



Serie: Das Grading of Recommendations, Assessment, Development and Evaluation (GRADE) System

## GRADE-Leitlinien: 19. Bewertung der Vertrauenswürdigkeit der Evidenz für die Bedeutung von Endpunkten oder Werten und Präferenzen – Risiko für Bias und Indirektheit<sup>☆</sup>



*GRADE Guidelines: 19. Assessing the certainty of evidence in the importance of outcomes or values and preferences: Risk of bias and indirectness*



Laura Kaiser<sup>a,b,\*</sup>, Markus Hübscher<sup>a</sup>, Olesja Rissling<sup>a</sup>, Sandra Schulz<sup>a</sup>, Gero Langer<sup>c</sup>, Jörg Meerpohl<sup>d,e</sup>, Lukas Schwingshackl<sup>d</sup>

<sup>a</sup> Abteilung Fachberatung Medizin, Gemeinsamer Bundesausschuss, Berlin, Deutschland

<sup>b</sup> Universität Witten/Herdecke, Deutschland

<sup>c</sup> Institut für Gesundheits- und Pflegewissenschaft, Medizinische Fakultät der Martin-Luther-Universität Halle-Wittenberg, Halle (Saale), Deutschland

<sup>d</sup> Institut für Evidenz in der Medizin, Universitätsklinikum und Medizinische Fakultät, Universität Freiburg, Freiburg, Deutschland

<sup>e</sup> Cochrane Deutschland, Cochrane Deutschland Stiftung, Freiburg, Deutschland

### ARTIKEL INFO

#### Artikel-Historie:

Eingegangen: 3. November 2020

Akzeptiert: 11. November 2020

Online gestellt: 16. Januar 2021

#### Schlüsselwörter:

GRADE

Qualität der Evidenz

Bedeutung von Endpunkten

Werte und Präferenzen

Risiko für Bias

Indirektheit

### ZUSAMMENFASSUNG

**Ziel:** Die Arbeitsgruppe *Grading of Recommendations Assessment, Development, and Evaluation* (GRADE) definiert Patient\*innenwerte und -präferenzen als die relative Bedeutung, die Patient\*innen gesundheitsbezogenen Endpunkten beimessen. Wir stellen GRADE-Empfehlungen zur Einschätzung des Risikos für Bias und der Indirektheit bereit. Diese können im Rahmen der Bewertung der Vertrauenswürdigkeit der Evidenz zur relativen Bedeutung von Endpunkten genutzt werden.

**Studiendesign und Setting:** Die GRADE-Domänen wurden in mehreren systematischen Übersichtsarbeiten angewendet, um die Vertrauenswürdigkeit der Evidenz zur Bedeutung von Endpunkten zu bewerten. Dabei wurden Entwürfe der GRADE-Empfehlungen iterativ geprüft und GRADE-Mitgliedern sowie andere Interessengruppen um Feedback gebeten.

**Ergebnisse:** Dies ist der erste von zwei Artikeln. Ein Evidenzkörper, der die Bedeutung von Endpunkten adressiert, wird bezüglich seiner Vertrauenswürdigkeit zunächst als „hoch“ eingestuft. Risiko für Bias, Indirektheit, Inkonsistenz, unzureichende Präzision oder Publikationsbias führen zu einer Herabstufung auf moderate, niedrige oder sehr niedrige Vertrauenswürdigkeit. Zur Einschätzung des Risikos für Bias schlagen wir die Subdomänen Auswahl der Studienpopulation, fehlende Daten, Art des Messinstruments und Beeinflussbarkeit durch Störfaktoren (Confounding) vor. Für jede dieser Subdomänen haben wir Items entwickelt. Die Population, die Intervention, die Vergleichsintervention und die Endpunkte, die mit der Evidenz zusammenhängen, bestimmen den Grad der Indirektheit.

**Fazit:** Dieser Artikel gibt Empfehlungen und Beispiele für die Einschätzung des Risikos für Bias und der Indirektheit eines Evidenzkörpers, der die Bedeutung von Endpunkten zusammenfasst.

<sup>☆</sup> Übersetzt und adaptiert von: Zhang Y, Alonso-Coello P, Guyatt GH, Yepes-Núñez JJ, Akl EA, Hazlewood G, Pardo-Hernandez H, Etzeandia-Ikobaltzeta I, Qaseem A, Williams Jr. JW, Tugwell P, Flottorp S, Chang Y, Zhang Y, Mustafa RA, Rojas MX, Schünemann HJ. GRADE Guidelines: 19. Assessing the certainty of evidence in the importance of outcomes or values and preferences – Risk of bias and indirectness. *J Clin Epidemiol* 2019; 111: 94–104. <https://doi.org/10.1016/j.jclinepi.2018.01.013>.

\* Korrespondenzadresse. Laura Kaiser, Abteilung Fachberatung Medizin, Gemeinsamer Bundesausschuss, Gutenbergstraße 13, 10587 Berlin, Deutschland.  
E-mail: [laura.kaiser@g-ba.de](mailto:laura.kaiser@g-ba.de) (L. Kaiser).

## ARTICLE INFO

## Article History:

Received: 3 November 2020

Accepted: 11 November 2020

Available online: 16 January 2021

## Keywords:

GRADE

Quality of evidence

Importance of outcomes

Value and preference

Risk of bias

Indirectness

## ABSTRACT

**Objectives:** The Grading of Recommendations Assessment, Development, and Evaluation (GRADE) working group defines patient values and preferences as the relative importance patients place on the main health outcomes. We provide GRADE guidance for assessing the risk of bias and indirectness domains for certainty of evidence about the relative importance of outcomes.

**Study Design and Setting:** We applied the GRADE domains to rate the certainty of evidence in the importance of outcomes to several systematic reviews, iteratively reviewed draft guidance and consulted GRADE members and other stakeholders for feedback.

**Results:** This is the first of two articles. A body of evidence addressing the importance of outcomes starts at “high certainty”; concerns with risk of bias, indirectness, inconsistency, imprecision, and publication bias lead to downgrading to moderate, low, or very low certainty. We propose the following subdomains of risk of bias: selection of the study population, missing data, the type of measurement instrument, and confounding; we have developed items for each subdomain. The population, intervention, comparison, and outcome elements associated with the evidence determine the degree of indirectness.

**Conclusion:** This article provides guidance and examples for rating the risk of bias and indirectness for a body of evidence summarizing the importance of outcomes.

## Was ist neu?

## Kernaussagen

- Die Domänen Risiko für Bias, Indirektheit, Inkonsistenz, unzureichende Präzision und Publikationsbias sind bei der Herabstufung der Vertrauenswürdigkeit von Evidenz zu berücksichtigen.

## Wie wird Bekanntes ergänzt?

- Die Subdomänen des Risikos für Bias umfassen die Auswahl von Studienteilnehmer\*innen, fehlende Daten, Messinstrumente und Datenanalyse. Das Risiko für Bias der Evidenz zur relativen Bedeutung von Endpunkten wird als nicht schwerwiegend, schwerwiegend oder sehr schwerwiegend bewertet, je nachdem wie groß der Anteil der Studien mit Risiko für Bias im Evidenzkörper ist. Die Bestimmung des Ausmaßes der Indirektheit erfolgt unter Berücksichtigung methodischer Aspekte und im Hinblick auf die Elemente des PICO Modells (Population, Intervention, Vergleichsintervention, Endpunkte) der bewerteten Evidenz.

## Was sollte sich jetzt ändern?

- Autor\*innen von Wissenssynthesen zur relativen Bedeutung von Endpunkten, inklusive Nutzen und Werte, sollten die Vertrauenswürdigkeit der zugrundeliegenden Evidenz bewerten.

## Einleitung

Entscheidungen im Gesundheitswesen erfordern nicht nur Evidenz für die Effekte von Interventionen (z.B. Reduktion oder Anstieg des absoluten Risikos eines Endpunktes in einer bestimmten Population durch eine spezifische Intervention gegenüber einer Kontrollintervention), sondern auch Kenntnisse über die relative Bedeutung oder Wichtigkeit der Endpunkte, die durch die Intervention(en) beeinflusst werden soll(en) (siehe Box 1 für ein hypothetisches Beispiel).

Die Berücksichtigung dieser Konzepte bei Entscheidungen über die gesundheitliche Versorgung geschieht häufig in Bezug auf Werte und Präferenzen [1–8]. Im Kontext der Entscheidungsfindung reflektieren Werte und Präferenzen die relative Bedeutung, die Betroffene den potentiellen Folgen einer Entscheidung und den damit verbundenen Endpunkten zusprechen (z. B. die Entscheidung für eine Behandlung oder Untersuchung) [1–8].

Methoden zur Bestimmung der relativen Bedeutung von Endpunkten umfassen a) die direkte Bestimmung des Nutzens (*utility*) oder Wertes (*value*) von Endpunkten, z. B. mittels Standardlotterie

(*standard gamble*) [9–11], der Methode der zeitlichen Abwägung (*time trade-off*) [12,13] oder Ratingskalen [10,13,14]. Die Verbundmessung (*Conjoint-Analyse*) ist ein weiteres Verfahren zur Ermittlung von Nutzen und Bedeutung von Endpunkten. Sie umfasst das diskrete Entscheidungsexperiment (*discrete choice experiment*) [15,16], die kontingente Bewertung und Zahlungsbereitschaft (*contingent valuation and willingness to pay*) [17], das Abwägen von Eintrittswahrscheinlichkeiten (*probability trade-off*) [18,19] sowie den paarweisen Vergleich (*paired comparison*). Die Methoden zur Bestimmung der relativen Bedeutung von Endpunkten umfassen darüber hinaus b) die indirekte Bestimmung des Nutzens durch Instrumente wie den EuroQol-Fragebogen (EQ-5D) oder die Short-Form-6-Dimension (SF-6D), wobei die Messwerte der einzelnen Domänen, d. h. Schmerz, Mobilität, zu einem Gesamtindex (EQ-5D *utility*, SF-6D *utility*) zusammengefasst werden [20,21] oder c) sonstige quantitative Surveys und Fragebögen ohne Nutzwertbezug [22,23]. Darüber hinaus können qualitative Studien Evidenz zur relativen Bedeutung von Endpunkten liefern [24,25] (siehe Anhang 1 unter [https://www.jclinepi.com/article/S0895-4356\(17\)31036-3/fulltext](https://www.jclinepi.com/article/S0895-4356(17)31036-3/fulltext)).

