnen. Der Effekt des Startens und Einhaltens der Intervention wie im Studienprotokoll spezifiziert entspricht dem Starten und Einhalten der experimentellen Intervention gegenüber dem Starten und Einhalten der Vergleichsintervention(en), solange kein Abbruch aus medizinischen Gründen (z.B. die Unverträglichkeit eines Medikaments) erforderlich ist.

Die unverzerrte Schätzung des Effekts des Startens und Einhaltens von Interventionen, die über einen bestimmten Zeitraum, das heißt nicht nur einmalig, durchgeführt bzw. verabreicht werden, erfordert sowohl bei RCTs als auch bei NRSI ein angemessenes Adjustieren für prognostische Faktoren („zeitabhängige Confounder“), die Abweichungen von den vorgesehenen Interventionen nach deren Beginn vorhersagen. Es wird empfohlen, sich bei der Bewertung von Interventionseffekten, die unter Verwendung von Methoden zur Adjustierung für zeitabhängiges Confounding geschätzt wurden, von Spezialist*innen mit entsprechender statistischer Expertise beraten zu lassen.

Vorläufige Beurteilung von Confoundern und Ko-Interventionen

Ziel der genaueren **Betrachtung der Confounder und Ko-Interventionen** vor der eigentlichen Bewertung des Biasrisikos ist eine vorläufige Einschätzung, inwieweit die im Protokoll spezifizierten wichtigen Confounder und Ko-Interventionen in der zu bewertenden Studie erhoben wurden, und ob darüber hinaus weitere Confounder und/oder Ko-Interventionen identifiziert wurden, die in der Studie verwendet oder von den Studienautor*innen als relevant erachtet wurden.

In ROBINS-I werden hierzu die folgenden Aspekte erfasst:

Confounder

- Die **Confounder-Domäne** (wie im Protokoll spezifiziert und ggf. wie zusätzlich in der Studie identifiziert)
- Die **gemessene(n) Variablen**
- Eine **Einschätzung der Notwendigkeit des Kontrollierens für diese Variable(n)**
- Eine **Einschätzung der Validität und Reliabilität der Messungen der Confounder-Domäne**

- Optional: Eine **Einschätzung der erwarteten Auswirkungen des Nicht-Kontrollierens für die Variable** auf den Effekt der Intervention (zugunsten der experimentellen Intervention oder des Vergleichs)

Ko-Interventionen

- Die **Ko-Intervention(en)** (wie im Protokoll spezifiziert und ggf. wie zusätzlich in der Studie identifiziert)
- Eine **Einschätzung der Notwendigkeit des Kontrollierens für die jeweilige Ko-Intervention**
- Optional: Eine **Einschätzung der erwarteten Auswirkungen der Präsenz der jeweiligen Ko-Intervention auf den Effekt der Intervention** (zugunsten der experimentellen Intervention oder des Vergleichs)

Bewertung des Biasrisikos

Nach Abschluss der Vorabspezifizierungen erfolgt die eigentliche Bewertung des Biasrisikos. Jede Studie soll umfassend und sorgfältig unter Berücksichtigung jeglicher potenzieller Biasquellen begutachtet werden. Die Bewertung soll sich auf die in Phase 1 spezifizierten Aspekte (z.B. zu Confoundern) stützen, damit ggf. wichtige unvorhergesehene Probleme identifiziert werden können.

Bias-Domänen in ROBINS-I

Tabelle 5 zeigt eine Übersicht über **die sieben in ROBINS-I enthaltenen Bias-Domänen** (für die meisten Arten von NRSI), die Bias-Kategorien, die sich ihnen zuordnen lassen, und eine kurze Erklärung zu jeder Domäne. Es sind alle Domänen zu bewerten, und es sollten keine zusätzlichen Domänen hinzugefügt werden.

Bias-Domäne	Bias-Kategorie
1. Bias durch Confounding	Confounding
2. Bias durch die Selektion der Teilnehmer*innen für den Einschluss in die Studie	Selection Bias
3. Bias durch die Klassifikation der Interventionen	Information Bias
4. Bias durch Abweichungen von den vorgesehenen Interventionen	Confounding
5. Bias durch fehlende Daten	Selection Bias
6. Bias durch die Ergebnismessung	Information Bias
7. Bias durch Selektion der berichteten Ergebnisse	Reporting Bias

Tab. 5: Bias-Domänen im RoB 2-Tool

Die ersten beiden Domänen befassen sich mit Aspekten, die den **Zeitraum vor dem Start der Intervention(en) („pre-intervention“)** betreffen. Die dritte Domäne befasst sich mit dem **Zeitraum während der Intervention(en) („at intervention“).** Die anderen vier Domänen befassen sich mit Aspekten, die den **Zeitraum nach dem Start der Intervention(en) („post-intervention“)** betreffen.

Die Bewertung der ersten drei Domänen unterscheidet sich weitgehend von der Bewertung von RCTs, weil die Randomisierung gegen Bias vor dem Interventionsstart schützt.

Die Bewertung der weiteren vier Domänen hingegen weist substantielle Überlappungen mit der Biasbewertung von RCTs auf.

Abbildung 10 zeigt einen exemplarischen Auszug aus dem originalen **ROBINS-I template**.[66]

Bewertung des Biasrisikos

Grün unterstrichene Antworten sind potenzielle Hinweise auf ein geringes Biasrisiko und Antworten in Rot sind potenzielle Hinweise auf ein Biasrisiko. Bei Fragen, die sich nur auf Verweise zu weiteren Fragen beziehen, wird keine Formatierung verwendet.

Signalfragen	Beschreibung	Antwortoptionen
Bias durch Confounding		
1.1 Besteht ein Potential für ein Confounding des Effekts der Intervention in dieser Studie? **Bei N/PN bei 1.1:** die Studie kann als von einem niedrigen Bias-Risiko durch Confounding betroffen betrachtet werden und es müssen keine weiteren Signalfragen berücksichtigt werden.		Y/ PY/ PN / N
Bei Y/PY bei 1.1: Ermitteln Sie, ob es erforderlich ist, zeitabhängiges Confounding zu bewerten.		
1.2 Erfolgte die Analyse auf Basis einer Aufsplittung der Nachverfolgungszeit der Teilnehmer*innen entsprechend der erhaltenen Intervention? **Bei N/PN**: beantworten Sie die Fragen zum Baseline-Confounding (1.4 bis 1.6) **Bei Y/PY**: fahren Sie mit Frage 1.3 fort.		NA / Y/ PY / PN / N /NI
1.3 Ist es wahrscheinlich, dass Unterbrechungen oder Wechsel der Intervention mit Faktoren zusammenhängen, die für das Ergebnis prognostisch sind? **Bei N/PN:** Beantworten Sie die Fragen zum Baseline-Confounding (1.4 bis 1.6) **Bei N/PN:** Beantworten Sie die Fragen zum Baseline-Confounding und zeitabhängigem Confounding (1.7 und 1.8)		NA / Y/ PY/ PN/ N /NI
Fragen, die sich nur auf Baseline-Confounding beziehen 1.4 Haben die Autoren eine angemessene Analysemethode angewandt, die für alle wichtigen Confounder-Domänen kontrolliert?		NA / Y/ PY / PN / N /NI

Abb. 10: Biasbewertung: Auszug aus dem ROBINS-I Tool[66]

Signalfragen

Für jede ROBINS-I Domäne wurden **Signalfragen** formuliert, die die Bewertung des Biasrisikos erleichtern sollen.

