Gute Praxis arbeitsepidemiologische systematische Reviews (GPAR), Stand: 22.03.2025

Langversion

Deutsche Gesellschaft für Epidemiologie (DGEpi)

Deutsche Gesellschaft für Arbeits- und Umweltmedizin (DGAUM)

Deutsche Gesellschaft für Medizinische Informatik, Biometrie und Epidemiologie (GMDS)

Deutsche Gesellschaft für Sozialmedizin und Prävention (DGSMP)

Netzwerk Evidenzbasierte Medizin (EbM-Netzwerk)

2025

Arbeitsgruppenleitung und redaktionelle Bearbeitung des Empfehlungspapiers

Name	Affiliation
Seidler, Andreas	Institut und Poliklinik für Arbeits- und Sozialmedizin (IPAS), TU Dresden
Hegewald, Janice	Bundesanstalt für Arbeitsschutz und Arbeitsmedizin

Mitarbeit (alphabetisch sortiert):

Name	Affiliation
Behrens, Thomas	Institut für Prävention und Arbeitsmedizin der Deutschen Gesetzlichen Unfallversicherung, Ruhr-Universität Bochum
Bolm-Audorff, Ulrich (Leitung Unterarbeitsgruppe Risk of Bias)	Institut und Poliklinik für Arbeits- und Sozialmedizin (IPAS), TU Dresden
Drossard, Claudia	Bundesanstalt für Arbeitsschutz und Arbeitsmedizin
Freiberg, Alice (Leitung Unterarbeitsgruppe GRADE, Beteiligung an der redaktionellen Bearbeitung des Empfehlungspapiers)	Institut und Poliklinik für Arbeits- und Sozialmedizin (IPAS), TU Dresden
Gabriel, Katharina	Bundesanstalt für Arbeitsschutz und Arbeitsmedizin
Gefeller, Olaf	Institut für Medizininformatik, Biometrie und Epidemiologie, Friedrich-Alexander-Universität Erlangen-Nürnberg
Girbig, Maria (Leitung Unterarbeitsgruppe Literaturrecherche)	Institut und Poliklinik für Arbeits- und Sozialmedizin (IPAS), TU Dresden
Jöckel, Karl-Heinz	Institut für Medizinische Informatik, Biometrie und Epidemiologie, Universität Duisburg Essen
Petereit-Haack, Gabriela	Dezernatsleiterin Landesgewerbeärztin Hessen
Romero Starke, Karla (Leitung Unterarbeitsgruppe Metaanalyse)	Institut und Poliklinik für Arbeits- und Sozialmedizin (IPAS), TU Dresden

Schlattmann, Peter	Institut für Medizinische Statistik, Informatik und Datenwissenschaften, Universitätsklinikum Jena
Schröder, Christin	Bundesanstalt für Arbeitsschutz und Arbeitsmedizin (BAuA)
Schubert, Melanie (Leitung Unterarbeitsgruppe Datenextraktion)	Institut und Poliklinik für Arbeits- und Sozialmedizin (IPAS), TU Dresden
Smolinska, Joanna	Bundesanstalt für Arbeitsschutz und Arbeitsmedizin (BAuA)
Wendt, Andrea	Bundesanstalt für Arbeitsschutz und Arbeitsmedizin (BAuA)
Wolf, Rebecca (Co-Leitung Unterarbeitsgruppe Literaturrecherche)	Bundesanstalt für Arbeitsschutz und Arbeitsmedizin (BAuA)

Wir danken herzlich **Isabelle Kaiser** und **Sabrina Schlesinger** für ihre hilfreichen Kommentare!

Inhaltsverzeichnis

Ziele und Zielgruppe der GPAR

Diese Empfehlungen richten sich primär an Epidemiolog:innen, Public Health-Expert:innen, Arbeitsmediziner:innen und Umweltmediziner:innen sowie weitere Fachwissenschaftler:innen, die auf der Grundlage arbeitsepidemiologischer oder umweltepidemiologischer Beobachtungsstudien systematische Reviews durchführen. Darüber hinaus richten sich diese Empfehlungen an politisch Entscheidungstragende, die auf der Grundlage arbeitsepidemiologischer systematischer Reviews Grenzwerte ableiten, neue Erkenntnisse zu etwaigen Berufskrankheiten bewerten oder präventive Maßnahmen begründen möchten. Schließlich richten sich diese Empfehlungen auch an forschungsfördernde Institutionen sowie Auftraggebende von arbeitsepidemiologischen systematischen Reviews.

Evidenzbasierte arbeitsmedizinische und arbeitsepidemiologische Forschung dient nicht nur dazu, ätiologische Zusammenhänge zwischen beruflicher Exposition und gesundheitlichen Folgen zu ermitteln und präventive Ansatzpunkte zu identifizieren, sondern ist auch eine wichtige Entscheidungshilfe für die Gesetzgebung. Evidenzbasierte arbeitsmedizinische Forschung stellt die bestmögliche Zusammenfassung des verfügbaren Wissens dar und bildet eine Grundlage, die die politische und juristische Entscheidungsfindung beeinflussen kann, z. B. wenn es darum geht, welche und wie viel Exposition in welcher Höhe zu einer Berufskrankheit führt oder welche Expositionsgrenzwerte für am Arbeitsplatz verwendete Chemikalien erforderlich sind.

Da evidenzbasierte Methoden aus dem Bereich der klinischen Medizin stammen, sind viele Instrumente, Handbücher und Leitfäden ursprünglich auf die Synthese randomisierter kontrollierter Studien (RCTs) und anderer Formen von Interventionsstudien ausgerichtet. Das bedeutet, dass ihre Eignung für arbeitsepidemiologische Forschungsfragen auf Beobachtungsstudien eingeschränkt ist. Aus diesem Grund wurden im Laufe der Jahre methodische Anpassungen entwickelt. Sie sind in zahlreichen Publikationen verstreut beschrieben, zum Teil in den Methodenteilen von systematischen Reviews. Diese Situation erschwert es selbst Personen mit umfassender Erfahrung in der Durchführung von systematischen Reviews, einen Überblick über die gute Praxis zu erlangen und zu behalten.

Darüber hinaus steht die evidenzbasierte Forschung wie alle Formen der Wissenschaft vor dem Dilemma des "Gartens der sich gabelnden Wege" (Kale et al. 2019). Obwohl die evidenzbasierte Forschung eine Systematik verwendet, um empirische Forschung zu identifizieren und zusammenzufassen, gibt es immer noch viele Optionen, wie die einzelnen Schritte durchgeführt werden.

Empfehlungen für die Gute Praxis können dazu beitragen, dass evidenzbasierte Forschung zu Gesundheit und Sicherheit am Arbeitsplatz nachvollziehbare, reproduzierbare und belastbare Ergebnisse liefert. Aus diesem Grund hat sich eine Arbeitsgruppe gebildet, um die derzeit verfügbaren Instrumente zu sichten und auf der Grundlage der kollektiven

Erfahrungen mit diesen Instrumenten Empfehlungen zu formulieren. Diese Empfehlungen sollen eine Orientierung für die Durchführung aussagekräftiger systematischer Reviews zu ätiologischen arbeitsepidemiologischen Forschungsfragen geben. Grundsätzlich können diese Empfehlungen auch für systematische Reviews zur ätiologischen Bedeutung von Umweltexpositionen auf konkrete Gesundheits-Outcomes Anwendung finden. Arbeitsbezogene systematische Reviews zu präventiven Interventionen gehören nicht zum Anwendungsbereich dieser Empfehlungen; diesbezüglich wird auf die entsprechenden aktuellen Cochrane-Empfehlungen verwiesen (siehe Cochrane Handbook, ROBINS-I).

In 2-3 Jahren nach Erstveröffentlichung soll eine Prüfung auf notwendige Aktualisierungen erfolgen, sofern erforderlich mit der Erstellung eines Nachtrags zu den Empfehlungen. In 5-6 Jahren nach Erstveröffentlichung soll eine überarbeitete neue Version der Empfehlungen erstellt werden. Die Weiterentwicklung des Verfahrens wird von den federführenden Autor:innen Andreas Seidler und Janice Hegewald zusammen mit den Ko-Autorinnen der Empfehlungen und der AG Epidemiologie in der Arbeitswelt von DGEpi, DGAUM, DGSMP und GMDS durchgeführt. Die Updates werden mit den beteiligten Fachgesellschaften abgestimmt. Im Rahmen der Überarbeitung der Empfehlungen nach 5-6 Jahren werden die zeitlichen Intervalle weiterer Anpassungen festgelegt. Sofern Andreas Seidler und/oder Janice Hegewald die weitere Anpassung der Living Guidelines nicht mehr federführend begleiten können, kümmern sie sich um die Übergabe der Verantwortlichkeiten an neue federführende Autor:innen.

Generell wird empfohlen, aktuelle Entwicklungen im Bereich der Durchführung systematischer Reviews bei der Erstellung arbeitsepidemiologischer Reviews zu berücksichtigen. Hier wird beispielsweise auf das Cochrane Handbuch, auf das angekündigte neue GRADE-Handbuch sowie auf wichtige Websites wie https://www.riskofbias.info/, https://www.equator-network.org/reporting-guidelines/, https://www.prisma-statement.org/ und https://www.latitudes-network.org/ hingewiesen.