Da die Gesundheit betreffende Entscheidungen sowohl durch gesundheitsbezogene Effekte der Interventionen als auch durch die relative Bedeutung interessierender Endpunkte beeinflusst werden, müssen angemessene Methoden eingesetzt werden, um die Vertrauenswürdigkeit der zugrundeliegenden Evidenz zu bewerten. Die GRADE-Arbeitsgruppe hat Methoden entwickelt, um die Vertrauenswürdigkeit der Evidenz zu Interventionseffekten [26,27], zur Testgenauigkeit [28], zur Ressourcennutzung [29,30], zu Prognose [31] und aus qualitativer Evidenz [32] zu bewerten. Ebenso hat die GRADE-Arbeitsgruppe die zunehmende Notwendigkeit erkannt, ein transparentes und strukturiertes Verfahren zur Bewertung der Vertrauenswürdigkeit der Evidenz zur relativen Bedeutung von Endpunkten zu entwickeln. Der GRADE-Ansatz zur Ableitung von Empfehlungen und Entscheidungen aus der Evidenz (*Evidence to Decision (EtD) Framework*) sowie deren Aufbereitung und Darstellung in EtD-Tabellen, erfordert eine Beurteilung des Vertrauens in die relative Bedeutung der Endpunkte [33–40]. Im Rahmen der letzten Iteration des EtD-Ansatzes wurde die Frage nach der relativen Bedeutung von Endpunkten folgendermaßen formuliert: „Besteht wesentliche Unsicherheit oder Variabilität dahingehend, welchen Wert Patient\*innen den gesundheitsbezogenen Endpunkten beimessen?“.

Nachdem wir die Gründe für diese Überlegungen dargelegt haben, werden wir zunächst die von uns verwendete Terminologie erörtern (siehe Box 2) [3,41,42]. Dabei stellen wir fest, dass die entsprechenden Begriffe in der wissenschaftlichen Gemeinschaft nicht einheitlich verwendet werden. Beispielsweise sind sich

nicht alle Wissenschaftler\*innen einig, dass eine visuelle Analogskala ein geeignetes Instrument zur Bestimmung des Nutzens ist, da bei dieser Methode eine Entscheidung unter Unsicherheit (choice under uncertainty) nicht verlangt wird. Trotz dieser Problematik verwenden wir die Terminologie „Bedeutung von Endpunkten“, da dieser Begriff über die strikte Definition des „Nutzens“ hinausgeht. Der Begriff „Bedeutung von Endpunkten“ bietet den Vorteil, dass er konzeptionell übereinstimmt mit dem Prozess des Abwägens zwischen gesundheitlichem Nutzen und Schaden. Außerdem fokussieren wir uns auf die „relative“ Bedeutung von Endpunkten, um deutlich zu machen, dass sich die Bedeutung auf einen Anker (z.B. „0“ für Tod und „1“ für optimale Gesundheit) oder andere Endpunkte bezieht, die durch eine Intervention kausal beeinflusst und gegeneinander abgewogen werden, um eine informierte Entscheidung zu treffen.

Ziel dieses und des nächsten Artikels zur relativen Bedeutung von Endpunkten ist es, GRADE-Empfehlungen für die Bewertung der Vertrauenswürdigkeit eines Evidenzkörpers zu geben, der sich mit der relativen Bedeutung von Endpunkten befasst. In diesem Artikel beschreiben wir sowohl die Definitionen und Methoden des Projektes als auch den GRADE-Ansatz bei der Bewertung des Risikos für Bias und der Indirektheit der Evidenz zur relativen Bedeutung von Endpunkten. Der zweite Artikel wird sich mit den Domänen Inkonsistenz, unzureichende Präzision und Publikationsbias sowie dem Heraufstufen der Vertrauenswürdigkeit der Evidenz beschäftigen. In diesem Kontext befasst sich der zweite Artikel auch mit der Bedeutung der Variabilität von Werten und Präferenzen bzw. der Variabilität der relativen Bedeutung von Endpunkten.

### Box 1. Ein hypothetisches Beispiel zur Bewertung der Bedeutung von Endpunkten

Die Evidenz aus dem Vergleich einer neuen Intervention gegenüber einer Standardbehandlung zeigt eine absolute Risikoreduktion von 10 pro 1000 in einem Schadensendpunkt „A“ und eine absolute Risikosteigerung von 10 pro 1000 in einem Schadensendpunkt „B“.

Wenn beide Endpunkte A und B als gleichbedeutend bewertet werden (z. B. Thrombose vs. Blutungen), dann führt die Nutzen-Schaden-Balance nicht zur Bevorzugung oder Ablehnung der neuen Intervention.

Wenn Endpunkt A als bedeutsamer eingeschätzt wird im Vergleich zu Endpunkt B (z. B. Mortalität vs. Blutungen), dann führt die Nutzen-Schaden-Balance zur Favorisierung der neuen Intervention.

## Methoden

Dieser Artikel stellt formale Empfehlungen der GRADE-Arbeitsgruppe zur Bewertung der relativen Bedeutung von Endpunkten dar. Diese Empfehlungen wurden unter Verwendung eines iterativen, mehrstufigen Verfahrens erarbeitet. Die Ergebnisse dieser Arbeit wurden auf Treffen der GRADE-Arbeitsgruppe vorgestellt und von Mitgliedern der GRADE-Arbeitsgruppe überprüft, bevor sie am 27. April 2017 bei einem Treffen der GRADE-Arbeitsgruppe in Rom durch Abstimmung beschlossen wurden. Anschließend wurden sie von der GRADE-Leitungsgruppe formal genehmigt.

## Box 2. Terminologie

Terminologie	Definition
Endpunkt	Der Begriff Endpunkt umschreibt den „Gesundheitszustand“ sowie Zustände ohne Bezug zur Gesundheit, welche im Hinblick auf die zur Diskussion stehenden Behandlungsalternativen relevant sind. Dazu zählt ein breites Spektrum an Endpunkten, die mit Gesundheit oder Krankheit, einer Intervention oder sonstigen Folgen ohne Gesundheitsbezug in direktem und indirektem Zusammenhang stehen. Endpunkte können einen mehr oder weniger starken Bezug zur Gesundheit haben. Beispielsweise werden die meisten Patient*innen ihre Ansichten im Hinblick auf die Bedeutung der folgenden Endpunkte haben (von größtenteils gesundheitsbezogen zu am wenigsten gesundheitsbezogen): Atemnot, Therapielast durch Warfarin oder Insulininjektionen, Einfachheit der Erreichbarkeit einer Klinik, um sich dort einem Bluttest oder anderen Tests zu unterziehen.
Relative Bedeutung von Endpunkten	Der Begriff <i>relative Bedeutung von Endpunkten</i> wird synonym verwendet mit den Begriffen <i>Werte und Präferenzen</i> , <i>Bedeutung von Endpunkten</i> oder <i>Wertung von Endpunkten</i> . Dennoch liegt der konzeptionelle Fokus auf Endpunkten, die sich auf die Folgen einer Intervention oder Entscheidung beziehen [3].
Instrument (zur Bestimmung der relativen Bedeutung von Endpunkten)	Dieser Begriff bezieht sich auf „Messinstrumente“, „Messverfahren“ oder „Messmethoden“ zur Erfassung der relativen Bedeutung von Endpunkten.
Aussagesicherheit der Evidenz	Die Begriffe „Aussagesicherheit der Evidenz“, „Qualität der Evidenz“, „Evidenzstärke“, „Vertrauen in den Effektschätzer“ werden synonym verwendet. Die <i>Aussagesicherheit der Evidenz</i> hat unterschiedliche Bedeutungen für systematische Reviews und Leitlinien. Die Definition für systematische Reviews lautet: das Ausmaß unseres Vertrauens, dass die relative Bedeutung der Endpunkte (und Variabilität) innerhalb eines bestimmten Bereichs liegen; die Definition für Leitlinien lautet: das Ausmaß unseres Vertrauens, dass die relative Bedeutung der Endpunkte (und Variabilität) ausreichen, um eine bestimmte Empfehlung zu stützen [41,42].

## Zusammenfassung von Domänen und Methoden zur Bewertung der Vertrauenswürdigkeit von Evidenz und Entwicklung des GRADE-Ansatzes

Basierend auf einem früheren systematischen Surveyprojekt [43] haben wir systematische Übersichtsarbeiten identifiziert, welche die relative Bedeutung von Endpunkten adressierten, und Methoden zur Beurteilung der Vertrauenswürdigkeit eines Evidenzkörpers sowie anderer potentieller Qualitätsindikatoren, d.h. aller Faktoren, die als Einflussfaktoren für die Vertrauenswürdigkeit angesehen werden, qualitativ zusammengefasst. Entsprechend haben wir eine Liste möglicher Faktoren erstellt und diese dann an die bestehenden GRADE-Domänen angepasst. Wir berücksichtigten die bestehenden GRADE-Domänen Risiko für Bias, Inkonsistenz, Indirektheit, unzureichende Präzision und Publikationsbias für das Herabstufen [41,42] bzw. große Effektstärke, Dosis-Wirkungs-Beziehung oder plausibles residuales Confounding für das Heraufstufen [44,45]. Darüber hinaus sollten auch weitere relevante Domänen erfasst werden.