In ROBINS-I gibt es fünf Antwortoptionen:

(1) **Ja** („yes“ → „Y“)
(2) Wahrscheinlich Ja („probably yes“ → „PY“)
(3) Wahrscheinlich Nein („probably no“ → „PN“)
(4) Nein („no“ → „N“)
(5) Keine Informationen („no information“ → „NI“)

Eine **Ausnahme** ist die Signalfrage 1.1, für die es die Option „Keine Informationen“ nicht gibt. Für einige Signalfragen, die nur unter bestimmten Umständen (in Abhängigkeit der für eine vorausgegangene Signalfrage gegebenen Antwort) zu beantworten sind, gibt es die Antwortoption **„Nicht relevant“ („not applicable“).**

Antworten im ROBINS-I Tool, die **grün unterstrichen** sind, kennzeichnen ein potenziell niedriges Biasrisiko, und Antworten, die **in roter Schrift** dargestellt sind, kennzeichnen ein potenzielles Biasrisiko. Bei Fragen, die sich nur auf Hinweise zu anderen Fragen beziehen, wird keine Formatierung verwendet.

Die Antwortoptionen „Ja“ und „Wahrscheinlich Ja“ (entsprechend „Nein“ und „Wahrscheinlich Nein“) führen zu derselben Biasrisiko-Bewertung, ermöglichen jedoch eine Unterscheidung zwischen einer Antwort, die gesichert ist (für die es Evidenz gibt), und einer Antwort, die wahrscheinlich, aber nicht gesichert ist (und die entsprechend auf einer Einschätzung der Gutachter*innen basiert).

In der Spalte neben jeder Signalfrage gibt es im ROBINS-I Tool **Platz für das Zufügen von Begründungen** für die Antworten bzw. Bewertungen, für die idealerweise relevante Zitate aus den Studien verwendet werden sollten.

Bewertung des Biasrisikos für die Domäne

Die Bewertungen des Biasrisikos für die einzelnen Domänen sollten sich über alle Domänen hinweg konsistent auf den potenziellen Einfluss von Bias auf die Vertrauenswürdigkeit des Ergebnisses beziehen. Wenn die Domänen-Bewertungen konsistent vorgenommen werden, ist die Gesamtbewertung des Biasrisikos für das zu bewertende Ergebnis relativ einfach.

Die **Kriterien für die Bewertung des Biasrisikos für die Domänen** sind wie folgt:

Wenn keine der Antworten auf die Signalfragen der zu bewertenden Domäne auf ein potenzielles Problem hindeutet, kann das Biasrisiko für die Domäne als „niedrig“ bewertet werden. Ist dies nicht der Fall, besteht ein Biasrisiko, bei dem die Gutachter*innen entscheiden müssen, als wie schwerwiegend („moderat“, „schwerwiegend“ oder „kritisch“) es einzuschätzen ist. Die Bewertungsoption “Keine Information” sollte nur dann verwendet werden, wenn die verfügbaren Daten unzureichend sind und kein Urteil über das Biasrisiko erlauben.

Für die Bewertung des Biasrisikos der einzelnen Domänen gibt es bei ROBINS-I fünf Optionen[57]:

(1) **Niedriges Biasrisiko („low risk of bias“):** Die Studie ist, bezogen auf diese Domäne, vergleichbar mit einem gut durchgeführten RCT.
(2) **Moderates Biasrisiko („moderate risk of bias“):** Für eine nicht-randomisierte Studie ist die Studie, bezogen auf diese Domäne, solide, jedoch kann sie nicht als mit einem gut durchgeführten RCT vergleichbar betrachtet werden.
(3) **Schwerwiegendes Biasrisiko („serious risk of bias“):** Die Studie weist einige bedeutsame Probleme auf.
(4) **Kritisches Biasrisiko („critical risk of bias“):** Die Studie ist zu problematisch, um nützliche Evidenz zu den Effekten der Intervention zu erbringen.
(5) **Keine Informationen („no information“):** Keine Informationen, auf deren Basis ein Urteil über das Biasrisiko für diese Domäne gefällt werden kann.

Im ROBINS-I Tool gibt es neben jeder Biasrisiko-Bewertung **Platz für das Einfügen der Gründe für die jeweilige Bewertung** und von Erläuterungen, deren Dokumentation bei „schwerwiegendem“ oder „kritischem“ Biasrisiko unerlässlich ist.

Beurteilung der prognostizierten Richtung des Bias

Am Ende jeder Domäne (sowie am Ende des Tools für die Gesamtbewertung) gibt es bei ROBINS-I optional die **Möglichkeit, eine Aussage über die erwartete Richtung des Bias** zu treffen. Die Kenntnis von Größe und Richtung eines identifizierten Bias ist für die Beurteilung von Studienergebnissen grundsätzlich sehr wünschenswert, stellt jedoch eine weitaus größere Herausforderung als die Bewertung des Biasrisikos dar. Bei einigen Domänen ist der Bias am einfachsten als **„in Richtung Null“ („towards the null“;** das beobachtete Ergebnis liegt näher an der Null, d.h. an „keinem Effekt“, als das wahre Ergebnis) oder **„weg von der Null“ („away from the null“**; das beobachtete Ergebnis liegt weiter entfernt von der Null, d.h. von „keinem Effekt“, als

der wahre Wert) einzuschätzen. So kann zum Beispiel der Verdacht auf ein selektives Nicht-Berichten von statistisch nicht signifikanten Ergebnissen als Bias „in Richtung Null“ betrachtet werden. Bei anderen Domänen (insbesondere Confounding, Selektionsbias und einigen Arten von Information (oder Measurement) Bias wie zum Beispiel der differentiellen Fehlklassifikation) ist der Bias nicht in Bezug auf die Null zu betrachten, sondern als Vergrößerung oder Verringerung des geschätzten Effektes zugunsten der experimentellen Intervention oder des Vergleichs. Zum Beispiel wäre ein Bias durch Confounding, der den geschätzten Effekt verringert, „gegen Null“, wenn das wahre Risikoverhältnis > 1 wäre, und „weg von der Null“, wenn das Risikoverhältnis < 1 wäre. Wenn Gutachter*innen keine klare Begründung für die Beurteilung der wahrscheinlichen Richtung des Bias haben, sollten sie nicht versuchen, eine solche auf der Basis einer Vermutung vorzunehmen.