Die Empfehlungen sind nach den Arbeitsschritten des Reviews gegliedert und orientieren sich an den Formulierungen von Leitlinien der Arbeitsgemeinschaft der Wissenschaftlichen Medizinischen Fachgesellschaften e. V. (AWMF):

- kann = offene Empfehlung
- sollte / sollte nicht = empfohlen
- soll / soll nicht = stark empfohlen.

1. Reviewprotokoll, Literatursuche und Screening

Innerhalb systematischer Reviews werden vorhandene Erkenntnisse zusammengefasst, um eine Forschungsfrage zuverlässig und transparent zu beantworten. Im Vordergrund stehen dabei die Transparenz und Reproduzierbarkeit, mit dem Ziel Fehler und Verzerrungen zu reduzieren und Updates von vorhandenen Reviews zu erleichtern.

Damit alle relevanten Studien systematisch identifiziert werden können, sind eine umfassende und sorgfältige Entwicklung der Suchstrategie und eine Definition der Auswahlkriterien Voraussetzung. Im nächsten Schritt des Reviewprozesses werden anhand einer klar formulierten Forschungsfrage alle primären Studien zum Forschungsthema systematisch gesucht und nach festgelegten Auswahlkriterien ausgewählt. Das Verfassen und die Veröffentlichung des Reviewprotokolls vor der eigentlichen Reviewdurchführung, die Literatursuche und das Screening bilden die Grundlage für die weiteren Schritte im Reviewprozess.

Allgemein ist darauf hinzuweisen, dass Autor:innen des systematischen Reviews entsprechend AMSTAR 2 (Shea et al. 2017, Item 16) ihre eigenen Interessenkonflikte und die Finanzierung ihres Reviews angeben sollen.

Empfehlung 1.1

Die Definition der Forschungsfrage sollte anhand von Schlüsselkomponenten erfolgen. Dazu wird die Nutzung der P(I)ECOS-Kriterien (P: Population; (I: Intervention); E: Exposition; C: Vergleichsgruppe; O: Outcome; S: Studiendesign) empfohlen.

Begründung

Jedes systematische Review wird auf Grundlage einer klar formulierten Forschungsfrage erstellt. Durch die Nutzung der P(I)ECOS-Kriterien soll eine hinreichende Eindeutigkeit und Fokussierung sowie eine Operationalisierung der Forschungsfrage erreicht werden (Morgan et al. 2018, Office of Health Assessment and Translation 2015, Thomas et al. 2022). Innerhalb arbeitsepidemiologischer Fragestellungen stehen i. d. R. die Schlüsselkomponenten Exposition am Arbeitsplatz (anstelle Intervention) und Outcome im Vordergrund. Arbeitsmedizinische Fragestellungen können Risiken am Arbeitsplatz, Prävention von arbeitsbedingten Erkrankungen sowie Maßnahmen zur Gesundheitsförderung am Arbeitsplatz umfassen. Die berücksichtigten Studiendesigns sollten für die Beantwortung der jeweiligen Fragestellung geeignet sein. Typische Studiendesigns für arbeitsepidemiologische und arbeitsmedizinische Fragestellungen sind Beobachtungs- und Interventionsstudien.

Empfehlung 1.2

Für alle systematischen Reviews soll *a priori* ein Reviewprotokoll erstellt und veröffentlicht werden.

Begründung

Das Reviewprotokoll dient der Erhöhung der Objektivität und der Transparenz im Reviewprozess, der Vermeidung selektiver Ergebnisberichterstattung sowie nicht notwendiger Duplizierung von Reviews (Hempel et al. 2016). Im Reviewprotokoll werden vorab die Forschungsfrage(n) und die geplanten einzelnen Phasen des Reviewprozesses detailliert dokumentiert. (Nachträgliche) Änderungen im Vorgehen sind transparent darzustellen (Lefebvre et al. 2022).

Die Veröffentlichung eines Reviewprotokolls erfolgt vorzugsweise auf der kostenlosen Plattform PROSPERO (internationales, prospektives Register für systematische Reviews; Link: https://www.crd.york.ac.uk/prospero/). Alternativ ist eine Registrierung auf der OSF-Plattform (Open Science Framework; Link: https://osf.io/) möglich (Pieper und Rombey 2022). Die PRISMA-Checkliste (Preferred Reporting Items for Systematic Reviews and Metaanalyses) verlangt für die Veröffentlichung eines systematischen Reviews Angaben zum Studienregister und zur Registrierungsnummer bzw. Registrierungslink.

Empfehlung 1.3

Unabhängig vom Thema sollen zumindest die Datenbanken Embase und MEDLINE sowie Web of Science in die Datenbankrecherche einbezogen werden.

Begründung

Ziel der Recherche ist es, so viele relevante Studien wie möglich zu identifizieren und einen Selektionsbias zu vermeiden. Es sollen daher stets mehre Fachdatenbanken durchsucht werden, um eine unterschiedliche Indexierung von Zeitschriften und Literaturquellen in den verschiedenen Fachdatenbanken zu berücksichtigen (Relevo 2012). Die Kombination der Datenbanken Embase, MEDLINE und Web of Science im Rahmen der Recherche gewährleisten eine angemessene und effiziente Auffindbarkeit relevanter Studien (Bramer et al. 2017). In begründeten Einzelfällen kann auf den Einbezug von Web of Science verzichtet werden. In Abhängigkeit des Forschungsfeldes und der Forschungsfrage ist zusätzlich eine Recherche in geeigneten fachspezifischen Datenbanken wünschenswert (Lefebvre et al. 2022). Bei der Recherche nach Interventionsstudien soll außerdem die Datenbank Cochrane Central Register of

Controlled Trials (CENTRAL) berücksichtigt werden (Bramer et al. 2017, Lefebvre et al. 2022).

Empfehlung 1.4

Neben der Recherche in Datenbanken sollen mindestens zwei weitere Informationsquellen bzw. Recherchestrategien genutzt werden.

Begründung

Die alleinige Suche in Datenbanken genügt nicht, um alle relevanten Referenzen einer Fragestellung zu identifizieren (Bramer et al. 2017, Lefebvre et al. 2022). Daher sind weitere Informationsquellen bzw. Recherchestrategien zu berücksichtigen.

Weitere mögliche Informationsquellen bzw. Recherchestrategien sind:

1. Suche nach weiteren Studien in Referenzlisten der identifizierten Studien und Schlüsselartikeln
2. Graue Literatur[1] (Berichte, Dissertationen, Diplom-/Master-/Bachelorarbeiten und Abstracts einschlägiger Konferenzen)
3. Expert:innen befragen
4. Suche nach Studien in früheren Reviews zum gleichen Thema
5. Citation Tracking (Vorwärts- und Rückwärtssuche)
6. Recherche in Google Scholar (erste 200 Referenzen)
7. laufende Studien (zum Verweis) und nicht publizierte Ergebnisse (z. B. Recherche in Studienregistern; Kontaktierung relevanter Autor:innen und Organisationen)
8. Suche auf Pre-Print Servern z. B. medRxiv oder bioRxiv

Nach Shea et al. (2017) müssen die unter 1. bis 3. aufgeführten zusätzlichen Suchverfahren durchgeführt werden, damit das AMSTAR 2-Kriterium 4 erfüllt ist. Außerdem ist Voraussetzung für die Erfüllung dieses Kriteriums, dass die Suche maximal 24 Monate vor Fertigstellung des systematischen Reviews durchgeführt wurde.

[1] z. B. über BASE (Link: https://www.base-search.net) oder OpenDOAR (Link: https://v2.sherpa.ac.uk/opendoar/)

Empfehlung 1.5

Es sollte eine sensitive Recherchestrategie entwickelt werden.

Begründung

Die Recherchen für systematische Reviews sollten so umfangreich wie möglich sein, um sicherzustellen, dass möglichst viele relevante Studien in die Überprüfung einbezogen werden. Dennoch ist eine angemessene Spezifität der Recherche anzustreben (Lefebvre et al. 2022).

Die Suchkomponenten orientieren sich an den P(I)ECOS-Kriterien, wobei Exposition und Outcome bei arbeitsepidemiologischen Fragestellungen i. d. R. die größte Bedeutung zukommen (Hempel et al. 2016). Innerhalb jeder Suchkomponente sind datenbankspezifische Schlagwörter (z. B. Medical Subject Headings (MeSH) in PubMed oder Embase Subject Headings (Emtree) in Embase) mit Suchbegriffen über die Freitextfunktion zu kombinieren. Für eine effektive Suchstrategie sollten Suchbegriffe und Sets von Suchbegriffen mit den Booleschen Operatoren „AND“ und „OR“ verbunden werden (Lefebvre et al. 2022). Zudem können datenbankspezifische Operatoren genutzt werden, z. B. Wordabstandsoperatoren (auch Proximity-Operatoren genannt) (Hempel et al. 2016).