### Beispielhafte Anwendung des GRADE-Ansatzes

Wir haben eine Stichprobe von 10 systematischen Übersichtsarbeiten [43] mit einer maximalen Varianzstichproben-Strategie (maximum variance sampling) ausgewählt, um sicherzustellen, dass wir durch diese Auswahl alle GRADE-Domänen veranschaulichen und eine Vielzahl von Gesundheitszuständen abbilden können. Zuerst haben wir geprüft, ob die bestehenden GRADE-Domänen alle Aspekte der Vertrauenswürdigkeit für die Bewertung der relativen Bedeutung der Endpunkte abdecken ohne dabei neue Domänen zu identifizieren. Erwägungen und Signalfragen für die Bewertung aus früheren GRADE-Handlungsempfehlungen, z.B. der Empfehlungen zu prognostischer Evidenz (e.g. guidance on prognostic evidence) [31], wurden angepasst. Für jedes bewertete Beispiel haben wir Entscheidungen zur Herabstufung erfasst und GRADE-Evidenzprofile entwickelt (siehe z.B. *Tabelle 1*) [4,46,47]. Sechs Untersucher\*innen (PA, HPH, IE, JJYN, YZ und YZ) bewerteten unter Verwendung der GRADE-Domänen paarweise unabhängig voneinander die Vertrauenswürdigkeit der Evidenz. Unstimmigkeiten wurden durch Diskussionen oder durch Feedback von erfahrenen GRADE-Methodikern (HJS, GG), die die Beispiele ebenso ausgewertet haben, gelöst. Die Ergebnisse wurden als cloudbasierte Dokumente gespeichert und für Kommentare und Feedbacks vorbereitet.

### Konsultation für Feedbacks

Um eine umfassende Perspektive zu gewährleisten, haben wir die Beispiele und Handlungsempfehlungen einer Gruppe von Personen aus Kanada, den USA und Europa zur Verfügung gestellt, darunter Leitlinienentwickler\*innen, Autoren\*innen von systematischen Übersichtsarbeiten oder HTA-Berichten, klinische Epidemiolog\*innen, Biostatistiker\*innen, Ärzt\*innen und Forscher\*innen mit Erfahrung in der Bewertung der relativen Bedeutung von Endpunkten. Die Gruppe erhielt Zugang zu den cloudbasierten Dokumenten und lieferte Feedback in sechs Online-Meetings Runden, ergänzt durch E-Mails, persönliche Treffen und Telefonate (siehe Anhang 3 unter [https://www.jclinepi.com/article/S0895-4356\(17\)31036-3/fulltext](https://www.jclinepi.com/article/S0895-4356(17)31036-3/fulltext) für Protokolle der Sitzungen). Wir haben die Handlungsempfehlungen überarbeitet, die vorgenommenen Anpassungen dokumentiert und die Aufzeichnungen im Rahmen einer GRADE-Projektgruppe zur Überprüfung und Kommentierung zirkulieren lassen. Nach jeder Feedbackrunde haben wir die vorläufigen GRADE-Empfehlungen zur Bewertung der Vertrauenswürdigkeit in Bezug auf die relative Bedeutung der Endpunkte iterativ verbessert und die Begründung anhand von Beispielen erläutert (siehe Anhang 4 unter [https://www.jclinepi.com/article/S0895-4356\(17\)31036-3/fulltext](https://www.jclinepi.com/article/S0895-4356(17)31036-3/fulltext) für die Empfehlungen in verschiedenen Phasen). Eine wesentliche Änderung, die wir vorgenommen haben, besteht beispielsweise darin, dass wir Überlegungen zur Glaubwürdigkeit von Subgruppen in der Inkonsistenz-Domäne hinzugefügt haben.

Als Teil des internen Review-Prozesses haben wir die Arbeit und Empfehlungen auf fünf der regulären Sitzungen der GRADE-Arbeitsgruppe präsentiert. Die Mitglieder hatten noch vor der formalen Genehmigung auf der GRADE-Arbeitsgruppensitzung am 27. April 2017 in Rom die Möglichkeit der Diskussion und des Feedbacks. Anschließend prüfte und genehmigte die GRADE-Leitungsgruppe das Dokument als offizielle Handlungsempfehlung, bevor es zur Begutachtung und Veröffentlichung eingereicht wurde.

### Empfehlungen für die GRADE-Domänen

Neben den von der GRADE-Arbeitsgruppe bereits vorgeschlagenen Domänen (Risiko für Bias, Inkonsistenz, Indirektheit, unzureichende Präzision, Publikationsbias und Domänen zum Heraufstufen der Vertrauenswürdigkeit der Evidenz [41,42]) konnten keine zusätzlichen Domänen von Relevanz zur Einschätzung der Vertrauenswürdigkeit eines Evidenzkörpers, der die relative Bedeutung der Endpunkte beschreibt, identifiziert werden. Nachfolgend wird auf die detaillierten Handlungsempfehlungen bezüglich der GRADE-Domänen Risiko für Bias und Indirektheit eingegangen.

### Risiko für Bias oder Limitationen beim Studiendesign oder der Studiendurchführung

Das Risiko für Bias kann in unterschiedlichen Phasen einer Studie zur relativen Bedeutung von Endpunkten ein Problem darstellen, inklusive dem Studiendesign, der Studiendurchführung, der Datenanalyse und dem Berichten von Studienergebnissen [48]. Die Einschätzung des Risikos für Bias für die relative Bedeutung von Endpunkten ist mit der Beurteilung des Risikos für Bias von Interventionseffekten vergleichbar, da zunächst eine Untersuchung des Risikos für Bias von Primärstudien, und anschließend des Risikos für Bias des gesamten Evidenzkörpers durchgeführt wird. Trotzdem bestehen in bestimmten Punkten Unterschiede: 1.) Anders als bei Studien zur Untersuchung eines Interventionseffekts existiert kein etabliertes oder häufig angewendetes Instrument zur Bestimmung des Risikos für Bias bezogen auf die relative Bedeutung von Endpunkten oder zur Untersuchung des Risikos für Bias bei unterschiedlichen Studiendesigns [48]. 2.) Die relative Bedeutung eines Endpunkts ist eine Schätzung, die nicht den Effekt repräsentiert, sondern sich konzeptionell näher an einer Schätzung der Testgenauigkeit oder eines Baseline-Risikos orientiert. Daher ist eine Randomisierung nicht erforderlich, um vor einem Bias durch Confounder zu schützen oder um bekannte und unbekannte prognostische Faktoren, die den Endpunkt beeinflussen können, auszugleichen. Entsprechend wird die Vertrauenswürdigkeit der Evidenz von nicht-randomisierten Studien zu Beginn als „hoch“ bewertet [28,31].

### Risiko für Bias von Subdomänen

Es wurden die folgenden Subdomänen, sowie für jede Subdomäne einzelne Signalfragen zur Untersuchung des Risikos für Bias entwickelt (siehe *Tabelle 2* und Anhang 5 unter [https://www.jclinepi.com/article/S0895-4356\(17\)31036-3/fulltext](https://www.jclinepi.com/article/S0895-4356(17)31036-3/fulltext) für eine detaillierte Anleitung):

1. Selektion von Studienteilnehmer\*innen: Inwiefern repräsentieren die in die Studie eingeschlossenen Teilnehmer\*innen die Zielpopulation? Eine nicht-adäquate Auswahl der Stichprobe führt zu einer Verzerrung in der Schätzung der relativen Bedeutung eines Endpunkts, falls die unterschiedlichen Charakteristika der Patient\*innen mit der relativen Bedeutung der Endpunkte bei den Teilnehmenden assoziiert sind.
2. Vollständigkeit der Daten: Inwiefern ähneln diejenigen Personen, die auf Fragen eingehen, denen, die nicht darauf eingehen? Ein hohes Ausscheiden während des Follow-up's oder niedrige Rücklaufquoten bei Querschnittsstudien können auf systematischen Unterschieden hinsichtlich der relativen Bedeutung von Endpunkten zwischen teilnehmenden und nicht-teilnehmenden Proband\*innen beruhen [49,50].
3. Messinstrumente: Inwiefern wurde ein validiertes Instrument gewählt, um die relative Bedeutung eines Endpunkts zu

**Tabelle 1**

Beispiel einer GRADE-Bewertung zur Aussagesicherheit. Evidenzprofil, **Autor\*innen:** Yuan Zhang, Pablo Alonso Coello, Holger Schünemann; **Datum:** 2017/05/01. **Frage:** Was sind die Ansichten über den relativen Wert/die Bedeutung von Endpunkten, die für die Entscheidungsfindung bei Patient\*innen mit antithrombotischer Behandlung von Interesse sind?. **Setting:** nicht spezifiziert; **Bibliographie:** MacLean S. Chest 2012; 141:e1S-e23S. [4] (siehe Anhang 2 unter [https://www.jclinepi.com/article/S0895-4356\(17\)31036-3/fulltext](https://www.jclinepi.com/article/S0895-4356(17)31036-3/fulltext) für die vollständige Zitierung der eingeschlossenen Studien dieses systematischen Reviews).