Die „Low risk of bias“ Kategorie existiert in erster Linie, um zwischen randomisierten und nicht-randomisierten Studien zu unterscheiden. Diese Unterscheidung betrifft vor allem die „pre-intervention“ und „at-intervention“ Domänen. Zu erwarten ist, dass das Biasrisiko von NRSI aufgrund von Confounding nur in seltenen Fällen als niedrig („low risk of bias“) bewertet wird. Da die Randomisierung nicht gegen einen nach der Intervention auftretenden („post-intervention“) Bias schützt, ist für die entsprechenden („post-intervention“) Domänen eine größere Überlappung zwischen den Bewertungen von randomisierten und nicht-randomisierten Studien zu erwarten. Andere Merkmale randomisierter Studien, wie die Verblindung von Teilnehmer*innen und dem Studienpersonal können gegen Bias nach der Intervention schützen.

Ein Gesamturteil über das Biasrisiko fällen

Die Bewertungen des Biasrisikos der sieben Domänen stellen die Grundlage für ein **Gesamturteil über das Biasrisiko** der begutachteten Studie, bzw. des begutachteten Ergebnisses, dar.

Für das Urteil zum Gesamt-Biasrisiko gibt es bei ROBINS-I fünf Optionen, die den Optionen für die Bewertung des Biasrisikos für die einzelnen Domänen entsprechen:

1. **Niedriges Biasrisiko („low risk of bias“)**
2. **Moderates Biasrisiko („moderate risk of bias“)**
3. **Schwerwiegendes Biasrisiko („serious risk of bias“)**
4. **Kritisches Biasrisiko („critical risk of bias“)**
5. **Keine Informationen („no information“)**

Die **grundlegenden Kriterien für die Ableitung des Gesamt-Biasrisikos** aus den Bewertungen der einzelnen Domänen sind in Tabelle 6 dargestellt.[57]

Tab. 6: Kriterien für das Gesamturteil über das Biasrisiko für das begutachtete Ergebnis[57] (orientiert an Tab. 5)

Option für das Gesamturteil	Aussage	Kriterien
Niedriges Biasrisiko	Die Studie ist vergleichbar mit einem gut durchgeführten RCT.	Das Biasrisiko wurde für alle Domänen als niedrig bewertet.
Moderates Biasrisiko	Für eine nicht-randomisierte Studie ist die Studie solide, jedoch kann sie nicht als mit einem gut durchgeführten RCT vergleichbar betrachtet werden.	Das Biasrisiko wurde für alle Domänen als niedrig oder moderat bewertet.
Schwerwiegendes Biasrisiko	Die Studie weist einige bedeutsame Probleme auf.	Das Biasrisiko wurde für mindestens eine Domäne als schwerwiegend, jedoch für keine Domäne als kritisch bewertet.
Kritisches Biasrisiko	Die Studie ist zu problematisch, um nützliche Evidenz zu den Effekten der Intervention zu erbringen).	Das Biasrisiko wurde für mindestens eine Domäne als kritisch bewertet.
Keine Informationen	Keine Informationen, auf deren Basis ein Urteil über das Biasrisiko gefällt werden kann.	Es gibt keine klaren Hinweise darauf, dass das Biasrisiko schwerwiegend oder kritisch ist, und mindestens eine Schlüssel- Domäne (hierfür ist eine Beurteilung erforderlich) wurde aufgrund fehlender oder unzureichender Informationen mit „Keine Informationen“ bewertet.

Die in der Tabelle dargestellten Kriterien stellen einen grundlegenden Rahmen für das Gesamturteil dar. Wird der begutachteten Studie (bezogen auf das interessierende Ergebnis) in einer Domäne ein bestimmtes Biasrisiko, zum Beispiel ein „schwerwiegendes“ Biasrisiko, attestiert, folgt daraus, dass das Gesamturteil bestenfalls diesem Biasrisiko entsprechen wird. Da es selten vorkommt, dass das Biasrisiko einer NSRI in Bezug auf Confounding als „niedrig“ bewertet wird, ist davon auszugehen, dass das Gesamt-Biasrisko für die meisten NSRIs bestenfalls als „moderat“ zu bewerten ist. In der Praxis können die der einzelnen Domänen mitunter als „additiv“ betrachtet werden, sodass ein „schwerwiegendes“ Biasrisiko in mehreren Domänen zu einem Gesamturteil eines „kritischen“ Biasrisikos, ein „moderates“ Biasrisiko in mehreren Domänen zu einem Gesamturteil eines „schwerwiegenden“ Biasrisikos führen kann.

Für die Darstellung der Gesamtergebnisse der Bewertung mehrerer Studien steht neuerdings mit **robvis**[55] eine web-basierte Anwendung zur Visualisierung von Bewertungen des Biasrisikos in systematischen Reviews zur Verfügung.

Phase 3: Formulierung von Schlussfolgerungen

Das Gesamturteil über das **Biasrisiko** für ein bestimmtes Studienergebnis sollte *mit diesem zusammen* dokumentiert und in den Analysen, der **Diskussion** und den **Schlussfolgerungen** eines

Berichtes (z.B. zu einem systematischen Review oder einer Leitlinie) angemessen berücksichtigt werden. Hilfestellung hierfür geben u.a. die verfügbaren Berichtsqualitäts-Instrumente[56].

Für die **Formulierung von Schlussfolgerungen in Evidenzsynthesen** darüber, inwieweit die beobachteten Effekte der interessierenden Intervention als kausal betrachtet werden können, sollten alle in den betreffenden Review eingeschlossenen Studien miteinander verglichen bzw. gegeneinander kontrastiert werden, sodass ihre Stärken und Schwächen in der Gesamtheit betrachtet werden können. Unterschiedliche Studiendesigns können zu unterschiedlichen Formen von Bias führen, und die „Triangulation“ der Bewertungsergebnisse über alle in einen Review eingeschlossenen Studien hinweg kann dabei helfen zu entscheiden, ob ein Bias unerheblich oder relevant ist.