Ein- und Ausschlusskriterien sollten parallel zur Entwicklung der Suchstrategie definiert werden. Es ist empfehlenswert die Suchstrategie z. B. anhand der PRESS-Checkliste (McGowan et al. 2016), durch eine andere Person (Hempel et al. 2016) oder durch einen Pre-Test zu überprüfen. Im Pre-Test kann beispielsweise überprüft werden, ob vorab festgelegte Schlüsselstudien mit der Suchstrategie identifiziert werden.

Es können auch bereits veröffentlichte Suchstrings im Bereich Arbeitsmedizin, Arbeitsepidemiologie, Arbeitssicherheit und Gesundheit herangezogen und für die jeweilige Fragestellung erweitert werden:

1) Identifikation von Artikeln zu arbeitsbedingten Erkrankungen für die Benutzeroberfläche PubMed (Mattioli et al. 2010)

2) Identifikation von Artikeln zu berufsbedingten Erkrankungen für Landarbeiter:innen (Mattioli et al. 2013)

3) berufliche Gesundheitsinterventionen in MEDLINE (Verbeek et al. 2005) (sensible und spezifische Suchstrategie)

Empfehlung 1.6

Das Screening (von Titel-Abstract & Volltext) soll anhand der festgelegten Ein- und Ausschlusskriterien erfolgen und durch mindestens zwei Reviewautor:innen unabhängig voneinander vorgenommen werden. Vor dem eigentlichen Screening soll das Verfahren in einer Pilotphase optimiert werden. Diskrepanzen werden zwischen den Reviewer:innen und/oder unter Einbezug weiterer Personen gelöst. Das Vorgehen zur Entscheidung über den Ein-/Ausschluss in der Screeningphase soll transparent beschrieben werden.

Begründung

Durch den Einbezug von zwei Reviewer:innen und ggf. einer weiteren Personen in den Screeningprozess werden zufällige und systematische Fehler reduziert (Muka et al. 2020, Page et al. 2021, Lefebvre et al. 2022). Der Grad der Übereinstimmung zwischen den Reviewerautor:innen kann berechnet werden (z. B. mit Hilfe des prozentualen Übereinstimmungsgrades oder des Cohens Kappa-Wertes).

Verschiedene Softwareanwendungen (Rayyan, EPPI-Reviewer, Covidence, Distiller SR) können das Screening erleichtern (Muka et al. 2020). Durch die Visualisierung und die verfügbare Vergleichsfunktion können Unstimmigkeiten zeitnah identifiziert und gelöst werden. Zudem verfügen diese Reviewsoftwaretools über erste künstliche Intelligenz-Ansätze, die besonders im Bereich der Literatursichtung nutzbar sind. Die Tools versuchen – nach bereits vorgenommenen Entscheidungen zu Ein- oder Ausschluss von Titeln/Abstracts – zukünftige Entscheidungen vorauszusagen (Lernalgorithmus).

Die Erfüllung des AMSTAR 2-Kriteriums 7 („critical domain“, siehe Shea et al. 2017) ebenso wie des Items 16b der PRISMA 2020 Checklist setzt voraus, dass der systematische Review eine Liste der ausgeschlossenen Studien enthalten muss.

Empfehlung 1.7

Der Prozess der Literaturrecherche soll in einem Flussdiagramm (Flow Chart) (inkl. der Ausschlussgründe der Volltexte) dokumentiert werden. Zudem sind für alle Datenbanken die Suchstrategie sowie eine Tabelle mit allen ausgeschlossenen Volltexten (inkl. der Referenz und dem Ausschlussgrund) zugänglich zu machen.

Begründung

Der Rechercheprozess soll von den Reviewautor:innen so detailliert dokumentiert werden, dass dieser in allen einbezogenen Datenbanken reproduzierbar ist (Lefebvre et al. 2022, Rethlefsen und Page 2022). Die Nutzung des PRISMA-Flow-Diagramms sowie der PRISMA-Checkliste (Rethlefsen und Page 2022) wird für die Veröffentlichung dringend empfohlen.

2. Datenextraktion

Eine qualitativ hochwertige Datenextraktion ist unerlässlich, um einen strukturierten Überblick über die vorhandene Evidenz zu ermöglichen. Die Datenerfassung sollte daher systematisch erfolgen, und es sollten möglichst standardisierte Informationen zu vordefinierten Schlüsselaspekten aus den identifizierten Studien extrahiert werden.

Empfehlung 2.1

Die Datenextraktionstabelle soll alle Daten enthalten, die zur Beantwortung der Forschungsfrage wichtig sind.

Begründung

Durch die Extraktion relevanter Informationen wird ein umfassender Überblick über die einzelnen Studien gegeben. Zudem enthält die Datenextraktionstabelle die grundlegenden Daten für die Metaanalyse. Weiterhin sollte die Extraktionstabelle so konzipiert sein, dass alle relevanten Informationen enthalten sind und damit die Notwendigkeit minimiert wird, auf das Quelldokument (also die eingeschlossenen Studien) zurückgreifen zu müssen. Die Erstellung der Extraktionstabelle sollte durch erfahrene Personen in Zusammenarbeit mit dem gesamten Forschungsteam erfolgen.

Die Extraktionstabelle sollte als Mindestanforderungen allgemeine Angaben zur Studie (Erstautor:in, Publikationsjahr, Studienname, Land, Studiendesign), zur Studienpopulation (Rekrutierung, Einschluss- und Ausschlusskriterien, Angaben zum Beruf, Anzahl der Teilnehmenden nach Geschlecht, Response, Loss-to-Follow-Up, Alter der Teilnehmenden, Differenzierung nach Studiengruppen), Exposition (Art der Exposition mit Einheit, Erhebungsmethode und -zeitpunkt ggf. Dosis-/Schwellenwerte, Definition der Belastungsgruppen), Outcome (Definition und Erhebungsmethode und -zeitpunkt) und Darstellung der Ergebnisse (z. B. Effektschätzer mit 95% Konfidenzintervallen) mit den statistischen Methoden und Confoundern enthalten. Weiterhin sollten die Art der Studienförderung und etwaige Interessenkonflikte dargestellt werden. Eine beispielhafte Extraktionstabelle ist im Anhang A dargestellt. Die Extraktionstabelle kann entsprechend den jeweiligen studienspezifischen Belangen angepasst werden.

Empfehlung 2.2

Die Datenextraktion soll standardisiert erfolgen.

Begründung

Eine nicht standardisierte Datenextraktion birgt die Gefahr, dass Daten nicht richtig extrahiert werden, was zu einem Informationsverlust führen kann. Die Extraktionstabelle soll so aufbereitet sein, dass eindeutig dargestellt ist, welche Daten extrahiert werden sollen. Ein Leitfaden für die Extraktion, der detaillierte Anweisungen und Definitionen enthält, kann ein standardisiertes Verfahren gewährleisten. Damit lassen sich Unterschiede in der Interpretation relevanter Informationen zwischen den extrahierenden Personen minimieren. Weiterhin sollte der Leitfaden den Umgang mit fehlenden/nichtzutreffenden Kategorien sowie (sofern notwendig) den Vorgang von ergänzenden Berechnungen durch die extrahierenden Personen beschreiben. Damit kann ein einheitliches Vorgehen gewährleistet werden, und es können Anpassungen der Extraktionstabelle an studienspezifische Belange vorgenommen werden.

Es ist darauf zu achten, dass die Daten nicht einfach aus der Publikation kopiert, sondern zuvor geprüft werden (beispielsweise können fehlerhafte Responseangaben publiziert worden sein, die nicht übernommen werden dürfen).

Es sind Programme erhältlich, mittels derer eine standardisierte Extraktion durchgeführt werden kann (z. B. Covidence, EPPI-Reviewer). Der Einbezug von Softwaretools mit künstlicher Intelligenz (KI) zur Datenextraktion befindet sich derzeit noch in der Entwicklung (Schmidt et al. 2021).

Empfehlung 2.3

Die Datenextraktionstabelle kann studientypenübergreifend angefertigt werden.

Begründung

In einem arbeitsepidemiologischen systematischen Review werden i. d. R. Studien mit unterschiedlichen epidemiologischen Studiendesigns zur Beantwortung einer Forschungsfrage herangezogen. Mit einer studientypenübergreifenden Extraktionstabelle können verschiedene Studiendesigns in einer Tabelle integriert werden. Da die Anforderungen an die zu extrahierende Informationen je nach Design unterschiedlich sind, sollte die Extraktionstabelle so aufgebaut sein, dass alle notwendigen Kategorien für die entsprechenden Studiendesigns abgebildet werden können. Für ein spezifisches Studiendesign nichtzutreffende Kategorien (z. B. Loss-to-Follow-up bei Querschnitts-studien) werden dann entsprechend gekennzeichnet.

Empfehlung 2.4

Mehrere Publikationen zu einer Studie sollen in einer Extraktion zusammengefasst werden.