Endpunkte	Studiendesign/ Messinstrument	Bewertung der Qualität					Vertrauen in die Effektschätzer (95% KI oder andere Variabilitätsmaße)	Vertrauenswürdigkeit
		Risiko für Bias	Inkonsistenz	Indirektheit	unzureichende Präzision	andere		
<i>Schlaganfall</i>								
Nicht tödlicher schwerer Schlaganfall	7 Querschnittsstudien, 580 Teilnehmer*innen  VAS, SG, TTO	kein schwerwiegendes Risiko <sup>1,2,3,4</sup>	keine schwerwiegende Inkonsistenz	keine schwerwiegende Indirektheit	keine schwerwiegende unzureichende Präzision	keine	0,1–0,39 (Bereich der Punktschätzungen)  0,149, 95% KI: 0,135–0,163	⊕⊕⊕⊕ Hoch
Moderater Schlaganfall	5 Querschnittsstudien, 339 Teilnehmer*innen  TTO, SG	kein schwerwiegendes Risiko	schwerwiegende Inkonsistenz <sup>5,6</sup>	keine schwerwiegende Indirektheit	keine schwerwiegende unzureichende Präzision	keine	0,29–0,77 (Bereich der Punktschätzungen)  0,664, 95% KI: 0,643 – 0,684	⊕⊕⊕○ Moderat
<i>Blutungen</i>								
Schwere (nicht spezifizierte) GI-Blutungen	3 Querschnittsstudien, 153 Teilnehmer*innen  VAS, TTO, SG	kein schwerwiegendes Risiko <sup>1,3</sup>	keine schwerwiegende Inkonsistenz	keine schwerwiegende Indirektheit	keine schwerwiegende unzureichende Präzision	keine	0,65–0,84 (Bereich der Punktschätzungen)  0,789, 95% KI: 0,758 – 0,820	⊕⊕⊕⊕ Hoch
<i>PPS</i>								
Schwere PPS	2 Querschnittsstudie, 66 Teilnehmer*innen  SG	kein schwerwiegendes Risiko <sup>7</sup>	keine schwerwiegende Inkonsistenz	schwerwiegende Indirektheit <sup>8</sup>	schwerwiegende unzureichende Präzision <sup>9</sup>	keine	0,93 – 0,982 (Bereich der Punktschätzungen)  0,973, 95% KI: 0,964 – 0,982	⊕⊕○○ Niedrig
<i>TVT</i>								
TVT und VTE, und Blutungen	1 Querschnittsstudie <sup>10, 124</sup> Teilnehmer*innen  TTO	kein schwerwiegendes Risiko	keine schwerwiegende Inkonsistenz	keine schwerwiegende Indirektheit	keine schwerwiegende unzureichende Präzision	keine	Wenn in den nächsten 2 Jahren eine 3%ige Chance für ein schwerwiegendes Blutungsereignis und eine 2%ige Chance für eine rezidivierende Episode einer venösen Thromboembolie vorliegt, schwanken die Rezidivraten der TVT ohne Behandlung zwischen 5%, 10% bis 15%. Der Prozentsatz der Teilnehmer*innen, die sich für eine Unterbrechung der VKA-Behandlungen entschieden haben, beträgt 21%, 23% bzw. 8%.	⊕⊕⊕⊕ Hoch



Tabelle 1 (Continued)

Bewertung der Qualität							Vertrauen in die Effektschätzer (95% KI oder andere Variabilitätsmaße)	Vertrauenswürdigkeit
Endpunkte	Studiendesign/ Messinstrument	Risiko für Bias	Inkonsistenz	Indirektheit	unzureichende Präzision	andere		
Belastung der Behandlung								
Belastung der Behandlung: Warfarin	7 Querschnittsstudien, 466 Teilnehmer*innen	kein schwerwiegendes Risiko <sup>1,2,3,4</sup>	keine schwerwiegende Inkonsistenz	keine schwerwiegende Indirektheit	keine schwerwiegende unzureichende Präzision	keine	0,66-1 (Bereich der Schätzungen für eingeschlossene Studien)	⊕⊕⊕⊕ Hoch
	VAS, SG, TTO						0,938, 95% KI: 0,934-0,942	
Belastung der Behandlung: Antikoagulans/ Warfarin	1 qualitative Studie, 21 Teilnehmer*innen  Halbstrukturiertes Interview <sup>11</sup>	kein schwerwiegendes Risiko	keine schwerwiegende Inkonsistenz	keine schwerwiegende Indirektheit	schwerwiegende unzureichende Präzision <sup>12</sup>	keine	Der Großteil (spezifischer Prozentsatz nicht angegeben) der Teilnehmer*innen hatte keine Komplikationen aufgrund von Warfarin. Viele Teilnehmer*innen berichteten über geringfügige Unannehmlichkeiten, wie etwa die tägliche Einnahme von Tabletten, regelmäßige Blutuntersuchungen und Ernährungsumstellungen.	⊕⊕⊕○ Moderat
GRADE Evidenzgrade der Arbeitsgruppe								
Hohe Vertrauenswürdigkeit: Das Panel ist sehr sicher, dass der wahre Effekt nahe bei dem Effektschätzer liegt.								
Moderate Vertrauenswürdigkeit: Das Panel hat mäßig viel Vertrauen in den Effektschätzer: Der wahre Effekt ist wahrscheinlich nahe bei dem Effektschätzer, aber es besteht die Möglichkeit, dass er grundlegend verschieden ist.								
Niedrige Vertrauenswürdigkeit: Das Vertrauen des Panels in den Effektschätzer ist begrenzt: Der wahre Effekt kann durchaus grundlegend verschieden vom Effektschätzer sein.								
Sehr niedrige Vertrauenswürdigkeit: Das Panel hat nur sehr wenig Vertrauen in den Effektschätzer: Der wahre Effekt ist wahrscheinlich grundlegend verschieden vom Effektschätzer.								

KI: Konfidenzintervall; GI-Blutung: Gastrointestinale Blutung; GRADE: Grading of Recommendations Assessment, Development, and Evaluation; PPS: postphlebitisches Syndrom; SG: standard gamble (Standardlotterie); TTO: Time Trade Off; VTE: tiefe Venenthrombose; VAS: Visuelle Analogskale; VKA: Vitamin K-Antagonisten; VTE: venöse Thromboembolie.

<sup>1</sup> Die Repräsentativität der Studien wurde durch einen geringen Rücklauf beeinflusst. Dies betraf jedoch nur einen kleinen Teil der eingeschlossenen Studienpopulation.

<sup>2</sup> In Protheroe 2000 antworteten 97 von 260 eingeschlossenen Patient\*innen.

<sup>3</sup> In Thomson 2000 beendeten 57 von 180 eingeschlossenen Patient\*innen das Interview.

<sup>4</sup> 17,4% der Studienteilnehmer\*innen in Gage 1995 haben die Time-Trade-Off Technik nicht verstanden.

<sup>5</sup> Große Abweichung der Punktschätzer.

<sup>6</sup> Die eingeschlossene Studienpopulation bestand aus Patient\*innen mit Vorhofflimmern (Gage 1996), 30 Freiwilligen aus der Gemeinde (Lenert 1997), drei verschiedenen Patient\*innengruppen in Locadia 2004 (Patient\*innen mit einer ersten oder zweiten Episode einer venösen Thromboembolie und begonnener oraler Antikoagulation; Patient\*innen, die während der Gabe von Antikoagulantien ein großes Blutungsereignis erlebt haben und Patient\*innen mit postthrombotischem Syndrom), Patient\*innen mit tiefer Venenthrombose und ohne tiefe Venenthrombose (O'Mera 1994) sowie Überlebende eines ischämischen Schlaganfalls und altersgerechten Kontrollpersonen (Slot 2009).

<sup>7</sup> Eine Studie (Lenert 1997) wurde mit einem hohen Risiko für Bias bewertet. Diese Studie hatte jedoch ähnliche Schätzer wie die andere eingeschlossene Studie mit geringem Verzerrungspotential.

<sup>8</sup> Die Vertrauenswürdigkeit der Evidenz wurde hinsichtlich der Indirektheit herabgestuft. Die eingeschlossenen Studien haben eine andere Population als die Patient\*innen, denen die Wahl gestellt wurde: 30 Freiwillige aus der Gemeinschaft (Lenert 1997), Patient\*innen mit tiefer Venenthrombose und ohne tiefe Venenthrombose (O'Mera 1994).

<sup>9</sup> Kleine Stichprobengröße: 66 Teilnehmer\*innen aus 2 Studien.

<sup>10</sup> Locadia 2004 ist eine Querschnittsstudie, in der die Teilnehmer\*innen mit einer Entscheidungsanalyse interviewt werden.

<sup>11</sup> Dantas 2004 ist eine qualitative Studie zur Belastung der Behandlung mit Antikoagulantien / Warfarin.

<sup>12</sup> Es wurde nur eine qualitative Studie (Dantas 2004) identifiziert, die dieses Phänomen adressiert.

**Tabelle 2**  
Risiko für Bias: Subdomänen und Signalfragen.

Subdomäne	Signalfragen
Auswahl der Teilnehmer*innen in die Studie	Wurde eine angemessene Fallzahl ausgewählt?
Vollständigkeit der Daten	Gab es einen geringen Verlust (dropout) an Teilnehmer*innen?
Messinstrumente	Wurde ein reliables und valides Instrument zur Messung der relativen Bedeutung des Endpunkts verwendet? Wurde das Instrument bestimmungsgemäß eingesetzt? Gab es eine valide Darstellung des Endpunkts (Gesundheitsstatus)? Haben die Forscher*innen untersucht, inwieweit das Instrument von den untersuchten Personen verstanden wurde?
Datenanalyse	Wurden die Ergebnisse angemessen analysiert, um einen Einfluss durch Störgrößen und eine Verzerrung zu vermeiden?

untersuchen und wie akkurat wurde das Instrument eingesetzt? Diese Subdomäne enthält 4 Items: Auswahl des Instruments, Einsatz des Instruments, Ergebnisdarstellung und Verständlichkeit des Instruments für die Studienteilnehmenden. Eine geringe Reliabilität oder Validität der Messung kann durch eine intrinsische Limitation des Messinstruments oder durch administrative Fehler zustande kommen. Zusätzlich kann die Untersuchung der Bedeutung von Endpunkten abhängig von den Eigenschaften des Endpunkts in zwei Kategorien eingeteilt werden: Untersuchung eines Ereignisses bei der betroffenen Person, das entweder bereits eingetreten ist oder gerade eintritt, oder die Untersuchung eines dargestellten (häufig hypothetischen) Ereignisses, das vielleicht oder vielleicht auch nicht in der Zukunft eintreten könnte. Dies beinhaltet die Beurteilung der Indirektheit, die später beschrieben wird. Für letzteres ist unabhängig von der Einschätzung der Indirektheit zusätzlich die Art der Ergebnispräsentation wichtig.

4. Datenanalyse: In welchem Ausmaß wird der Schätzer durch eine nicht geeignete Datenanalyse beeinflusst? War die Adjustierung oder Stratifikation der Analyse sowie die Auswahl des Modells angemessen, um eine Verzerrung der Ergebnisse durch Confounder zu vermeiden?