6.4 Die Newcastle-Ottawa Scale (NOS)

Die Newcastle-Ottawa Scale (NOS) ist ein für die Bewertung von nicht-randomisierten Studien häufig verwendetes Bewertungsinstrument. Die NOS wurde im Rahmen einer fortlaufenden Zusammenarbeit zwischen Wissenschaftler*innen der Universitäten von Newcastle/Australien und Ottawa/Kanada entwickelt.[34] Die NOS kann jedoch allenfalls eingeschränkt empfohlen werden, da sie interne und externe Validität vermischt und aufgrund ihres Alters neuere Erkenntnisse zu Bias nicht berücksichtigt.

Die Bewertung mit der NOS erfolgt anhand eines „Sterne-Systems“, bei dem die zu bewertende Studie in Bezug auf drei übergeordnete Gesichtspunkten betrachtet wird: die Selektion der Studiengruppen, die Vergleichbarkeit der Studiengruppen und die Erfassung der Exposition (in Fall-Kontroll-Studien) bzw. des interessierenden Ergebnisses (in Kohortenstudien). Das Verzerrungspotential einer Fall-Kontrollstudie (s. Abschnitt 6.4.1) oder einer Kohortenstudie (s. Abschnitt 6.4.2) ist insbesondere abhängig von der Strukturgleichheit der beiden zu vergleichenden Gruppen beziehungsweise Kohorten. Die Bewertung von nicht-randomisierten Studien nach der NOS vergibt daher ein Maximum von zwei Sternen für den Aspekt ‚Vergleichbarkeit‘ (siehe II., unten). Eine Studie erhält dagegen nur einen Stern für jeden gelisteten Unterpunkt der Aspekte „Selektion der Studienteilnehmer*innen“ und „Expositionserfassung“ (bei Fall-Kontrollstudien) beziehungsweise „Endpunkterfassung“ (bei Kohortenstudien). Insgesamt kann eine Fall-Kontrollstudie beziehungsweise eine Kohortenstudie neun Sterne erhalten.

6.4.1 Bewertung von Fall-Kontrollstudien

*Selektion der Studienteilnehmer*innen*

1) Wurden die Fälle adäquat definiert?
    a) Ja (unabhängige Validierung, z.B. durch 2 Personen, Patient*innenakte(n), Dokumentation anhand bildgebender Verfahren) *
    b) Nein (z.B. Falldefinition anhand „record linkage“ [z.B. anhand von ICD-Kodierung] oder Angaben der Patient*innen ohne vorhandene Patient*innenakte oder wenn keine Angaben vorhanden sind)

2) Sind die Fälle repräsentativ?
    a) Ja (z.B. konsekutive oder alle Fälle, die in einem bestimmten Bezirk, Einzugsbereich oder einer vordefinierten Zeitspanne aufgetreten sind; randomisiertes Sample der vorliegenden Fälle) *

b) Nein (Potential für selection bias wahrscheinlich oder wenn keine Angaben vorhanden sind)

3) Sind die Kontrollen repräsentativ, erfolgte eine adäquate Auswahl der Kontrollen?
   a) Ja (die Kontrollen stammen aus einer vergleichbaren Population wie die Fälle [jedoch ist der Endpunkt bei den Kontrollen nicht aufgetreten]) *
   b) Nein (Kontrollen aus dem Krankenhaus (Patient*innen) oder wenn keine Angaben vorhanden sind)

4) Wurden die Kontrollen adäquat definiert?
   a) Ja (der Endpunkt [z.B. Krebserkrankung], der bei den Fällen zum ersten Mal eingetreten ist, darf bei den Kontrollen bei der Endpunkterhebung nicht vorhanden sein) *
   b) Nein (keine Angabe vorhanden, ob der Endpunkt bereits in der Kontrollgruppe aufgetreten ist)

Vergleichbarkeit

1) Ist die Vergleichbarkeit der Fälle und Kontrollen gegeben?
   a) Ja (die Fälle und Kontrollen wurden bereits bei der Auswahl ziemlich genau aufeinander abgestimmt [Matching] oder die Studie kontrolliert die wichtigsten Störfaktoren in der Datenanalyse [z.B. Alter, Geschlecht, Häufigkeit der Medikamenteneinnahme, Ko-Morbidität]) * (an dieser Stelle können maximal 2 Sterne vergeben werden: Ein Stern, wenn für den wichtigsten Störfaktor kontrolliert wurde, und ein weiterer Stern, wenn für einen weiteren entscheidenden Störfaktor kontrolliert wurde [siehe Abschnitt 6.4.3: Tabellenvorlage])
   b) Nein (die Aussage: „no differences between groups or that differences were not statistically significant“ sind nicht ausreichend, um von einer Vergleichbarkeit auszugehen)

Expositionserfassung

1) Erfolgte eine valide Erfassung der Exposition?
   a) Ja (z.B. anhand der Patient*innenakte) *

b) Ja (durch ein Interview z.B. der Kolleg*innen oder Freund*innen, die im Hinblick auf den Fall-Kontroll-Status verblindet waren) *

c) Nein (durch ein Interview z.B. der Kolleg*innen oder Freund*innen, die im Hinblick auf den Fall-Kontroll-Status nicht verblindet waren)

d) Nein (Angaben der Patient*innen ohne vorhandene Patient*innenakte oder wenn keine Angaben vorhanden sind)

2) Erfolgte die Erfassung der Fälle und Kontrollen identisch?

a) Ja (z.B. durch identische standardisierte diagnostische Methoden) *

b) Nein

3) Kann die „Non-Response-Rate“ als valide betrachtet werden?

a) Ja (für beide Gruppen liegt die Rate vor) *

b) Nein (keine Ereignisraten angegeben)

c) Nein (unterschiedliche Ereignisraten, in der Studie wird jedoch nicht weiter darauf eingegangen)

6.4.2 Bewertung von Kohortenstudien

Selektion der Studienteilnehmer*innen

1) Ist die exponierte Kohorte repräsentativ für die zu untersuchende Intervention/Exposition?

a. Ja und Wahrscheinlich Ja (sollen z.B. unerwünschte Wirkungen der Östrogen-Exposition in der Post-Menopause untersucht werden, muss eine Frauen-Kohorte ausgewählt werden, die repräsentativ für die Einnahme dieser Östrogene ist. Frauen, die z.B. einer ethnischen Minderheit angehören, wären in diesem Fall nicht repräsentativ). *

b. Nein (selektiertes Sample wie z.B. freiwillig oder Pflegefachpersonen oder wenn keine Angaben vorhanden sind)

2) Ist die nicht-exponierte Kohorte repräsentativ, wurde sie adäquat ausgewählt?