Begründung

Werden mehrere Publikationen zu einer Studie eingeschlossen, sollen diese zusammengefasst extrahiert werden. Es ist nicht korrekt, die einzelnen Publikationen als getrennte Studien zu extrahieren. In der Regel wird die aktuellste bzw. die thematisch relevanteste Publikation als zentrale Quelle betrachtet. Weitere für den systematischen Review relevante Publikationen aus der Studie sollten jedoch nicht verworfen werden, da sie oftmals relevante Informationen zur Beschreibung der Studiendurchführung und des Studiendesigns enthalten oder die Ergebnisse mit anderen statistischen Verfahren analysieren. Wenn mehrere Publikationen zusammen extrahiert werden, sollten in der allgemeinen Beschreibung alle Publikationen mit Erstautor:in und Jahr angegeben werden. Weiterhin empfiehlt sich, die Ergebnisse publikationsspezifisch darzustellen. Kommen innerhalb einer Studie unterschiedliche methodische Ansätze zum Einsatz, sollten diese auch gesondert nach Publikation dargestellt werden.

Empfehlung 2.5

Die Datenextraktion soll von mindestens zwei erfahrenen Personen unabhängig voneinander durchgeführt werden.

Begründung

Studien zeigen, dass eine einfache Datenextraktion durch eine Person fehleranfälliger ist, als eine doppelte Datenextraktion (Buscemi et al. 2006). Um eine hohe Qualität zu gewährleisten, sollte die Datenextraktion unabhängig von mindestens zwei erfahrenen Personen durchgeführt werden.[2] Die Einbindung von weniger erfahrenen Personen sollte nur nach Training unter Anleitung erfolgen. Es ist ein einheitliches Vorgehen der extrahierenden Personen zu gewährleisten (z. B. durch Leitfaden und regelmäßige Besprechungen). Unstimmigkeiten in den Extraktionen sollten in regelmäßigen Treffen der

[2] In einigen Projekten (z. B. in einem Rapid Review) kann eine doppelte Datenextraktion durch zwei Personen unabhängig voneinander nicht realisiert werden (Seidler et al. 2021). In solchen Fällen wird empfohlen, dass die Datenextraktion hauptsächlich durch eine erfahrene Person durchgeführt wird und dass eine zweite erfahrene Person die Extraktionen auf Vollständigkeit und Korrektheit überprüft.

extrahierenden Personen, ggf. unter Einbezug einer weiteren Person, gelöst werden.

Empfehlung 2.6

Die Datenextraktion sollte vorab in einer Pilotphase getestet und optimiert werden.

Begründung

Um eine hohe Qualität der Datenextraktion zu erreichen, soll eine Pilotierung stattfinden (Li et al. 2022). Die Pilotierung sollte auf der Grundlage einer Auswahl, der im Projekt einbezogenen Publikationen durchgeführt werden. Eine erneute Pilotierung sollte bei wesentlichen Änderungen im Prozess der Datenextraktion durchgeführt werden.

Empfehlung 2.7

Fehlende Angaben in den Publikationen können durch Kontaktaufnahme mit den Autor:innen eingeholt werden.

Begründung

Im Regelfall sollten bei fehlenden relevanten Angaben die Autor:innen der entsprechenden Publikationen kontaktiert und um Informationsübermittlung gebeten werden (Li et al. 2022, zum Vorgehen siehe auch Goossen et al. 2021). Die Bereitstellung relevanter Informationen durch die Autor:innen kann Einfluss auf die Ergebnisse, die Risk of Bias-Bewertung und das GRADE-Verfahren haben (Meursinge Reynders et al. 2019). Da dieser Prozess oftmals sehr ressourcen- und zeitintensiv ist (Young und Hopewell 2011), empfiehlt es sich, im Rahmen des Projektes vorab die Kosten und Nutzen abzuwägen und bei der Projektplanung zu berücksichtigen.

3. Risk of Bias

Die Thematik Risk of Bias in epidemiologischen Studien betrifft systematische Studienfehler (interne Validität). Nach Last (2000) wird Bias definiert als „Any trend in the collection, analysis, interpretation, publication, or review of data that can lead to conclusions that are systematically different from the truth." Ähnlich verstehen Boutron et al. (2023) unter Bias einen systematischen Fehler einer Studie, der zu einer Abweichung der Studienergebnisse von der Wahrheit führt. Die Abschätzung des Risk of Bias bei der Bewertung von Einzelstudien im Rahmen eines systematischen Reviews soll somit dazu führen, dass die systematischen Fehler einer Studie erkannt werden können. Damit soll vermieden werden, dass der systematische Review zu verfälschten Ergebnissen führt. Die

Risk of Bias-Bewertung hat für jede untersuchte Expositions-Outcome-Kombination gesondert zu erfolgen.

Die Ergebnisse der Risk of Bias-Bewertung der Studien in einem systematischen Review fließen in den Schritt der Metaanalyse ein sowie in den Schritt der Bewertung der Vertrauenswürdigkeit der Evidenz (GRADE).

Die nachfolgenden Ausführungen zur Risk of Bias-Bewertung beziehen sich vorrangig auf Beobachtungsstudien zu ätiologischen Fragestellungen. Für die Risk of Bias-Bewertung von Interventionsstudien wird auf das *Cochrane Handbook for Systematic Reviews of Interventions* (https://training.cochrane.org/handbook) verwiesen.

Empfehlung 3.1

In einem systematischen Review soll das Risk of Bias der eingeschlossenen Primärstudien von mindestens zwei Personen mit fundierten epidemiologischen Kenntnissen unabhängig voneinander beurteilt werden. Diskrepanzen werden zwischen den Reviewautor:innen und/oder unter Einbezug weiterer Personen gelöst. Vor der eigentlichen Risk of Bias-Bewertung soll das Verfahren in einer Pilotphase optimiert werden.

Begründung

Ein systematischer Review ohne Risk of Bias-Bewertung läuft Gefahr, dass seine Schlussfolgerungen nicht belastbar sind, wenn sie auf Studien mit hohem Risk of Bias basieren, die den wahren Effektschätzer deutlich unter- oder überschätzen können. Daher ist die Bewertung des Risk of Bias ein wichtiger Bestandteil systematischer Reviews von Interventionsstudien (Shea et al. 2017, Higgins et al. 2023b) und Beobachtungsstudien (Lash et al. 2014, Hempel et al. 2016, Howard et al. 2017, Arroyave et al. 2021).

Zur Wahrung einer hohen Qualität soll die Beurteilung des Risk of Bias der eingeschlossenen Primärstudien von mindestens zwei Personen mit fundierten epidemiologischen Kenntnissen unabhängig voneinander durchgeführt werden. Diskrepante Bewertungen sollen durch Diskussion zwischen den beiden Reviewautor:innen und ggf. durch das Hinzuziehen einer weiteren Person zur Einigung gelöst werden. Um den Prozess der Risk of Bias-Beurteilung zu optimieren, soll er in einer Pilotphase mit einer Auswahl an eingeschlossenen Studien getestet werden.

Empfehlung 3.2

Es sind nur Instrumente zur Beurteilung des Risk of Bias geeignet, die alle in Primärstudien potenziell auftretenden Fehlerquellen berücksichtigen.

Begründung

Die wesentlichen Fehlerquellen sind Rekrutierung und Follow-Up (Selektionsbias), die Messung der beruflichen Exposition inklusive des Umgangs mit fehlenden Werten, die objektive und präzise Messung der Erkrankung inklusive des Umgangs mit fehlenden Werten (Informationsbias), Confounding (angemessene Berücksichtigung der inhaltlich relevanten Confounder), die Qualität der epidemiologischen Analyse, Chronologie (d. h. Exposition tritt vor der Erkrankung auf), Verblindung (wenn möglich), Studienförderung und Interessenskonflikte. Diese Liste der Fehlerquellen orientiert sich insbesondere an folgenden Risk of Bias-Instrumenten: Higgins et al. (2023a), Joanna Briggs Institute (2020), Romero Starke et al. (2020), WHO (2020), Woodruff und Sutton (2014) bzw. Lam et al. (2016).

Empfehlung 3.3

Es sind nur Instrumente geeignet, die ein Gesamturteil über den Risk of Bias der jeweiligen Primärstudie ermöglichen.

Begründung

Nach AMSTAR 2, einem Instrument zur Qualitätsbewertung von systematischen Reviews, soll das Ergebnis der Risk of Bias-Bewertung der einbezogenen Primärstudien bei der Interpretation und Diskussion der Studienergebnisse, der Metaanalyse oder der Evidenzsynthese berücksichtigt werden [AMSTAR 2-Items 12 und 13, Shea et al. (2017)]. Diese Berücksichtigung wird erleichtert, wenn das Risk of Bias-Instrument ein Gesamturteil (low Risk of Bias oder high Risk of Bias) über den Risk of Bias der jeweiligen Primärstudien ermöglicht. In begründeten Fällen kann es sinnvoll sein, Studien mit einem niedrigem (*low*) und hohem (*high*) Risk of Bias in Einzeldomänen zu vergleichen.