Auf Grundlage der Antworten auf die oben genannten Signalfragen wird für jede Studie in Abhängigkeit der Wahrscheinlichkeit einer Verzerrung sowie des Einflusses auf die Schätzer das Risiko für Bias jeder Subdomäne als niedrig, moderat, schwerwiegend und kritisch eingestuft (siehe Box 3). Die Einstufung des Risikos für Bias von individuellen Studien (siehe Box 4 für alle Subdomänen innerhalb einer Studie) ist hilfreich, um eine Studie zu beschreiben und darüber hinaus notwendig für die Bewertung des Risikos für Bias des Evidenzkörpers.

#### Zusammenfassung des Risikos für Bias für einen Evidenzkörper

Die Einschätzung des Risikos für Bias für einen Evidenzkörper erfordert die Betrachtung des Gesamtbildes der Ergebnisse über alle Subdomänen und Studien hinweg. Darüber hinaus ist es notwendig, die relative Gewichtung bzw. den Einschluss von Studien mit einem möglichen Risiko für Bias zu prüfen und einzuschätzen, inwiefern dieses Risiko für Bias das Gesamtergebnis beeinflussen kann. Die Einschätzung des Evidenzkörpers hinsichtlich des Risikos für Bias wird entweder als nicht schwerwiegend, schwerwiegend oder sehr schwerwiegend bewertet.

#### Box 3. Beurteilung des Risikos für Bias von Subdomänen

Antwortmöglichkeit	Interpretation
niedriges Risiko für Bias	Die Schätzung der relativen Bedeutung der Endpunkte ist höchstwahrscheinlich nicht verzerrt in Bezug auf diese Subdomäne.
moderates Risiko für Bias	Die Schätzung der relativen Bedeutung der Endpunkte ist wahrscheinlich verzerrt in Bezug auf diese Subdomäne, jedoch ist der Einfluss der Verzerrung begrenzt.
schwerwiegendes Risiko für Bias	Die Schätzung der relativen Bedeutung der Endpunkte ist höchstwahrscheinlich verzerrt in Bezug auf diese Subdomäne und der Einfluss der Verzerrung ist erheblich.
kritisches Risiko für Bias	Die Schätzung der relativen Bedeutung der Endpunkte ist mit Sicherheit verzerrt in Bezug auf diese Subdomäne und der Schätzer ist nicht vertrauenswürdig.

#### Box 4. Gesamtrisiko für Bias

Antwortmöglichkeit	Interpretation
niedriges Risiko für Bias	Die Studie hat ein niedriges Risiko für Bias über alle Subdomänen hinweg.
moderates Risiko für Bias	Die Studie hat ein niedriges bis moderates Risiko für Bias über alle Subdomänen hinweg.
schwerwiegendes Risiko für Bias	Die Studie hat ein schwerwiegendes Risiko für Bias in mindestens eine Subdomäne, das aber nicht als kritisch über alle Subdomänen hinweg eingeordnet wird.
kritisches Risiko für Bias	Die Studie hat ein kritisches Risiko für Bias in mindestens einer Subdomäne.

Wir ermutigen Bewerter\*innen, eine Beurteilung auf Grundlage der verfügbaren Informationen (entweder aus dem Studienbericht oder anhand zusätzlicher von den Autor\*innen zur Verfügung gestellter Informationen) sowie auf der Basis von Schlussfolgerungen über Aspekte, die zwar nicht angegeben wurden, aber sehr wahrscheinlich sind, vorzunehmen.

In Übereinstimmung mit dem GRADE-Ansatz für andere Evidenztypen wird das Risiko für Bias für jeden Endpunkt ermittelt [51,52]. Falls überwiegend Informationen aus Studien mit niedrigem Risiko für Bias in allen Subdomänen zur Verfügung stehen, sollte auch die Gesamtbeurteilung zum Risiko für Bias „niedriges Risiko für Bias“ lauten und die Vertrauenswürdigkeit durch GRADE-Anwender\*innen nicht heruntergestuft werden. Sobald allerdings der Einschluss von Studien mit einem Risiko für Bias dazu führt, dass Bedenken hinsichtlich der Vertrauenswürdigkeit des Evidenzkörpers entstehen, wird diese um eine oder mehrere Stufen herabgestuft [53,54]. Die Beurteilung des Risikos für Bias für jede Subdomäne findet auf einem Kontinuum statt, und Begutachter\*innen sollten dies bei ihrer zusammenfassenden Beurteilung mitberücksichtigen. Falls notwendig, können Anwender\*innen Sensitivitätsanalysen durchführen, um herauszufinden, ob das Risiko für Bias einzelner Studien das Gesamtergebnis über alle Subdomänen und Studien hinweg beeinflusst.

Beispiel: Eine systematische Übersichtsarbeit fasste den Nutzen zusammen, den Patient\*innen mit schwerem, nicht-tödlichem

Schlaganfall ihrem Gesundheitszustand beimaßen [4]. Zwei von sieben eingeschlossenen Studien berichteten von einer niedrigen Rücklaufquote, und 17% der Patient\*innen einer dritten Studie gaben Verständnisschwierigkeiten beim Einsatz des Instruments an. Die Daten dieser drei Studien machten 35% der insgesamt ausgewerteten Daten aus. Allerdings gab es keine Hinweise zu anderen Subdomänen für Risiko für Bias und die Ergebnisse der Studien ähnelten denen mit niedrigem Risiko für Bias. Daher erfolgte keine Herabstufung hinsichtlich eines möglichen Risikos für Bias [4]. In einer anderen systematischen Übersichtsarbeit zur Erfassung von Patient\*innenpräferenzen zu behandlungsrelevanten Endpunkten bei Diabetes-Typ 2 zeigten nur 6 von 61 eingeschlossenen Studien eine ausreichende Vergleichbarkeit zwischen Befragten und Nicht-Befragten [55]. Daher wurde die Vertrauenswürdigkeit der Evidenz wegen eines Risikos für Bias hinsichtlich der Auswahl der Patient\*innen in die Studien herabgestuft.

## Indirektheit

In Bezug auf die relative Bedeutung von Endpunkten kann auch Indirektheit ein Grund für die Herabstufung der Vertrauenswürdigkeit sein [56,57]. Die Einschätzung der Indirektheit im Hinblick auf die relative Bedeutung von Endpunkten weist spezifische Merkmale auf. Zum einen konzentrieren sich Studien für gewöhnlich auf Endpunkte, statt direkte Vergleiche zwischen den Interventionen vorzunehmen. Zum anderen sind Surrogat-Endpunkte und Endpunkte, die nicht patient\*innenrelevant sind, eine Ursache der Indirektheit bezogen auf Behandlungsfragen. Dies trifft möglicherweise in dieser Form nicht auf Evidenz zur relativen Bedeutung von Endpunkten zu. Aber auch in diesen Studien können Endpunkte, wenn sie nicht interessierende Endpunkte repräsentieren, indirekt sein. Wenn wir an der Bedeutung eines Surrogat-Endpunktes aus Sicht der Patient\*innen interessiert sind, sollte die Vertrauenswürdigkeit also nicht herabgestuft werden, nur weil es sich um einen Surrogat-Endpunkt handelt. Darüber hinaus gibt es keinen indirekten Vergleich in der Evidenz zur relativen Bedeutung von Endpunkten. Schlussendlich können auch Methoden, die zur Ermittlung der relativen Bedeutung von Endpunkten verwendet werden, eine Ursache von Indirektheit sein. Nachfolgend werden die Gründe und Beispiele für diese Überlegungen dargestellt. Sie werden in zwei Kategorien eingeteilt: Indirektheit aufgrund der Population, Intervention, des Vergleichs und der Endpunkte (PICO) sowie Indirektheit aufgrund von methodologischen Aspekten (siehe [Tabelle 3](#)).

### PICO-Elemente

Für eine systematische Übersichtsarbeit, die die relative Bedeutung von Endpunkten adressiert, kann die Forschungsfrage wie folgt definiert werden: „Was ist die relative Bedeutung, die Patient\*innen Endpunkten zusprechen, wenn sie eine auf diese Endpunkte bezogene Entscheidung treffen?“. Auch hierfür werden klar definierte PICO-Elemente benötigt. PICO-Elemente können eine Ursache der Indirektheit sein, wenn der Evidenzkörper nicht die Elemente widerspiegelt, die von Interesse sind (siehe Anhang 6 unter [https://www.jclinepi.com/article/S0895-4356\(17\)31036-3/fulltext](https://www.jclinepi.com/article/S0895-4356(17)31036-3/fulltext)).

Wenn die Endpunkte, die in den verfügbaren Studien adressiert werden, nicht die interessierenden Endpunkte repräsentieren, dann ist die Vertrauenswürdigkeit der Evidenz notwendigerweise niedriger. Ob die Prüf- und Vergleichsinterventionen die Ursache der Indirektheit sind, ist davon abhängig, inwieweit Unterschiede in dem betrachteten Endpunkt auf Unterschiede zwischen den Interventionen zurückgeführt werden können. Interventionen können sich hinsichtlich vieler Aspekte unterscheiden – chirurgi-

**Tabelle 3**  
Signalfragen zur Indirektheit.

Ursachen von Indirektheit	Signalfragen
Indirektheit aufgrund der PICO-Elemente	<p>War die in der Studie untersuchte Population ausreichend vergleichbar in Bezug zur interessierenden Population?</p> <p>Waren die in der Studie untersuchten Endpunkte ausreichend vergleichbar in Bezug zu den interessierenden Endpunkten?</p> <p>Waren die in der Studie untersuchten Kontrollinterventionen ausreichend vergleichbar in Bezug zu den interessierenden alternativen Vergleichsmöglichkeiten?</p> <p>Haben die Teilnehmer*innen Fragen beantwortet, die sich direkt auf die relative Bedeutung von Endpunkten beziehen?</p> <ul style="list-style-type: none"> <li>• Wurden direkte statt indirekter Methoden zur Erfassung des Nutzens von Endpunkten verwendet?</li> <li>• Wurde der Nutzen mittels eines Instruments geschätzt, dessen direktes Ziel die Erfassung von Nutzen ist, oder basierte die Schätzung auf Instrumenten, die nicht explizit für diesen Zweck entwickelt wurden?</li> </ul>
Indirektheit aufgrund von methodologischen Aspekten	

sche Fähigkeiten oder Ansätze, Medikamentendosierungen, -dauer oder -verabreichungswege –, aber Bedenken sind nur angebracht, wenn die Unterschiede in den Interventionen wahrscheinlich auch zu Unterschieden in den Endpunkten führen.