a. Ja (die nicht-exponierte Kohorte stammt aus einer vergleichbaren Grundgesamtheit wie die exponierte Kohorte) *

b. Nein (die nicht-exponierte Kohorte stammt aus einer anderen Grundgesamtheit, z.B. aus dem Krankenhaus [Patient*innen] oder wenn keine Angaben vorhanden sind)

3) Erfolgte eine valide Erfassung der Exposition?
   a. Ja (z.B. anhand der Patient*innenakte) *
   b. Ja (durch ein strukturiertes Interview) *
   c. Nein (narrative Angaben der Studienteilnehmer*innen [ohne vorhandene Patient*innenakte] oder wenn keine Angaben vorhanden sind)

4) Ist es wahrscheinlich, dass der gemessene Endpunkt nicht zu Studienbeginn vorhanden war?
   a. Ja (z.B. diagnostische Maßnahmen erfolgten)
   b. Nein (keine Angabe vorhanden, ob der Endpunkt bereits am Anfang der Studie vorhanden war)

Vergleichbarkeit

1) Ist die Vergleichbarkeit der exponierten und nicht-exponierten Kohorte gegeben?
   a. Ja (die exponierte und nicht-exponierte Kohorte wurden bereits bei der Auswahl ziemlich genau aufeinander abgestimmt [Matching] oder die Studie kontrolliert für die wichtigsten Störfaktoren in der Datenanalyse [z.B. Alter, Geschlecht, Häufigkeit der Medikamenteneinnahme, Ko-Morbidität, ethnische Herkunft]) * (an dieser Stelle können maximal 2 Sterne vergeben werden: Ein Stern, wenn für den wichtigsten Störfaktor kontrolliert wurde, und ein weiterer Stern, wenn für einen weiteren entscheidenden Störfaktor kontrolliert wurde [siehe Abschnitt 6.3: Tabellenvorlage])
   b. Nein (die Aussage: „no differences between groups or that differences were not statistically significant“ sind nicht ausreichend, um von einer Vergleichbarkeit auszugehen)

Endpunkterfassung

1) Erfolgte eine valide Erfassung der Endpunkte?
   a. Ja (unabhängige oder verblindetete Erhebung, z.B. durch 2 Personen, Patient*innenakte(n), Dokumentation anhand bildgebender Verfahren) *
   b. Ja (anhand „record linkage“ [z.B. anhand von ICD-Kodierung] in der Patient*innenakte) *
   c. Nein (Angaben der/des Patient*in [ohne dass eine Patient*innenakte verfügbar ist] oder wenn keine Angaben vorhanden sind)

2) Konnte in der Beobachtungszeit der Endpunkt überhaupt auftreten?
   a. Ja (a priori muss eine ausreichende Beobachtungszeit festgelegt werden, damit sichergestellt ist, dass der Endpunkt auch in dieser Zeit auftreten kann, z.B. sollte bei der Bewertung der Verträglichkeit von Brustimplantaten ein Minimum von 5 Jahren angesetzt werden) *
   b. Nein

3) Wurden fehlende Daten adäquat berücksichtigt? (siehe Abschnitt 5.1.1: Domänen der RoB Bewertung/ Fehlende Daten bei der Endpunkterhebung)
   a. Ja (es liegen keine fehlenden Daten bei der Endpunkterhebung vor) *
   b. Ja (der Einfluss der fehlenden Daten auf den Effekt ist statistisch und/oder klinisch nicht relevant und/oder steht wahrscheinlich nicht in Zusammenhang mit der Exposition) *
   c. Nein (der Einfluss der fehlenden Daten auf den Effekt ist statistisch und/oder klinisch relevant und/oder steht wahrscheinlich in Zusammenhang mit der Exposition)
   d. Nein (keine Angaben dazu vorhanden)

Tabelle 6.4.3 zeigt eine Tabellenvorlage für die Bewertungsergebnisse einer Bewertung mit der NOS.

6.4.3 NOS Tabellenvorlage

Tab. 7: NOS RoB Tabelle für nicht-randomisierte Studien

Fall-Kontroll-Studien			
	Selektion	Vergleichbarkeit	Expositionserfahrung
Studie 1	****	**	***
Studie 2	**	*	*
Kohortenstudien			
	Selektion	Vergleichbarkeit	Expositionserfahrung
Studie 1	*	**	***
Studie 2	***	**	**

7 BEWERTUNG DES BIAS-RISIKOS IN NICHT-VERGLEICHENDEN STUDIEN

Unter nicht-vergleichende Studien fallen insbesondere einarmige Kohortenstudien ohne klar definierte Vergleichsgruppe und Fallserien (beziehungsweise Verlaufsbeobachtungen). Da die Bewertung des Nutzen- und Schadensverhältnisses einer Intervention eine Kontrollgruppe erfordert, lässt sich aus nicht-vergleichenden Studien in der Regel keine Aussage zur Wirksamkeit einer Intervention ableiten. Nicht-vergleichenden Studien sollten als erster Informationsgewinn, vor allem im Hinblick auf potenzielle Schäden, zu einer Intervention betrachtet werden. Ausnahmen dabei bilden Interventionen bei Krankheitsbildern, die dramatische Effekte zeigen, wie zum Beispiel die Substitution von Insulin bei Patient*innen mit hyperglykämischer Krise bei Diabetes mellitus Typ 1. Dies setzt jedoch voraus, dass der (natürliche) Verlauf der Erkrankung ohne die Intervention hinreichend sicher bekannt ist (ähnlich einem indirekten Vergleich von Fallserien). Nicht-vergleichende Querschnittstudien sind zum Beispiel für die Schätzung von Krankheitsprävalenzen geeignet, nicht jedoch für die Ableitung von Wirksamkeit. Für nicht-vergleichende Studien liegen in der Literatur (noch) keine eindeutigen Kriterien vor, nach denen das Verzerrungspotenzial auf Studienebene beurteilt werden soll. Aus methodischer Sicht kann jedoch festgehalten werden, dass auch bei nicht-vergleichenden Studien folgende Merkmale das Vertrauen in die Studienergebnisse erhöhen:

- prospektive Planung mit Protokoll, in dem Einschlusskriterien und Interventionen sowie interessierende Endpunkte hinterlegt sind
- konsekutiver Patient*inneneinschluss
- transparentes, nicht-selektives Berichten in Bezug auf Patient*innencharakteristika, Intervention und Ergebnis.