Empfehlung 3.4

Es sollen nur Instrumente zur Beurteilung des Risk of Bias verwendet werden, die die Fehlerquellen einzeln bewerten und ein Gesamturteil unabhängig von einem Summenscore (numerische Gesamtpunktzahl) bilden.

Begründung

Risk of Bias-Instrumente, die einen Summenscore bezüglich der Studienqualität bilden, laufen Gefahr, dass mehrere Fehlerquellen mit geringem Fehlerrisiko einer Primärstudie eine Fehlerquelle mit massivem Fehlerrisiko verdecken (Arroyave et al. 2021, Seidler et al. 2021). Ein Beispiel für ein Risk of Bias-Instrument, das einen Summenscore bildet und das häufig in systematischen Reviews in der Arbeitsmedizin eingesetzt wird, ist die Newcastle-Ottawa Scale, abgekürzt auch NOS (Wells GA 2009). Die NOS und ihre Einbeziehung von Parametern, die für die methodische Bewertung epidemiologischer Studien ungeeignet sind, wurde von Stang (2010) kritisch hinterfragt.

Empfehlung 3.5

Das Instrument zur Beurteilung des Risk of Bias der eingeschlossenen Primärstudien soll sich für alle gängigen Studientypen in der Arbeitsepidemiologie eignen.

Begründung

Bei den gängigen Studientypen in arbeitsepidemiologischen systematischen Reviews handelt es sich um Kohorten-, Fall-Kohorten-, Fall-Kontroll- und Querschnittsstudien. Ein geeignetes Instrument muss in der Lage sein, den Risk of Bias in Primärstudien mit diesen Studientypen zu beschreiben. Bei ätiologischen Fragestellungen soll das Studiendesign in der Risk of Bias-Bewertung Beachtung finden.

Empfehlung 3.6

Sofern im Rahmen der Risk of Bias-Analyse eine Studie mit der Gesamtbewertung „high risk of bias“ bewertet wurde, sollte es das Risk of Bias-Instrument ermöglichen bzw. dazu auffordern, die Richtung und das Ausmaß der Verzerrung des wahren Effektschätzers anzugeben.

Begründung

Die Abschätzung der Richtung und des Ausmaßes der Verzerrung durch eine Fehlerquelle auf den Effektschätzer in einer Primärstudie und in der gesamten Evidenz ist eine wesentliche Aufgabe im Rahmen eines systematischen Reviews (Howard et al. 2017, Savitz et al. 2019, Steenland et al. 2020, Arroyave et al. 2021). Diese Abschätzung ermöglicht die Aussage, ob durch eine Fehlerquelle der relative Effektschätzer in Richtung 1 (Nulleffekt) oder in die Gegenrichtung verzerrt wird. Im ersten Fall wäre ein erhöhter (bei Risikoschätzern über 1) oder erniedrigter (bei Risikoschätzern unter 1) Effektschätzer eher

als konservativ anzusehen, im zweiten Fall wenig glaubwürdig.

Ein Beispiel für eine Fehlerquelle, die den relativen Effektschätzer i. d. R. in Richtung 1 (Nulleffekt) verzerrt, ist ein nichtdifferenzieller Messfehler der Expositionshöhe, also ein Messfehler, der von der Erkrankung unabhängig ist. Dagegen wird der Effektschätzer nicht verzerrt, wenn der Messfehler der Exposition vom Berksontyp[3] ist (Armstrong 1998). Ausnahmen von der Regel sind bei Armstrong (1998), Lash et al. (2021) und Yland et al. (2022) dargestellt, ebenso wie analytische Möglichkeiten zur Biasabschätzung in Einzelstudien.

Ein Beispiel für eine weitere Fehlerquelle mit besonderem arbeitsepidemiologischem Bezug ist der *Healthy Worker Effect*. Bei Beschäftigten finden sich häufig im Vergleich zur Gesamtbevölkerung erniedrigte Erkrankungsrisiken, z. B. für Herz-Kreislauf-, Krebs- und muskuloskelettale Erkrankungen. Diese werden meist als Ausdruck eines Selektionsbias aufgefasst, weil Beschäftigte durch den Auswahlprozess bei der Einstellung gesünder sind als die Gesamtbevölkerung. Auch verbleiben gesündere Beschäftigte länger in einem Beruf als Beschäftigte, die erkranken und ausscheiden. Zum Teil wird der *Healthy Worker Effect* auch als Ausdruck eines Confounding Bias aufgefasst, weil Gesunde eher eine exponierte Beschäftigung finden und ein geringeres Krankheitsrisiko aufweisen (Checkoway et al. 2004, Lash und Rothman 2021). Der Healthy Worker Effect kann zu einer u.U. erheblichen konservativen Verzerrung („Bias towards the null“) der Risikoschätzer führen.

Empfehlung 3.7

Das Instrument zur Beurteilung des Risk of Bias der eingeschlossenen Primärstudien sollte anwenderfreundlich sein.

Begründung

Insbesondere wenn im Rahmen des systematischen Reviews eine große Zahl von Primärstudien bezüglich des Risk of Bias zu beurteilen ist, sollte das Risk of Bias-Instrument anwenderfreundlich sein und eine Beurteilung mit einem Zeitbedarf von deutlich unter einer Stunde[4] ermöglichen.

[3] Armstrong beschreibt den Berkson-Fehler als einen Fehler, der auftritt, wenn „dieselbe approximative Exposition (Proxy) für viele Probanden verwendet wird; die tatsächlichen Expositionen variieren zufällig um diesen Proxy, wobei der Mittelwert gleich diesem Proxy ist“.

[4] im Durchschnitt und für Personen mit Erfahrung

Empfehlung 3.8

Am ehesten geeignet für die Beurteilung des Risk of Bias der eingeschlossenen Primärstudien sind derzeit – nach erfolgter Konkretisierung der Beurteilungskriterien durch die Reviewautor:innen eines systematischen Reviews – folgende Instrumente: Woodruff und Sutton (2014) bzw. Lam et al. (2016) [Navigation Guide], Romero Starke et al. (2020) und in zweiter Linie die WHO Global Air Quality Guidelines (WHO 2020). Das Risk of Bias-Instrument von Higgins et al. (2024a) (ROBINS-E) ist gut geeignet, wenn ausschließlich Kohortenstudien in einem systematischen Review zu beurteilen sind.[5]

Begründung

Hinsichtlich der Empfehlungen 3.1 bis 3.7 wurden fünf existierende Risk of Bias-Instrumente miteinander verglichen. Keines der fünf geprüften Risk of Bias-Instrumente (Woodruff und Sutton (2014) bzw. Lam et al. (2016) [Navigation Guide], WHO Global Air Quality Guidelines 2020 (WHO 2020), Romero Starke et al. (2020), Joanna Briggs Institute Appraisal Tool, Higgins et al. (2023a) [ROBINS-E]) erfüllt alle der sieben o. g. Voraussetzungen. Nach Abwägung des Für und Wider erscheinen die o. g. vier Instrumente am ehesten geeignet, den Risk of Bias in arbeitsepidemiologischen Studien abzuschätzen.

Abschlussbemerkung

Für die Abschätzung des Risk of Bias von Primärstudien im Rahmen eines systematischen Reviews ist sowohl methodische als auch themenspezifische Expertise erforderlich, um alle relevanten Fehler einer Studie im Hinblick auf die beschriebenen Risk of Bias-Domänen zu identifizieren.

Die Bedeutung der Fehlerquellen und ihr Einfluss auf die Beurteilung des Risk of Bias ist kontextabhängig, so dass in der praktischen Anwendung in der Regel kontextspezifische Konkretisierungen notwendig sind.

[5] Zukünftige Weiterentwicklungen der genannten Instrumente und Neuentwicklungen werden im Sinne eines „lebenden Dokuments“ in die Empfehlungen aufgenommen.

4. Metaanalyse

Ein systematisches Review kann sich auf eine narrative Zusammenfassung der Studienmerkmale, der Ergebnisse und des Risk of Bias begrenzen. Durch die Berechnung einer Metaanalyse, sofern die Studien inhaltlich vergleichbar sind und die weiteren Voraussetzungen erfüllt sind, wird dem Review ein wichtiger quantitativer Aspekt hinzugefügt. Bei einer Metaanalyse werden die Ergebnisse einzelner Studien statistisch kombiniert. Durch „Poolen“ der Studienergebnisse lassen sich die Richtung und die Größe des Gesamteffekts quantifizieren. Außerdem lässt sich feststellen, ob dieser Effekt in allen Studien einheitlich ist (Higgins und Green 2008).

Empfehlung 4.1

Um eine Metaanalyse durchzuführen, sollten mindestens zwei Studien vorliegen, deren Methoden (z. B. hinsichtlich Exposition und Outcome) hinreichend ähnlich sind, so dass eine Zusammenfassung sinnvoll möglich ist.