Wir schließen Interventionen und Kontrollinterventionen, das „I“ und das „C“ von PICO, aus folgenden Gründen mit in die Bewertung der Indirektheit ein (und ebenso in die Domäne zur Inkonsistenz). Erstens können die Unterschiede in den zu vergleichenden Interventionen auf potenzielle Unterschiede in der Art der Endpunkte hindeuten. Zweitens deuten empirische Belege darauf hin, dass die Befragten denselben Endpunkt unterschiedlich beurteilen, wenn dieser als Folge unterschiedlicher Interventionen beschrieben wird [58].

Beispiel der Indirektheit auf der Basis der PICO-Elemente: Eine systematische Übersichtsarbeit fasste die relative Bedeutung zusammen, die Patient\*innen mit gutartiger Prostatahyperplasie ihrem Gesundheitszustand beimaßen: die Einschätzung der Symptomverbesserung, die Verringerung der Prostatagröße, die Risiken der akuten Harnretention und der Operation [59]. Die Autor\*innen schlussfolgerten, dass Männer länger auf eine Symptomverbesserung warten würden wenn sie im Gegenzug eine verringerte Prostatagröße erreichten (13 Monate), als sie es für eine absolute Risikoreduktion um 1% hinsichtlich der akuten Harnretention (2 Monate) oder Operation (8 Monate) täten. Diese Bewertung basierte allerdings auf lediglich einer Studie, die in das systematische Review eingeschlossen wurde. In dieser Studie wurden 208 Männer aus der Allgemeinbevölkerung im Alter von mindestens 40 Jahren eingeschlossen. Wir betrachten die optimale Studienpopulation in diesem Fall als eine alternde männliche Population mit einem Risiko für eine gutartige Prostatahyperplasie. Da die in die Studie eingeschlossenen Männer ( $\geq 40$  Jahre) im Allgemeinen jünger waren als Personen, die typischerweise mit dieser Entscheidung konfrontiert wären, haben die Autor\*innen dieser Studie keine ausreichend vergleichbare Population gewählt. Wir haben die Vertrauenswürdigkeit der Evidenz für die Indirektheit bezüglich der Population herabgestuft, da die Bewertung der Endpunkte die akute Harnretention und Operation mit eingeschlossen hat, obwohl dies für gewöhnlich keine Entscheidungen sind, die die meisten Männer der Allgemeinbevölkerung im Alter von 40 Jahren und älter treffen müssen [59]. Auf der anderen Seite sind alternde Männer der Allgemeinbevölkerung von einer Prostatahyperplasie bedroht, sodass die dargestellten Überlegungen für sie nicht völlig irrele-



vant sind. Bezogen auf die PICO-Elemente konnten keine weiteren Bedenken zur Indirektheit identifiziert werden. Wie dieses Beispiel zeigt, besteht der Vorteil des GRADE-Ansatzes nicht darin, Unstimmigkeiten zu beseitigen, sondern einen transparenten und expliziten Bewertungsprozess zu ermöglichen.

#### Methodologische Aspekte

Die Methoden, die für die Ermittlung der relativen Bedeutung von Endpunkten verwendet werden, können ebenso eine Ursache für Indirektheit darstellen (siehe Anhang 7 unter [https://www.jclinepi.com/article/S0895-4356\(17\)31036-3/fulltext](https://www.jclinepi.com/article/S0895-4356(17)31036-3/fulltext)). Diese Überlegung ist immer dann relevant, wenn Forscher\*innen ein indirektes Messverfahren verwendet haben (z.B. multifaktorieller Nutzenindex), um den Nutzen von Endpunkten darzustellen (aus dem EQ-5D, SF-6D, Qualität des Wohlbefindens (quality of wellbeing), oder Gesundheits-Nutzen-Index) oder wenn ein zusammenführender Algorithmus zur Schätzung eines generischen Nutzens angewendet wurde, der auf den Schätzungen anderer Messinstrumente (z.B. die Schätzung des Nutzens nach dem EQ-5D aus dem St. George Respiratory Questionnaire) beruht. Dies kann auf einer Verlinkung oder mathematischen Transformationsfunktionen basieren, die für die Berechnung der relativen Bedeutung von Endpunkten auf der Grundlage von Instrumenten (z.B. Instrumenten zur Messung der Lebensqualität) verwendet werden [60].

Wenn man Patient\*innen bittet, den Wert anzugeben, den sie dem Gesundheitszustand oder einem klinischen Szenario beimessen, dann kann man sie dies auch direkt fragen (unter Verwendung der SG, TTO und VAS). Multifaktorielle Nutzen-Messinstrumente (z.B. EQ-5D Nutzenscore, SF-6D Nutzenscore) verwenden ein solches direktes Messinstrument zusammen mit der Messung verschiedener Gesundheitsdomänen (z.B. Schmerz, Mobilität usw.), um Bewertungssysteme für die Einschätzung des Gesundheitszustandes zu entwickeln. Dies ist der Algorithmus zur Transformation der Messungen von Gesundheitsdomänen in den Nutzen. Anwender\*innen multifaktorieller Nutzen-Messinstrumente bitten die Befragten lediglich, ihren eigenen Gesundheitszustand anhand der Gesundheitsdomänen zu beschreiben. Auf diesem Weg geben die Befragten nicht ihre eigene Einschätzung der Bedeutung an, sondern berichten vielmehr über ihre Erfahrungen. Um daraus den Nutzen des individuellen Gesundheitszustandes zu erhalten, benötigen Forscher\*innen einen Algorithmus, der auf der Grundlage einer anderen Population geschätzt wird. Diese Werte kommen dann von jemand anderem, und je nach Art der Population kann die Bewertung des Nutzens indirekt sein.

Im Wesentlichen besteht die gleiche Situation, wenn Forscher\*innen Werte zur krankheitsbezogenen Lebensqualität (z.B. St. George Respiratory Questionnaire) in generische Nutzenwerte umwandeln. In diesem Fall werden indirekte Nutzenwerte nicht geschätzt sondern auf der Basis von Forschungsdaten vorhergesagt, die mit einem Instrument erhoben wurden, dessen primäre Zweckbestimmung es war, das Ausmaß der Behinderung zu beurteilen und nicht die Nutzenwerte zu schätzen. Auch hier wurden die Werte von einer anderen Person erhoben und sind damit indirekt.

In Abhängigkeit der Perspektive, die im Entscheidungsprozess im Gesundheitswesen eingenommen wird, entweder in einem gesundheitspolitischen Entscheidungsszenario, einem klinischen Leitlinienentwicklungsprojekt oder einer Entscheidung für einzelne Patient\*innen, können Anwender\*innen eine Herabstufung der Indirektheit bezogen auf die Vertrauenswürdigkeit der Evidenz jedoch kaum einschätzen. Wenn man annimmt, dass die Population, die mittels multifaktorieller Nutzen-Messinstrumente befragt wurde, Endpunkten dieselbe relative Bedeutung beimisst wie die Individuen, die an der Szenarienbewertung teilgenommen haben, die in erster Linie zu dem gewichteten Algorithmus geführt hat,

kann man daraus schließen, dass die Bewertungen identisch sind mit einer direkten Einschätzung der relativen Bedeutung der Endpunkte. Unter dieser Annahme würde man keine Herabstufung aufgrund von Indirektheit vornehmen.

#### Verschiedene Strategien für Autor\*innen systematischer Übersichtsarbeiten sowie Leitliniengruppen

In den meisten Fällen würden Autor\*innen systematischer Übersichtsarbeiten ausschließlich Studien einschließen, die die Einschlusskriterien für die Population, Intervention, Vergleichsgruppe und Endpunkte erfüllen und damit Direktheit gewährleisten [56,57]. In einigen Situationen kann es allerdings sein, dass Autor\*innen systematischer Übersichtsarbeiten indirekte Evidenz einschließen und aufgrund der Population und des primären Endpunkts die Indirektheit herabstufen. Im Gegensatz zu systematischen Übersichtsarbeiten ist die Verwendung indirekter Evidenz bei Leitlinien für die klinische Praxis weit verbreitet.

Diese unterschiedlichen Verwendungszwecke und Betrachtungsweisen von Evidenz könnten bei gleichem Evidenzkörper zu unterschiedlichen Einschätzungen der Indirektheit führen. Wie vorab diskutiert, ist in systematischen Übersichtsarbeiten die Bedeutsamkeit eines schweren Blutungsereignisses, das nach der Einnahme von Aspirin auftritt, genauso indirekt wie ein Blutungsereignis in Folge einer Warfarin-Einnahme. Im Gegensatz dazu kann es bei der Leitlinienentwicklung von Bedeutung sein, ob die Patient\*innen die Bedeutung des Blutungsereignisses nach der Einnahme von Warfarin oder der Einnahme von Aspirin bewerteten, insbesondere wenn sich die Schwere oder die Art der Blutung unterscheidet.

#### Zusammenfassung

Dieser Artikel beschreibt die Verwendung von GRADE für die Einschätzung der Vertrauenswürdigkeit von Evidenz zur relativen Bedeutung von Endpunkten unter Berücksichtigung des Risikos für Bias und der Indirektheit. Bei der Beurteilung der Vertrauenswürdigkeit für die relative Bedeutung von Endpunkten beginnt die Bewertung für alle Studiendesigns bei „hoch“. Bei schwerwiegenden Einschränkungen hinsichtlich des Risikos für Bias oder der Indirektheit wird die Vertrauenswürdigkeit herabgestuft. Abhängig von den spezifischen Einschränkungen stufen die Bewerter\*innen die Vertrauenswürdigkeit in den zwei Domänen um eine oder mehrere Stufen herab. Die Bewertung des Risikos für Bias stellt in diesem Zusammenhang eine Herausforderung dar. Wir haben hierfür eine Reihe von Signalfragen bereitgestellt, die die entsprechenden Probleme beim Risiko für Bias berücksichtigen. Die Reliabilität oder Validität der von uns vorgeschlagenen Ansätze wurde noch nicht untersucht, jedoch existiert bis jetzt kein anderes gut validiertes Instrument. Anhand der von uns zur Verfügung gestellten Signalfragen und Beispiele können Beurteilungen des Risikos für Bias transparent gemacht werden.