8 QUELLEN

1. Sackett DL, Rosenberg WM. The need for evidence-based medicine. *J R Soc Med*. 1995;88(11):620-624.
2. Cochrane Deutschland Stiftung, Institut für Evidenz in der Medizin, Institut für Medizinische Biometrie und Statistik, Freiburg, Arbeitsgemeinschaft der Wissenschaftlichen Medizinischen Fachgesellschaften - Institut für Medizinisches Wissensmanagement, Ärztliches Zentrum für Qualität in der Medizin. Manual Systematische Recherche für Evidenzsynthesen und Leitlinien. 2.1 Auflage (14.12.2020). Verfügbar: Cochrane Deutschland: https://www.cochrane.de/de/literaturrecherche; AWMF: https://www.awmf.org/leitlinien/awmf-regelwerk/ll-entwicklung.html; ÄZQ: https://www.aezq.de/aezq/publikationen/azq-partner#literaturrecherche. doi: 10.6094/UNIFR/174468.
3. Arbeitsgemeinschaft der Wissenschaftlichen Medizinischen Fachgesellschaften (AWMF) - Ständige Kommission Leitlinien. AWMF-Regelwerk “Leitlinien”. 2. Auflage 2020. https://www.awmf.org/leitlinien/awmf-regelwerk.html. [Zugriff 18.03.2021].
4. Juni P. Systematic reviews in health care: Assessing the quality of controlled clinical trials. *BMJ*. 2001;323(7303):42-46. doi:10.1136/bmj.323.7303.42.
5. Guyatt GH, Oxman AD, Kunz R, et al. GRADE guidelines: 8. Rating the quality of evidence—indirectness. *J Clin Epidemiol*. 2011;64(12):1303-1310. doi:10.1016/j.jclinepi.2011.04.014.
6. Ciani O, Buyse M, Garside R, et al. Comparison of treatment effect sizes associated with surrogate and final patient relevant outcomes in randomised controlled trials: meta-epidemiological study. *BMJ*. 2013;346(jan29 1):f457-f457. doi:10.1136/bmj.f457.
7. Higgins JPT, Altman DG, Gotzsche PC, et al. The Cochrane Collaboration’s tool for assessing risk of bias in randomised trials. *BMJ*. 2011;343(oct18 2):d5928-d5928. doi:10.1136/bmj.d5928.
8. Deutsches Register Klinischer Studien (DRKS). Website. https://www.drks.de/drks_web/. [Zugriff 18.03.2021].
9. ClinicalTrials.gov. Website. https://www.clinicaltrials.gov/. [Zugriff 18.03.2021].
10. Deutsche Forschungsgemeinschaft (DFG), Bundeministerium für Bildung und Forschung. Grundsätze und Verantwortlichkeiten bei der Durchführung klinischer Studien. https://www.dfg.de/download/pdf/foerderung/programme/klinische_studien/klinische_studien_grundsaetze_verantwortlichkeiten.pdf. [Zugriff 18.03.2021].
11. ICH Guidelines - The International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use. https://www.ich.org/. [Zugriff 18.03.2021].
12. Deutsche Gesellschaft für Epidemiologie. http://www.gesundheitsforschung-bmbf.de/_media/Empfehlungen_GEP.pdf. [Zugriff 18.03.2021].
13. Blümle A, von Elm E, Antes G, Meerpohl JJ. Messung und Bewertung der Studienqualität und Berichtsqualität. *Z Evid Fortbild Qual Gesundhwes*. 2014;108(8-9):495-503. doi:10.1016/j.zefq.2014.09.022.
14. Consort Statement. Website. http://www.consort-statement.org/. [Zugriff 18.03.2021].
15. Schulz K, Altman D, Moher D. CONSORT 2010 statement: Updated guidelines for reporting parallel group randomised trials. *J Pharmacol Pharmacother*. 2010;1(2):100. doi:10.4103/0976-500X.72352.
16. Turner L, Shamseer L, Altman DG, et al. Consolidated standards of reporting trials (CONSORT) and the completeness of reporting of randomised controlled trials (RCTs) published in medical journals. *Cochrane Database Syst Rev*. 2012;11:MR000030. doi:10.1002/14651858.MR000030.pub2.
17. Equator Network. Equator Network Webpage. https://www.equator-network.org/. [Zugriff 18.03.2021].
18. Langer G, Meerpohl JJ, Perleth M, Gartlehner G, Kaminski-Hartenthaler A, Schünemann H. GRADE-Leitlinien: 1. Einführung – GRADE-Evidenzprofile und Summary-of-Findings-Tabellen. *Z Evid Fortbild Qual Gesundhwes*. 2012;106(5):357-368. doi:10.1016/j.zefq.2012.05.017.
19. Centre for Evidence-Based Medicine; University of Oxford. Catalogue of Bias. https://catalogofbias.org/. [Zugriff 18.03.2021].
20. Savović J, Jones HE, Altman DG, et al. Influence of Reported Study Design Characteristics on Intervention Effect Estimates From Randomized, Controlled Trials. *Ann Intern Med*. 2012;157(6):429. doi:10.7326/0003-4819-157-6-201209180-00537.

21. Hemkens LG, Contopoulos-Ioannidis DG, Ioannidis JPA. Agreement of treatment effects for mortality from routinely collected data and subsequent randomized trials: meta-epidemiological survey. *BMJ*. 2016;352:i493. doi:10.1136/bmj.i493.

22. Correction notice to paper “Agreement of treatment effects for mortality from routinely collected data and subsequent randomized trials: meta-epidemiological survey.” *BMJ*. August 2018:k3210. doi:10.1136/bmj.k3210.

23. 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.

24. Otto C, Schiffer G, Tjardes T, Kunter H, Eysel P, Paffrath T. Blood loss and operative duration using monopolar electrosurgery versus ultrasound scissors for surgical preparation during thoracoscopic ventral spondylodesis: results of a randomized, blinded, controlled trial. *Eur Spine J*. 2014;23(8):1783-1790. doi:10.1007/s00586-014-3303-1.

25. Wood L, Egger M, Gluud LL, et al. Empirical evidence of bias in treatment effect estimates in controlled trials with different interventions and outcomes: meta-epidemiological study. *BMJ*. 2008;336(7644):601-605. doi:10.1136/bmj.39465.451748.

26. Bero LA. Why the Cochrane Risk of Bias Tool Should Include Funding Source as a Standard Item. In: Tovey D, ed. *Cochrane Database of Systematic Reviews*. Vol 12. 2014/02/28. Chichester, UK: John Wiley & Sons, Ltd; 2013:ED000075. doi:10.1002/14651858.ED000075.

27. Sterne JA. Why the Cochrane Risk of Bias Tool Should not Include Funding Source as a Standard Item. In: Tovey D, ed. *Cochrane Database of Systematic Reviews*. Vol 12. 2014/02/28. Chichester, UK: John Wiley & Sons, Ltd; 2013:ED000076. doi:10.1002/14651858.ED000076.