Begründung

Eine Metaanalyse ist ein statistisches Verfahren bei dem für von einzelnen Studien beobachteten Effekten ein gemeinsamer gewichteter Schätzer berechnet wird (Higgins und Green 2008, Valentine et al. 2010). Im Cochrane Handbuch heißt es dazu: „Eine Metaanalyse ist die statistische Kombination der Ergebnisse aus zwei oder mehr separaten Studien.“ (Deeks et al. 2022). In eine Metaanalyse sollten nach Möglichkeit mehr als zwei Studien eingeschlossen werden. In den Ausnahmefällen, in denen nur zwei Studien einbezogen werden, sollten die Ergebnisse kritisch diskutiert werden.

Empfehlung 4.2

Wenn keine inhaltlichen oder methodischen Gründe dagegensprechen, können verschiedene Effektmaße unter Berücksichtigung weiterer Informationen in einer Metaanalyse kombiniert werden.

Begründung

Um verschiedene Effektmaße in einer Metaanalyse kombinieren zu können, bedarf es zuvor einer Vereinheitlichung. Es existieren Formeln, um Effektmaße in andere Effektmaße zu konvertieren (Borenstein et al. 2009).

Ist das untersuchte Outcome in den untersuchten Populationen selten, können die Unterschiede zwischen den Maßen des relativen Risikos (Odds Ratio, relatives Risiko,

Rate Ratio und Hazard Ratio) als vernachlässigbar angesehen werden (Rothman et al. 2008, Cummings 2009).

Ein Outcome wird als selten angesehen, wenn es in der untersuchten Population bei weniger als 10% der Personen auftritt. Gilt diese Annahme auch für Subgruppen und verändert sich das Auftreten des Outcomes nicht über die Zeit, kann die „*rare disease assumption*“ angenommen werden (Cornfield 1951, Greenland et al. 1986).

Sollte die „*rare disease assumption*“ nicht zutreffen, schlägt Zhang (1998) ein Verfahren vor, das eine bekannte Häufigkeit des Outcomes voraussetzt. Dieses Verfahren wird mehrfach kritisiert (vgl. McNutt et al. 1999; McNutt et al. 2003). Für eine anzunehmende Prävalenz des Outcomes von 20 bis 80% schlägt VanderWeele (2020) ein konservatives Verfahren vor. Bei eingebetteten Fall-Kontroll-Studien ist keine Transformation notwendig (Greenland et al. 1982).

Empfehlung 4.3

Die Auswahl des Modells [Fixed Effect-Modell (ein fester Effekt) versus Random Effects-Modell (mehrere zufällige Effekte)] sollte in Abhängigkeit von der Studienlage (u.a. Anzahl, Setting der Einzelstudien) getroffen werden.

Begründung

Wird eine Metaanalyse durchgeführt, muss die Entscheidung getroffen werden, ob die Daten mit einem Modell mit einem festen Effekt (Fixed Effect-Modell) oder mit einem Modell mit mehreren zufälligen Effekten (Random Effects-Modell) analysiert werden soll. Da bei der Entscheidung verschiedene Aspekte bedacht werden müssen, kann keine allgemeine Empfehlung ausgesprochen werden (Higgins und Green 2008). Nach Berücksichtigung dieser Aspekte erweist sich jedoch häufig ein Random-Effects-Modell als angemessen.

Zudem kann es sinnvoll sein, Studienergebnisse von separat angegebenen Subgruppen (z. B. Ergebnisse für Männer und Frauen getrennt) vorab mit einem Fixed Effect-Modell zu poolen, bevor die Ergebnisse in der Gesamtmetaanalyse aufgenommen werden können. Sollen Studienergebnisse von Subgruppen (z. B. verschiedene Berufstätigkeiten) mit einer identischen oder sich überschneidenden Vergleichsgruppe gemeinsam in die Metaanalyse aufgenommen werden, so sind i.d.R. spezielle Auswertungsmethoden erforderlich (Higgins et al. 2024) (vgl. Empfehlung 4.4).

Empfehlung 4.4

Drei-Ebenen-Metaanalysen sollten angewendet werden, wenn abhängige Effektgrößen in den eingeschlossenen Studien vorhanden sind. Dieses Modell ermöglicht eine differenzierte Schätzung der Varianz auf verschiedenen Ebenen und gewährleistet eine präzisere Analyse bei abhängigen Datenstrukturen.

Begründung

Abhängigkeiten zwischen Effektgrößen entstehen, wenn innerhalb derselben Studie mehrere Effektgrößen berichtet werden und diese auf derselben Vergleichsgruppe basieren. Weitere Abhängigkeiten können auftreten, wenn für dieselben Individuen mehrere Outcomes analysiert oder wiederholte Messungen zu verschiedenen Zeitpunkten ausgewertet werden (Cheung, 2014; Hedges et al., 2010).

Das Drei-Ebenen-Modell erweitert klassische Modelle mit zufälligen Effekten, indem es die Heterogenität getrennt auf der Ebene innerhalb und zwischen Studien quantifiziert. Durch die Schätzung von zwei Varianzkomponenten minimiert es Verzerrungen und erhöht die Validität und Robustheit der Ergebnisse (Assink & Wibbelink, 2016; Pastor & Lazowski, 2018).

Die Anwendung von Drei-Ebenen-Metaanalysen erfordert sorgfältige Überlegungen, insbesondere bei der Modellspezifikation. Eine ausreichende Anzahl an Studien und Effektgrößen ist notwendig, um zuverlässige Schätzungen der Varianzkomponenten zu gewährleisten (Konstantopoulos, 2011). Zudem wird empfohlen, die Ergebnisse solcher Analysen mit herkömmlichen zweistufigen Modellen zu vergleichen, um sicherzustellen, dass die zusätzliche Modellkomplexität durch eine verbesserte Datenanpassung gerechtfertigt ist (Cheung, 2014).

Empfehlung 4.5

Die Bewertung der Heterogenität einer Metaanalyse sollte nicht ausschließlich am I^2-Wert festgemacht werden; als grobe Orientierung ist dieser Wert aber durchaus sinnvoll (siehe dazu auch die entsprechenden Cochrane-Ausführungen).

Begründung

In die Bewertung der Heterogenität können eine visuelle Prüfung des Forest Plots, die 95%-Prädiktionsintervalle (bei ≥5 Studien) und weitere Größen zur Abschätzung der Variabilität des Effektmaßes herangezogen werden. Der I^2-Wert liefert eine Abschätzung zur Variabilität des Effektmaßes und ist damit eine wichtige Größe zur Identifikation und Kommunikation von Heterogenität in Metaanalysen. Ein Nachteil besteht darin, dass der

I^2-Wert direkt vom Wert der Teststatistik des Cochran's Q-Tests abhängt, und daher u. U. eine geringe Power hat. Daher sind zusätzlich modellbasierte Überlegungen sinnvoll, da Metaanalysen als Spezialfall von Auswertungen mit generalisierten linearen gemischten Modellen angesehen werden können. In diesem Zusammenhang kann auf den Likelihood Ratio Test zum Modellvergleich und auf den Schätzer für die Heterogenitätsvarianz τ^2 anhand dieses Modells zurückgegriffen werden. Hier empfiehlt sich die Verwendung des *restricted maximum likelihood* (REML) Schätzers (Langan et al. 2019).

Empfehlung 4.6

Publikationsbias sollte anhand von Funnel Plots bewertet werden, mit dem Wissen, dass Symmetrie nicht unbedingt das Fehlen von Publikationsbias und Asymmetrie nicht unbedingt das Vorhandensein von Publikationsbias bedeutet. Formale Tests auf Asymmetrie sollten durchgeführt werden, wenn mindestens zehn Studien zur Verfügung stehen. Wenn Funnel Plots (ohne Durchführung formaler Tests auf Asymmetrie) mit weniger als zehn Studien erstellt werden, sollten diese mit Vorsicht interpretiert werden.

Begründung

Reporting Bias in Form von Publikationsbias wird von der Cochrane Collaboration definiert als die Veröffentlichung oder Nicht-Veröffentlichung von Forschungsergebnissen, je nach Art und Richtung der Ergebnisse (Higgins und Green 2008). Verzerrungen durch kleine Studien können auftreten, wenn kleinere Studien, die keine statistisch signifikanten Effekte zeigen, möglicherweise unveröffentlicht bleiben (Sterne et al. 2000). Kleinere Studien können auch im Vergleich zu gut finanzierten großen Studien eine geringere methodische Qualität aufweisen, was zu einer Überschätzung der untersuchten Effekte führen kann (Sterne et al. 2000). Funnel Plots ermöglichen den Reviewautor:innen eine visuelle Bewertung, ob ein "kleine Studien-Bias" in einer Metaanalyse vorhanden sein könnten. In einem Funnel Plot werden die Effektschätzer der einzelnen Studien gegen die Größe oder Präzision der einzelnen Studien dargestellt. Kleine Studien weisen eine breitere Streuung am unteren Rand des Diagramms auf, während größere Studien eine schmalere Streuung aufweisen. Das Diagramm sollte einem symmetrischen umgekehrten Trichter ähneln, wenn kein "kleine Studien-Bias" vorliegt.