Im nächsten Artikel wird die Anwendung anderer GRADE-Domänen (unzureichende Präzision, Inkonsistenz, Publikationsbias und Domänen für das Heraufstufen) in der Bewertung der Aussagesicherheit der relativen Bedeutung von Endpunkten diskutiert.

#### Danksagung

Die Autorengruppe des Originalartikels ist Dr. Amiram Gafni von der McMaster University für Kommentare zu den Manuskripten dankbar.

Beiträge der Autoren des Originalartikels: Y.Z., P.A., G.G. und H.J.S. entwickelten die Methodik für dieses Projekt. H.J.S. hat das

Projekt und den Ansatz konzipiert. Y.Z., J.J.Y.N. und Y.C. haben die Bewertung zur Aussagesicherheit relevanter Items in systematischen Übersichtsarbeiten zusammengefasst; Y.Z., P.A., G.G. und H.J.S. haben einen Vorschlag zu den Subdomänen für die Einschätzung der Aussagesicherheit von Evidenz erarbeitet; alle Autoren haben an den methodischen Diskussionen teilgenommen. Alle Autoren haben das finale Manuskript gelesen und genehmigt. Die Autoren sind Mitglieder der GRADE-Arbeitsgruppe.

## Interessenkonflikt

Die Autor\*innen erklären, dass kein Interessenkonflikt besteht.

## Literatur

- [1] Schünemann HJ, Fretheim A, Oxman AD. Improving the use of research evidence in guideline development: 10 Integrating values and consumer involvement. *Health Res Policy Syst* 2006;4:22, doi: 10.1186/1478-4505-4-22.
- [2] Murad MH, Montori VM, Guyatt GH. Incorporating patient preferences in evidence-based medicine. *Jama* 2008;300(21):2483, <http://dx.doi.org/10.1001/jama.2008.730>, author reply -4.
- [3] Schünemann HJ, Wiercioch W, Etseandía I, Falavigna M, Santesso N, Mustafa R, et al. Guidelines 2.0: systematic development of a comprehensive checklist for a successful guideline enterprise. *CMAJ* 2014;186(3):E123–42, doi: 10.1503/cmaj.131237.
- [4] MacLean S, Mulla S, Akl EA, Jankowski M, Vandvik PO, Ebrahim S, et al. Patient values and preferences in decision making for antithrombotic therapy: A systematic review; Antithrombotic Therapy and Prevention of Thrombosis, 9th ed: American College of Chest Physicians Evidence-Based Clinical Practice Guidelines. *Chest* 2012;141(2 Suppl):e15–235, doi: 10.1378/chest.11-2290.
- [5] Andrews JC, Schünemann HJ, Oxman AD, Pottie K, Meerpohl JJ, Coello PA, et al. GRADE guidelines: 15 Going from evidence to recommendation—determinants of a recommendation's direction and strength. *J Clin Epidemiol* 2013;66(7):726–35, <http://dx.doi.org/10.1016/j.jclinepi.2013.02.003>.
- [6] Nussbaumer B, Gartlehner G, Kien C, Kaminski-Hartenthaler A, Langer G, Meerpohl JJ, et al. Grade Leitlinien: 15 Von der Evidenz zur Empfehlung – Determinanten, die Richtung und Stärke einer Empfehlung bestimmen. *Z Evid Fortbild Qual Gesundhwes* 2014;108(7):421–31, <http://dx.doi.org/10.1016/j.zefq.2014.08.004>.
- [7] Krahn M, Naglie G. The next step in guideline development: incorporating patient preferences. *Jama* 2008;300(4):436–8, <http://dx.doi.org/10.1001/jama.300.4.436>.
- [8] van der Weijden T, Legare F, Boivin A, Burgers JS, van Veenendaal H, Stiggelbout AM, et al. How to integrate individual patient values and preferences in clinical practice guidelines? A research protocol. *Implement Sci* 2010;5:10, doi: 10.1186/1748-5908-5-10.
- [9] Gafni A. The standard gamble method: what is being measured and how it is interpreted. *Health services research* 1994;29(2):207–24.
- [10] Torrance GW. Measurement of health state utilities for economic appraisal. *Journal of health economics* 1986;5(1):1–30.
- [11] Torrance GW. Utility measurement in healthcare: the things I never got to. *Pharmacoeconomics* 2006;24(11):1069–78, doi: 10.2165/00019053-200624110-00004.
- [12] Churchill DN, Torrance GW, Taylor DW, Barnes CC, Ludwin D, Shimizu A, et al. Measurement of quality of life in end-stage renal disease: the time trade-off approach. *Clinical and investigative medicine Medecine clinique et experimentale* 1987;10(1):14–20.
- [13] Torrance GW, Feeny D, Furlong W. Visual analog scales: do they have a role in the measurement of preferences for health states? *Medical decision making: an international journal of the Society for Medical Decision Making* 2001;21(4):329–34, doi: 10.1177/0272989x0102100408.
- [14] Morimoto T, Fukui T. Utilities measured by rating scale, time trade-off, and standard gamble: review and reference for health care professionals. *Journal of epidemiology* 2002;12(2):160–78, doi: 10.2188/jea.12.160.
- [15] Ryan M, Gerard K. Using discrete choice experiments to value health care programmes: current practice and future research reflections. *Applied health economics and health policy* 2003;2(1):55–64.
- [16] Ryan M. Discrete choice experiments in health care. *BMJ* 2004;328(7436):360–1, doi: 10.1136/bmj.328.7436.360.
- [17] Stevens TH, Belkner R, Dennis D, Kittredge D, Willis C. Comparison of contingent valuation and conjoint analysis in ecosystem management. *Ecological Economics* 2000;32(1):63–74, [http://dx.doi.org/10.1016/S0921-8009\(99\)00071-3](http://dx.doi.org/10.1016/S0921-8009(99)00071-3).
- [18] Alonso-Coello P, Montori VM, Diaz MG, Devereaux PJ, Mas G, Diez AJ, et al. Values and preferences for oral antithrombotic therapy in patients with atrial fibrillation: physician and patient perspectives. *Health expectations: an international journal of public participation in health care and health policy* 2015;18(6):2318–27, doi: 10.1111/hex.12201.
- [19] Devereaux PJ, Anderson DR, Gardner MJ, Putnam W, Flowerdew GJ, Brownell BF, et al. Differences between perspectives of physicians and patients on anticoagulation in patients with atrial fibrillation: observational study. *Bmj* 2001;323(7323):1218–22, doi: 10.1136/bmj.323.7323.1218.
- [20] Craig BM, Busschbach JJ, Salomon JA. Modeling ranking, time trade-off, and visual analog scale values for EQ-5D health states: a review and comparison of methods. *Medical care* 2009;47(6):634–41.
- [21] Rabin R, de Charro F. EQ-5D: a measure of health status from the EuroQol Group. *Annals of medicine* 2001;33(5):337–43, doi: 10.3109/07853890109002087.
- [22] Sepucha K, Ozanne EM. How to define and measure concordance between patients' preferences and medical treatments: A systematic review of approaches and recommendations for standardization. *Patient Educ Couns* 2010;78(1):12–23, <http://dx.doi.org/10.1016/j.pec.2009.05.011>.
- [23] King M, Nazareth I, Lampe F, Bower P, Chandler M, Morou M, et al. Conceptual framework and systematic review of the effects of participants' and professionals' preferences in randomised controlled trials. *Health technology assessment (Winchester, England)* 2005;9(35):1–186, iii-iv.
- [24] Cronin M, Meaney S, Jepson NJ, Allen PF. A qualitative study of trends in patient preferences for the management of the partially dentate state. *Gerodontology* 2009;26(2):137–42, doi: 10.1111/j.1741-2358.2008.00239.x.
- [25] Dejean D, Giacomini M, Vanstone M, Brundisini F. Patient experiences of depression and anxiety with chronic disease: a systematic review and qualitative meta-synthesis. *Ontario health technology assessment series* 2013;13(16):1–33.
- [26] Atkins D, Best D, Briss PA, Eccles M, Falck-Ytter Y, Flottorp S, et al. Grading quality of evidence and strength of recommendations. *BMJ* 2004;328(7454):1490, doi: 10.1136/bmj.328.7454.1490.
- [27] Guyatt GH, Oxman AD, Vist GE, Kunz R, Falck-Ytter Y, Alonso-Coello P, et al. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. *Bmj* 2008;336(7650):924–6, doi: 10.1136/bmj.39489.470347.AD.
- [28] Schünemann HJ, Oxman AD, Brozek J, Glasziou P, Jaeschke R, Vist GE, et al. Grading quality of evidence and strength of recommendations for diagnostic tests and strategies. *Bmj* 2008;336(7653):1106–10, doi: 10.1136/bmj.39500.677199.AE.
- [29] Brunetti M, Shemilt I, Pregno S, Vale L, Oxman AD, Lord J, et al. GRADE guidelines: 10 Considering resource use and rating the quality of economic evidence. *J Clin Epidemiol* 2013;66(2):140–50, <http://dx.doi.org/10.1016/j.jclinepi.2012.04.012>.
- [30] Perleth M, Matthias K, Langer G, Meerpohl JJ, Gartlehner G, Kaminski-Hartenthaler A, et al. GRADE-Leitlinien: 10 Den Ressourcenverbrauch berücksichtigen und die Qualität ökonomischer Evidenz bewerten. *Z Evid Fortbild Qual Gesundhwes* 2013;107(3):256–68, <http://dx.doi.org/10.1016/j.zefq.2013.04.006>.
- [31] Iorio A, Spencer FA, Falavigna M, Alba C, Lang E, Burnand B, et al. Use of GRADE for assessment of evidence about prognosis: rating confidence in estimates of event rates in broad categories of patients. *Bmj* 2015;350:h870, doi: 10.1136/bmj.h870.
- [32] Lewin S, Glenton C, Munthe-Kaas H, Carlsen B, Colvin CJ, Gulmezoglu M, et al. Using qualitative evidence in decision making for health and social interventions: an approach to assess confidence in findings from qualitative evidence syntheses (GRADE-CERQual). *PLoS medicine* 2015;12(10):e1001895, doi: 10.1371/journal.pmed.1001895.
- [33] Akl EA, Grant BJ, Guyatt GH, Montori VM, Schünemann HJ. A decision aid for COPD patients considering inhaled steroid therapy: development and before and after pilot testing. *BMC medical informatics and decision making* 2007;7:12, doi: 10.1186/1472-6947-7-12.
- [34] Guyatt GH, Oxman AD, Kunz R, Falck-Ytter Y, Vist GE, Liberati A, et al. Going from evidence to recommendations. *BMJ* 2008;336(7652):1049–51, doi: 10.1136/bmj.39493.646875.AE.
- [35] Alonso-Coello P, Schünemann HJ, Moberg J, Brignardello-Petersen R, Akl EA, Davoli M, et al. GRADE Evidence to Decision (EtD) frameworks: a systematic and transparent approach to making well informed healthcare choices. 1 Introduction. *BMJ* 2016;353:i2016, doi: 10.1136/bmj.i2016.
- [36] Nussbaumer-Streit B, Grillich L, Glehner A, Affengruber L, Gartlehner G, Morche J, et al. GRADE: Von der Evidenz zur Empfehlung oder Entscheidung – ein systematischer und transparenter Ansatz, um gut informierte Entscheidungen im Gesundheitswesen zu treffen. 1 Einleitung. *Z Evid Fortbild Qual Gesundhwes* 2018;134:57–66, <http://dx.doi.org/10.1016/j.zefq.2018.05.004>.
- [37] Schünemann HJ, Mustafa R, Brozek J, Santesso N, Alonso-Coello P, Guyatt G, et al. GRADE Guidelines: 16 GRADE evidence to decision frameworks for tests in clinical practice and public health. *J Clin Epidemiol* 2016;76:89–98, <http://dx.doi.org/10.1016/j.jclinepi.2016.01.032>.
- [38] Morche J, Conrad S, Passon A, Perleth M, Gartlehner G, Meerpohl JJ, et al. GRADE-Leitlinien: 16 Von der Evidenz zur Empfehlung oder Entscheidung – Vorgehen nach GRADE für Tests in der klinischen Praxis und in Public Health. *Z Evid Fortbild Qual Gesundhwes* 2018;133:58–66, <http://dx.doi.org/10.1016/j.zefq.2018.03.004>.
- [39] Schünemann HJ, Hill SR, Kakad M, Vist GE, Bellamy R, Stockman L, et al. Transparent development of the WHO rapid advice guidelines. *PLoS medicine* 2007;4(5):e119, doi: 10.1371/journal.pmed.0040119.
- [40] Kelson M, Akl EA, Bastian H, Cluzeau F, Curtis JR, Guyatt G, et al. Integrating values and consumer involvement in guidelines with the patient at the center: article 8 in Integrating and coordinating efforts in COPD guideline development