28. Lundh A, Lexchin J, Mintzes B, Schroll JB, Bero L. Industry sponsorship and research outcome. *Cochrane Database Syst Rev*. 2017. doi:10.1002/14651858.MR000033.pub3.

29. Boutron, I; Page, MJ; Higgins, JPT; Altman, DG; Lundh, A; Hrobjartsson A on behalf of the CBMG. Chapter 7: Considering bias and conflicts of interest among the included studies. In: *Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, Welch VA (Editors). Cochrane Handbook for Systematic Reviews of Interventions. 2nd Ed. Wiley Blackwell*. 2019.

30. Dreier M, Borutta B, Stahmeyer J, Krauth C, Walter U. Vergleich von Bewertungsinstrumenten für die Studienqualität von Primär- und Sekundärstudien zur Verwendung für HTA-Berichte im deutschsprachigen Raum. *Schriftenr Heal Technol Assement*. 2010;Bd. 102. doi:10.3205/hta000085L.

31. Moher D, Cook DJ, Jadad AR, et al. Assessing the quality of reports of randomised trials: implications for the conduct of meta-analyses. *Heal Technol Assess*. 1999;3(12):i-iv, 1-98.

32. Juni P, Witschi A, Bloch R, Egger M. The hazards of scoring the quality of clinical trials for meta-analysis. *Jama*. 1999;282(11):1054-1060.

33. Jadad AR, Moore RA, Carroll D, et al. Assessing the quality of reports of randomized clinical trials: is blinding necessary? *Control Clin Trials*. 1996;17(1):1-12.

34. Wells, G.A.; Shea, B.; O’Connell, D.; Peterson, J.; Welch, V.; Losos, M.; Tugwell P. The Newcastle-Ottawa Scale (NOS) for assessing the quality of nonrandomised studies in meta-analyses. http://www.ohri.ca/programs/clinical_epidemiology/oxford.asp. [Zugriff 18.03.2021].

35. Scottish Intercollegiate Guidelines Network. Checklists. https://www.sign.ac.uk/what-we-do/methodology/checklists/. [Zugriff 18.03.2021].

36. Lundh A, Gøtzsche PC. Recommendations by Cochrane Review Groups for assessment of the risk of bias in studies. *BMC Med Res Methodol*. 2008;8(1):22. doi:10.1186/1471-2288-8-22.

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

38. Sterne JA, Hernán MA, Reeves BC, et al. ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. *BMJ*. 2016:i4919. doi:10.1136/bmj.i4919.

39. Higgins, JPT, Savovic, J; Page, MJ; Sterne J (eds.) on behalf of the R 2 DG. RoB 2 - Full guidance document (Version 22 August 2019) (riskofbias.info). https://drive.google.com/file/d/19R9savfPdCHC8XLz2iiMvL_71lPJERWK/view. [Zugriff 18.03.2021].

40. Higgins, J.P.T.; Savovic, J.; Page, M.J.; Elbers, R.G.; Sterne JAC. Chapter 8: Assessing risk of bias in a randomized trial. In: *Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, Welch VA (Editors). Cochrane Handbook for Systematic Reviews of Interventions. 2nd Ed. Wiley Blackwell*. 2019.

41. ROB-ME - A tool for assessing Risk Of Bias due to Missing Evidence in a synthesis. https://www.riskofbias.info/welcome/rob-me-tool. [Zugriff 18.03.2021].

42. ROB-ME Development Group. *ROB_ME Detailed Guidance (Version 24 October 2020) (Riskofbias.Info)*. https://drive.google.com/file/d/1BaF3lZ6j1ZIx208gsoYab8uGzk6zlqvw/view. [Zugriff 18.03.2021].

43. Jørgensen L, Paludan-Müller AS, Laursen DRT, et al. Evaluation of the Cochrane tool for assessing risk of bias in randomized clinical trials: overview of published comments and analysis of user practice in Cochrane and non-Cochrane reviews. *Syst Rev*. 2016;5(1):80. doi:10.1186/s13643-016-0259-8.

44. Williams CA, Wadey C, Pieles G, Stuart G, Taylor RS, Long L. Physical activity interventions for people with congenital heart disease. *Cochrane database Syst Rev*. 2020;10:CD013400. doi:10.1002/14651858.CD013400.pub2.

45. RoB 2 - A revised Cochrane risk of bias tool for randomized trials. Website. (riskofbias.info). https://sites.google.com/site/riskofbiastool/welcome/rob-2-0-tool?authuser=0. [Zugriff 18.03.2021].

46. RoB 2 - Current version of RoB 2 (22. August 2019). Cribsheet. (riskofbias.info). https://sites.google.com/site/riskofbiastool/welcome/rob-2-0-tool/current-version-of-rob-2. [Zugriff 18.03.2021].

47. RoB 2 - Revised Cochrane risk-of-bias tool for randomized trials (RoB 2).Template for completion (Version 22 August 2019) (riskofbias.info). https://drive.google.com/file/d/18Zks7k4kxhbUUlbZ51Ya5xYa3p3ECQV0/view [Zugriff 18.03.2021].

48. RoB 2: Excel tool to implement RoB 2 (riskofbias.info). https://drive.google.com/file/d/1uwAVr-wKE3elEzcsVOBGLzJOVhbpf321/view. [Zugriff 18.03.2021].

49. Cochrane Methods. https://methods.cochrane.org/risk-bias-2. [Zugriff 18.03.2021].

50. Higgins, J.P.T.; Eldridge, S.; Tianjing L. Chapter 23: Including variants on randomized trials. In: *Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, Welch VA (Editors). Cochrane Handbook for Systematic Reviews of Interventions. 2nd Ed. Wiley Blackwell*. 2019.

51. Hernán MA, Robins JM. Per-Protocol Analyses of Pragmatic Trials. *N Engl J Med*. 2017;377(14):1391-1398. doi:10.1056/NEJMsm1605385.

52. Cochrane RevMan (Review Manager). Cochrane Training. https://training.cochrane.org/online-learning/core-software-cochrane-reviews/revman. [Zugriff 18.03.2021].

53. Gartlehner G, Nussbaumer-Streit B, Gaynes BN, et al. Second-generation antidepressants for preventing seasonal affective disorder in adults. *Cochrane Database Syst Rev*. 2019. doi:10.1002/14651858.CD011268.pub3.

54. Cochrane RevMan Web (Review Manager Web- Version). Website. https://revman.cochrane.org/#/myReviews. [Zugriff 18.03.2021].

55. Cochrane robvis. Visualization tool. (riskofbias.info). https://www.riskofbias.info/welcome/robvis-visualization-tool. [Zugriff 18.03.2021].