Die Cochrane Collaboration weist jedoch darauf hin, dass ein symmetrischer Funnel Plot nicht zwangsläufig bedeutet, dass kein Publikationsbias vorliegt (Higgins und Green 2008). Wenn Studien auf der Grundlage von p-Werten veröffentlicht werden, werden Studien, die ganz links oder rechts liegen, mit größerer Wahrscheinlichkeit veröffentlicht als solche in der Mitte. Eine Asymmetrie ist anders herum nicht zwangsläufig mit einem Publikationsbias verbunden: Kleine Studien können von schlechter methodischer Qualität

sein, was zu überhöhten Effekten führt, es könnte eine echte Heterogenität vorliegen, oder die Asymmetrie könnte ein Artefakt oder zufällig bedingt sein (Sterne et al. 2011).

Formale Tests für Funnel-Plot-Asymmetrie können verwendet werden, wenn mindestens zehn Studien zur Verfügung stehen, da die Aussagekraft der Tests bei weniger Studien zu gering ist und es möglicherweise nicht gelingt, zwischen Zufall und echter Asymmetrie zu unterscheiden (Higgins und Green 2008). Die Testergebnisse sollten im Zusammenhang mit der visuellen Inspektion der Funnel Plots interpretiert werden. Selbst wenn ein Asymmetrietest statistisch signifikant ist, kann ein Publikationsbias wahrscheinlich ausgeschlossen werden, wenn kleine Studien tendenziell zu niedrigeren Effektschätzungen führen als größere Studien oder wenn es keine Studien mit signifikanten Ergebnissen gibt (Sterne et al. 2011). Der Egger-Test wird von Cochrane für kontinuierliche Ergebnisse empfohlen, bei denen die Auswirkungen als Mittelwertunterschiede gemessen werden (Egger et al. 1997). Für dichotome Outcomes werden die Tests von Harbord et al. (2006) und Peters et al. (2006) empfohlen. Bei erheblicher Heterogenität zwischen den Studien wird der von Rücker et al. (2008) vorgeschlagene Test empfohlen. Bei fehlender Heterogenität ist dieser Test allerdings konservativ und seine Interpretation ist noch wenig bekannt. Verfahren, einen möglichen Publication Bias (z.B. durch Trim-and-fill-Methoden oder Rosenbergs Fail-Save-Methoden) auszugleichen, werden nicht empfohlen.

Empfehlung 4.7

Wenn die Anzahl der Studien dies zulässt, sollten Sensitivitätsanalysen und/oder Subgruppenanalysen durchgeführt werden. Solche Analysen sollten *a priori* im Reviewprotokoll definiert werden, und ihre Durchführung sollte begründet werden. Mögliche Unterschiede in den Effektschätzungen zwischen Studien mit hohem und niedrigem Risk of Bias sollen untersucht werden. Analysen, die nicht im Protokoll enthalten sind (*Post-hoc*-Analysen), und Analysen, die spezifische Risk of Bias-Domänen untersuchen, sollten als explorativ angesehen werden.

Begründung

Sensitivitätsanalysen (inklusive „Leave-One-Out“-Analyse) und/oder Subgruppenanalysen können helfen, kategoriale Faktoren zu identifizieren, die einen Einfluss auf die Effektgröße haben könnten. Eine Sensitivitätsanalyse, die Studien mit hohem und niedrigem Verzerrungsrisiko vergleicht, hilft bei der GRADE-Bewertung (Guyatt et al. 2008, Hulshof et al. 2019) und sollte als nahezu unverzichtbar angesehen werden, sofern eine ausreichende Anzahl von Studien vorliegt. Insbesondere wenn ganz überwiegend Studien mit einem hohen Risk of Bias vorliegen, ist es sinnvoll zu prüfen, ob es einzelne Domänen mit bedeutendem Einfluss auf den Effektschätzer von Interesse gibt.

Um ein "data dredging" zu vermeiden, sollten solche Analysen im Reviewprotokoll vorab festgelegt werden. Andere Sensitivitätsanalysen, wie z. B. die Untersuchung spezifischer Risk of Bias-Domänen, sollten als explorativ angesehen werden (hypothesengenerierend).

Empfehlung 4.8

Meta-Regressionen sind ein Hilfsmittel, um Quellen der Heterogenität zu ermitteln. Sinnvoll für eine Meta-Regression ist es, die Auswirkung einer prädefinierten Variablen wie z. B. des Studienjahrs oder der Studienregion auf den Effektschätzer zu analysieren. Die Meta-Regression von Studientypen oder Studienqualität wird nicht empfohlen. Diese können vielmehr in stratifizierten Sensitivitätsanalysen nach Risk of Bias analysiert werden (siehe Empfehlung 4.7). Die Ergebnisse von Meta-Regressionen sind bei einer geringen Anzahl von Studien mit Vorsicht zu interpretieren.

Begründung

Meta-Regressionsanalysen können verwendet werden, um wichtige Quellen der Heterogenität zu erkennen. Diese Analysen können helfen, Faktoren zu identifizieren, die mit den beobachteten Effekten in Zusammenhang stehen (Higgins und Green 2008). Es ist jedoch zu beachten, dass wahrscheinlich nicht alle Quellen der Heterogenität gefunden werden und dass die Möglichkeit einer residualen Heterogenität besteht. Daher ist die geeignete Analyse eine Analyse mit zufälligen Effekten (Thompson und Higgins 2002).

Als Faustregel gilt, dass für die Analyse mindestens zehn Studien zur Verfügung stehen sollten (Higgins und Green 2008). Reviewautor:innen sollten im Reviewprotokoll angeben, welche Merkmale später in einer Meta-Regression untersucht werden. Die Interpretation der Ergebnisse wird erleichtert, wenn der untersuchte Faktor eine höhere Variabilität zwischen den Studien als innerhalb der Studien aufweist (Thompson und Higgins 2002). Sinnvolle zu untersuchende Faktoren sind z. B. Studienjahr oder Studienregion. Wenn ein Merkmal übersehen wurde, aber von großer Bedeutung und gerechtfertigt ist, können Reviewautor:innen es *post hoc* untersuchen. Allerdings sollten diese *Post-hoc*-Analysen als solche gekennzeichnet werden (Higgins und Green 2008) und sollten eher als hypothesengenerierend und nicht als hypothesenprüfend betrachtet werden (Thompson und Higgins 2002). Aufgrund der meist geringen Anzahl von Datenpunkten (Studien) in der Regression und weil nur Variablen untersucht werden können, die verfügbar sind, sollten die Ergebnisse der Meta-Regression im Allgemeinen mit Vorsicht interpretiert werden.

Empfehlung 4.9

Dosis-Wirkungs-Metaanalysen sollten vorzugsweise mit einem modellbasierten Ein-Schritt-Verfahren durchgeführt werden.

Begründung

Für eine Dosis-Wirkungs-Metaanalyse stehen Ein-Schritt- und Zwei-Schritt-Verfahren zur Verfügung. Bei Letzterem werden für jede individuelle Studie im einfachsten Fall lineare Dosis-Wirkungs-Beziehungen aus den Dosiskategorien bestimmt. Die individuellen Regressionskoeffizienten werden dann metaanalytisch zusammengefasst (Greenland und Longnecker 1992). Ein Nachteil dieses Vorgehens besteht darin, dass mindestens drei Dosiskategorien benötigt werden. Dieser Nachteil besteht bei dem Ein-Schritt-Verfahren nicht (Crippa et al. 2019). Hier können alle Studien genutzt werden und auch komplexe Dosis-Wirkungs-Beziehungen, z. B. mit quadratischen Termen oder Splines, modelliert werden.

5. Bewertung der Vertrauenswürdigkeit der Evidenz (GRADE)

Um die Vertrauenswürdigkeit der Evidenz in systematischen Reviews (certainty of evidence) zu beurteilen, wird die GRADE-Methode (Grading of Recommendations Assessment, Development and Evaluation) eingesetzt (Guyatt et al. 2011a). In arbeitsepidemiologischen Zusammenhängen geht es häufig darum, die Vertrauenswürdigkeit der Evidenz zur Existenz eines Zusammenhangs zwischen einer Exposition und einem Outcome zu beurteilen. Die GRADE-Methodik wurde ursprünglich für klinische Fragen im Bereich Diagnostik, Screening, Prävention und Therapie entwickelt, kann aber auch für Problemstellungen im Kontext von Gesundheitssystemen und Versorgungsforschung verwendet werden (Guyatt et al. 2011a). Für systematische Reviews aus dem Bereich der Umweltepidemiologie wurde mit dem Navigation Guide eine modifizierte GRADE-Version empfohlen (Woodruff und Sutton 2014), die mittlerweile auch in arbeitsepidemiologischen systematischen Reviews Anwendung findet (Johnson et al. 2014, Hulshof et al. 2019, Hulshof et al. 2021b, Lam et al. 2021). Die GRADE Working Group publiziert seit dem Jahr 2011 fortlaufend (bereits über 40) spezifische Empfehlungen zur Anwendung der GRADE-Methode in systematischen Reviews, die teilweise auch umwelt- und arbeitsbezogene Fragestellungen betreffen.