- An official ATS/ERS workshop report. *Proc Am Thorac Soc* 2012;9(5):262–8, doi: 10.1513/pats.201208-061ST.
- [41] Balshem H, Helfand M, Schunemann HJ, Oxman AD, Kunz R, Brozek J, et al. GRADE guidelines: 3 Rating the quality of evidence. *J Clin Epidemiol* 2011;64(4):401–6, <http://dx.doi.org/10.1016/j.jclinepi.2010.07.015>.
- [42] Meerpohl JJ, Langer G, Perleth M, Gartlehner G, Kaminski-Hartenthaler A, Schunemann H. GRADE-Leitlinien: 3 Bewertung der Qualität der Evidenz (Vertrauen in die Effektschätzer). *Z Evid Fortbild Qual Gesundheitswes* 2012;106(6):449–56, <http://dx.doi.org/10.1016/j.zefq.2012.06.013>.
- [43] Yepes-Nunez JJ, Zhang Y, Xie F, Alonso-Coello P, Selva A, Schunemann H, et al. Forty-two systematic reviews generated 23 items for assessing the risk of bias in values and preferences' studies. *J Clin Epidemiol* 2017;85:21–31, <http://dx.doi.org/10.1016/j.jclinepi.2017.04.019>.
- [44] Guyatt GH, Oxman AD, Sultan S, Glasziou P, Akl EA, Alonso-Coello P, et al. GRADE guidelines: 9 Rating up the quality of evidence. *J Clin Epidemiol* 2011;64(12):1311–6, <http://dx.doi.org/10.1016/j.jclinepi.2011.06.004>.
- [45] Kien C, Gartlehner G, Kaminski-Hartenthaler A, Meerpohl JJ, Flamm M, Langer G, et al. GRADE-Leitlinien: 9 Heraufstufen der Qualität der Evidenz. *Z Evid Fortbild Qual Gesundheitswes* 2013;107(3):249–55, <http://dx.doi.org/10.1016/j.zefq.2013.04.007>.
- [46] Guyatt G, Oxman AD, Akl EA, Kunz R, Vist G, Brozek J, et al. GRADE guidelines: 1 Introduction-GRADE evidence profiles and summary of findings tables. *J Clin Epidemiol* 2011;64(4):383–94, <http://dx.doi.org/10.1016/j.jclinepi.2010.04.026>.
- [47] Langer G, Meerpohl JJ, Perleth M, Gartlehner G, Kaminski-Hartenthaler A, Schunemann HJ. GRADE-Leitlinien: 1 Einführung - GRADE-Evidenzprofile und Summary-of-Findings-Tabellen. *Z Evid Fortbild Qual Gesundheitswes* 2012;106(5):357–68, <http://dx.doi.org/10.1016/j.zefq.2012.05.017>.
- [48] Karanickolas PJ, Montori VM, Devereaux PJ, Schunemann H, Guyatt GH. A new 'mechanistic-practical' framework for designing and interpreting randomized trials. *J Clin Epidemiol* 2009;62(5):479–84, <http://dx.doi.org/10.1016/j.jclinepi.2008.02.009>.
- [49] Levin KA. Study design III: Cross-sectional studies. Evidence-based dentistry 2006;7(1):24–5, <http://dx.doi.org/10.1038/sj.ebd.6400375>.
- [50] Fincham JE. Response rates and responsiveness for surveys, standards, and the journal. *American journal of pharmaceutical education* 2008;72(2):43, doi: 10.5688/aj720243.
- [51] Guyatt GH, Oxman AD, Kunz R, Atkins D, Brozek J, Vist G, et al. GRADE guidelines: 2 Framing the question and deciding on important outcomes. *J Clin Epidemiol* 2011;64(4):395–400, <http://dx.doi.org/10.1016/j.jclinepi.2010.09.012>.
- [52] Langer G, Meerpohl JJ, Perleth M, Gartlehner G, Kaminski-Hartenthaler A, Schunemann H. GRADE-Leitlinien: 2 Formulierung der Fragestellung und Entscheidung über wichtige Endpunkte. *Z Evid Fortbild Qual Gesundheitswes* 2012;106(5):369–76, <http://dx.doi.org/10.1016/j.zefq.2012.05.018>.
- [53] Guyatt GH, Oxman AD, Vist G, Kunz R, Brozek J, Alonso-Coello P, et al. GRADE guidelines: 4 Rating the quality of evidence—study limitations (risk of bias). *J Clin Epidemiol* 2011;64(4):407–15, <http://dx.doi.org/10.1016/j.jclinepi.2010.07.017>.
- [54] Meerpohl JJ, Langer G, Perleth M, Gartlehner G, Kaminski-Hartenthaler A, Schunemann H. GRADE-Leitlinien: 4 Bewertung der Qualität der Evidenz - Studienlimitationen (Risiko für Bias). *Z Evid Fortbild Qual Gesundheitswes* 2012;106(6):457–69, <http://dx.doi.org/10.1016/j.zefq.2012.06.014>.
- [55] Joy SM, Little E, Maruthur NM, Purnell TS, Bridges JF. Patient preferences for the treatment of type 2 diabetes: a scoping review. *Pharmacoeconomics* 2013;31(10):877–92, <http://dx.doi.org/10.1007/s40273-013-0089-7>.
- [56] Guyatt GH, Oxman AD, Kunz R, Woodcock J, Brozek J, Helfand M, et al. GRADE guidelines: 8 Rating the quality of evidence—indirectness. *J Clin Epidemiol* 2011;64(12):1303–10, <http://dx.doi.org/10.1016/j.jclinepi.2011.04.014>.
- [57] Rasch A, Perleth M, Langer G, Meerpohl JJ, Gartlehner G, Kaminski-Hartenthaler A, et al. GRADE Leitlinien: 8 Einschätzung der Qualität der Evidenz - Indirektheit. *Z Evid Fortbild Qual Gesundheitswes* 2012;106(10):745–53, <http://dx.doi.org/10.1016/j.zefq.2012.10.019>.
- [58] Holbrook A, Labiris R, Goldsmith CH, Ota K, Harb S, Sebaldt RJ. Influence of decision aids on patient preferences for anticoagulant therapy: a randomized trial. *Cmaj* 2007;176(11):1583–7, doi: 10.1503/cmaj.060837.
- [59] Emberton M. Medical treatment of benign prostatic hyperplasia: physician and patient preferences and satisfaction. *International journal of clinical practice* 2010;64(10):1425–35, doi: 10.1111/j.1742-1241.2010.02463.x.
- [60] Brazier JE, Yang Y, Tsuchiya A, Rowen DL. A review of studies mapping (or cross walking) non-preference based measures of health to generic preference-based measures. *The European journal of health economics*: HEPAC: health economics in prevention and care 2010;11(2):215–25, <http://dx.doi.org/10.1007/s10198-009-0168-z>.