56. Equator Network. Website. https://www.equator-network.org/. [Zugriff 18.03.2021].

57. Sterne, J.A.C.; Higgins, J.P.T.; Elbers, R.G.; Reeves BC. (eds.). on behalf of thedevelopment group forROBINS-I. ROBINS-I - Detailed guidance document (zuletzt aktualisiert 20. Oktober 2016) (riskofbias.info). http://www.bristol.ac.uk/media-library/sites/social-community-medicine/images/centres/cresyda/ROBINS-I_detailed_guidance.pdf. [Zugriff 18.03.2021].

58. Sterne, J.A.C.; Hernán, M.A.; McAleenan, A.; Reeves, B.C.; Higgins JPT. Chapter 25: Assessing risk of bias in a non-randomized study. In: *Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, Welch VA (Editors). Cochrane Handbook for Systematic Reviews of Interventions. 2nd Ed. Wiley Blackwell*. 2019.

59. Hernán MA, Robins JM. Using Big Data to Emulate a Target Trial When a Randomized Trial Is Not Available: Table 1. *Am J Epidemiol*. 2016;183(8):758-764. doi:10.1093/aje/kwv254.

60. Velie EM, Shaw GM. Impact of Prenatal Diagnosis and Elective Termination on Prevalence and Risk Estimates of Neural Tube Defects in California, 1989-1991. *Am J Epidemiol*. 1996;144(5):473-479. doi:10.1093/oxfordjournals.aje.a008953.

61. Cochrane Bias Methods Group. Website. https://methods.cochrane.org/bias/home [Zugriff 18.03.2021].

62. Cochrane (NRS) Non-Randomized Studies for Interventions Methods Group. Website.

https://methods.cochrane.org/nrsi/welcome. [Zugriff 18.03.2021].

63. Centre for Research Synthesis and Decision Analysis - University of Bristol. Archived tool: A Cochrane Risk Of Bias Assessment Tool for Non-Randomized Studies of Interventions (ACROBAT-NRSI). https://www.bristol.ac.uk/population-health-sciences/centres/cresyda/barr/riskofbias/robins-i/acrobat-nrsi/. [Zugriff 18.03.2021].

64. ROBINS-I - Risk Of Bias In Non-randomized Studies of Interventions. Website (riskofbias.info). https://www.riskofbias.info/welcome/home. [Zugriff 18.03.2021].

65. ROBINS-I - Version mit Erläuterungen (Version 01. August 2016). (riskofbias.info). https://www.riskofbias.info/welcome/home/current-version-of-robins-i/robins-i-tool-2016. [Zugriff 18.03.2021].

66. ROBINS-I - Template (Version 19 September 2016) (riskofbias.info). https://www.riskofbias.info/welcome/home/current-version-of-robins-i/robins-i-template-2016. [Zugriff 18.03.2021].

67. riskofbias.info. *Risk Of Bias In Non-Randomized Studies of Interventions (ROBINS-I): Detailed Guidance (Version 20 October 2016)*. https://www.riskofbias.info/welcome/home/current-version-of-robins-i/robins-i-detailed-guidance-2016. [Zugriff 28.04.2021].

68. Jeyaraman MM, Rabbani R, Copstein L, et al. Methodologically rigorous risk of bias tools for nonrandomized studies had low reliability and high evaluator burden. *J Clin Epidemiol*. 2020;128:140-147. doi:10.1016/j.jclinepi.2020.09.033.

9 WEITERFÜHRENDE INFORMATIONEN UND PRAXISHILFEN

Leitlinien zur Verbesserung der Berichterstattung verschiedener Studientypen einschließlich Systematischer Übersichtsarbeiten sind abrufbar unter: **https://www.equator-network.org/** (siehe Abbildung 11).

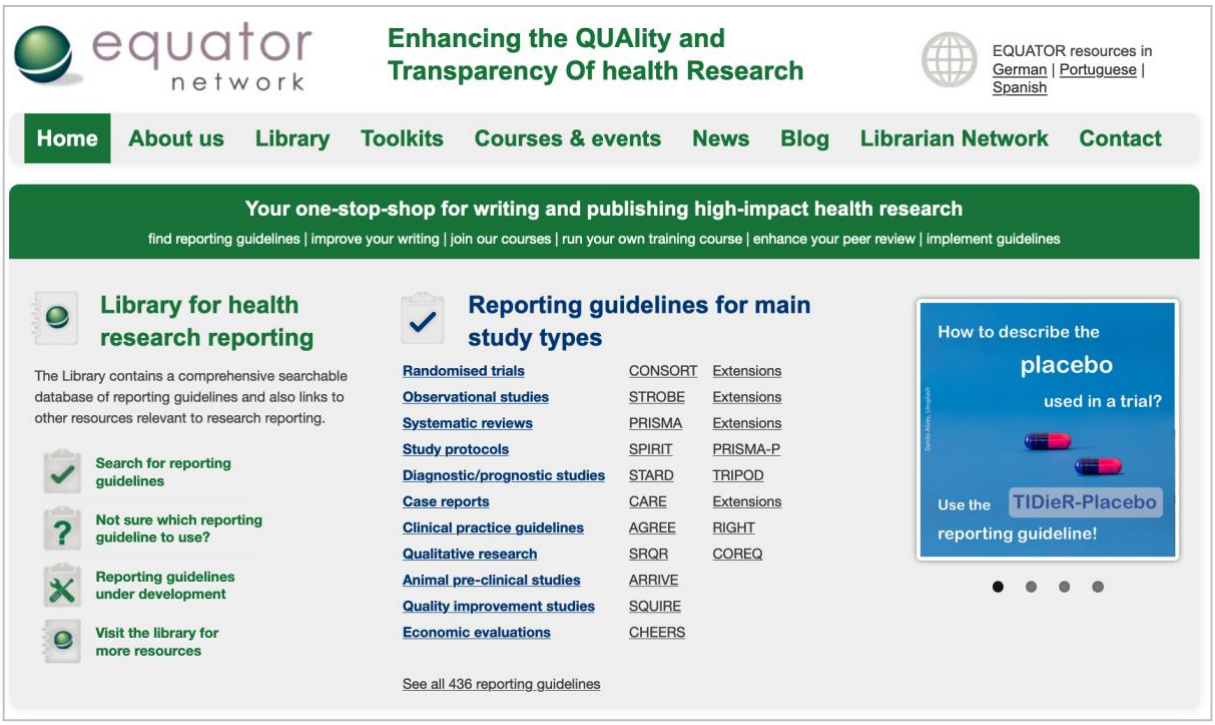

Abb. 11: Screenshot der Equator Webseite (https://www.equator-network.org/).