Bei der Anwendung formalisierter Regeln sollten immer auch die Besonderheiten der Wissensdomäne beachtet werden. Insofern sind die hier dargestellten Empfehlungen nicht zwingend bindend für die Anwendenden. Allerdings sollten Abweichungen immer wissenschaftlich begründet werden.

In den letzten Jahren hat die GRADE Working Group konzeptuelle Veränderungen publiziert, die insbesondere zur Beurteilung der Vertrauenswürdigkeit der Evidenz für

präventive Interventionen in der Arbeitswelt von Bedeutung sind. Bei systematischen Reviews werden nicht bzw. minimal kontextualisierte von partiell kontextualisierten Ansätzen („non-contextualized“ bzw. „minimally contextualized approach“ und „partially contextualized approach“) unterschieden; siehe Hultcrantz et al. (2017) bzw. Zeng et al. (2021). Zur Beantwortung der Frage nach der Existenz eines positiven (oder negativen) Zusammenhangs eignet sich der minimal kontextualisierte Ansatz. Das Nullrisiko (RR von 1,0) wird hier als Schwelle definiert, für deren Überschreitung (oder Unterschreitung) die Vertrauenswürdigkeit der Evidenz eingestuft werden soll. Sofern es um die Identifizierung eines positiven Zusammenhangs geht, wird im Folgenden auch von „hazard identification“ gesprochen.

Grundsätzlich könnte auch das für Fragestellungen zu Berufskrankheiten bedeutsame „Verdopplungsrisiko“ (RR von 2,0) als Schwelle definiert werden. Hier können Modifikationen des GRADE-Verfahrens (z. B. bei der Bewertung der Effektstärke) erforderlich werden; diesbezügliche nähere Ausführungen sind einem späteren Update dieser Empfehlungen vorbehalten.

Auf den partiell kontextualisierten Ansatz soll im Rahmen dieser Empfehlungen zunächst nicht weiter eingegangen werden. Das Grundprinzip dieses Ansatzes besteht in der Beurteilung, ob der tatsächliche Effekt in einem definierten Bereich („category of magnitude of effect“, siehe Schünemann et al. (2022b)) liegt. Die Bereiche werden von Schwellen begrenzt („thresholds“: small effect, moderate effect, large effect; siehe insbesondere Zeng et al. (2021)). Im teilweise kontextualisierten Ansatz hat die Zahl der Überschreitung solcher Schwellen durch das Konfidenzintervall Bedeutung für die Bewertung der Präzision (siehe Empfehlung 5.6). Bei Überschreitung von mindestens drei Schwellen wird von der GRADE Working Group ein Downgrade um drei Stufen vorgeschlagen (Schünemann et al. 2022b).

Empfehlung 5.1

In arbeitsepidemiologischen systematischen Reviews soll die Vertrauenswürdigkeit der Evidenz möglichst für jede gestellte Forschungsfrage gesondert (i. d. R. für jede Expositions-Wirkungs-Beziehung) eingeschätzt werden. Dabei sollte eine Einteilung in eine von drei Stufen erfolgen: hoch, moderat oder niedrig.

Begründung

Die Einschätzung der Vertrauenswürdigkeit der Evidenz erfolgt studienübergreifend jeweils zu einem Endpunkt einer spezifischen Forschungsfrage (Guyatt et al. 2011a, Hempel et al. 2016). Die Vertrauenswürdigkeit der Evidenz kann durch die fünf Domänen „Risiko für Bias“ (risk of bias), „Inkonsistenz“ (inconsistency), „Indirektheit“ (indirectness), fehlende „Präzision“ (imprecision) und „Publikationsbias“ (publication bias) herabgestuft

und durch die drei Domänen „großer Effekt“ (large magnitude of effect), „Expositions-Wirkungs-Beziehung“ (dose response) und „Residual Confounding“ (residual confounding) heraufgestuft werden. Die einzelnen Domänen sollten dabei nicht in Isolation betrachtet werden, da sie häufig miteinander zusammenhängen und sich gegenseitig beeinflussen. Bei einer bestehenden gegenseitigen Beeinflussung der Domänen sollte es nicht zu einer doppelten Auf- bzw. Abwertung kommen, sondern der vorrangige Einflussfaktor soll identifiziert und bewertet werden. Zur Beantwortung der Frage nach dem Kausalzusammenhang zwischen einer Exposition und einem Outcome sollte über die Anwendung des GRADE-Verfahrens hinaus noch die Erfüllung der von Sir Austin Bradford Hill formulierten Kausalitätskriterien geprüft werden (Hill 1965). Wenn die Reviewautor:innen abschätzen, dass eine Domäne die Vertrauenswürdigkeit der Evidenz beeinflusst, kommt es zu einem Herauf- bzw. Herunterstufen um jeweils ein bis drei Grade (+1, +2 oder +3 bzw. -1, -2 oder -3). Diese Beurteilungskriterien sind fragestellungsbezogen zu konkretisieren. Heraufstufungen und Herabstufungen sollten aufsummiert werden und bei einer resultierenden negativen Summe zu einer Absenkung der Vertrauenswürdigkeit der Evidenz und bei einer positiven Summe zu einer Heraufstufung der Vertrauenswürdigkeit der Evidenz führen. Dabei ist auch das Ausgangslevel entscheidend (vgl. Empfehlung 5.2).

Die Einstufung der Vertrauenswürdigkeit der Evidenz einer Expositions-Wirkungs-Beziehung erfolgt in Anlehnung an den Navigation Guide (Johnson et al. 2014, Woodruff und Sutton 2014, Hulshof et al. 2019, Hulshof et al. 2021b, Lam et al. 2021) in die drei Stufen „hoch“, „moderat“ und „niedrig“. Hierbei werden die von GRADE vorgesehenen Stufen „niedrig“ und „sehr niedrig“ zu „niedrig“ zusammengefasst. Grundsätzlich ist auch in Anlehnung an das Vorgehen der GRADE Working Group die Verwendung einer zusätzlichen Stufe „sehr niedrig“ möglich.

Es kann Szenarien geben, in denen die Anwendung des GRADE-Verfahrens nur eingeschränkt sinnvoll ist. Ein Beispiel für solche Szenarien stellt ein gepoolter relativer Risikoschätzer von < 1,0 (wenn es um die Frage nach einem positiven Zusammenhang geht) dar [6]. Hier werden zwei Vorgehensweisen für möglich gehalten:

1. Vorgehensweise: Das GRADE-Verfahren wird ausnahmslos für jede gestellte Forschungsfrage (also für alle gepoolten Risikoschätzer) angewendet. Bei einem gepoolten Risikoschätzer < 1,0 wird die Vertrauenswürdigkeit der Evidenz für einen negativen Zusammenhang angegeben (ggf. mit der Zusatzbemerkung, dass das Ergebnis mit einem positiven wie mit einem negativen Zusammenhang vereinbar ist).

[6] Sofern die Vertrauenswürdigkeit der Evidenz für einen positiven Zusammenhang bei einem gepoolten Risikoschätzer von genau 1,0 höher als „niedrig“ eingestuft wird, ist das GRADE-Verfahren kritisch zu überprüfen (da ein solche Einstufung nicht plausibel zu machen ist).

2. Vorgehensweise: Das GRADE-Verfahren wird lediglich angewendet, wenn sich bei der Einschätzung der Vertrauenswürdigkeit für einen positiven (bzw. negativen) Zusammenhang für eine konkrete Exposition ein gepoolter Risikoschätzer > 1,0 (bzw. < 1,0) findet. Insbesondere wenn in diesem Fall das GRADE-Verfahren nicht angewendet wird, sollte darauf in der „Summary of Findings"-Tabelle hingewiesen werden, und es sollte auf die Richtung des gepoolten Risikoschätzer in der Diskussion eingegangen werden.

Empfehlung 5.2

Die GRADE-Bewertung der Vertrauenswürdigkeit der Evidenz sollte[7] mit einem hohen Ausgangslevel beginnen.

Begründung

Diesbezüglich wird auf die Ausführungen von Steenland et al. (2020) hingewiesen. Steenland et al. (2020) zufolge sei abzulehnen, randomisierte kontrollierte Studien (RCTs) als Goldstandard für epidemiologische Beobachtungsstudien anzusehen und von daher Beobachtungsstudien von vornherein mit der Prämisse eines Bias zu belegen[8]. Diese Position wird von dem grundlegenden Beitrag von Arroyave et al. (2021) gestützt[9]. Die arbeitsepidemiologischen ILO/WHO-Reviews vertreten in der Frage des Ausgangslevels der GRADE-Bewertung keine durchgängig konsistente Position. Im Reviewprotokoll von Hulshof et al. (2019) wird zwar ein hohes Ausgangslevel für nicht-randomisierte Studien angegeben; allerdings ist aus dem Text nicht eindeutig abzuleiten, ob hier epidemiologische Beobachtungsstudien oder nicht-randomisierte Interventionsstudien adressiert sind. Im zugehörigen Ergebnispapier von Hulshof et al. (2021a) wird für randomisierte Studien ein hohes, für Beobachtungsstudien ein moderates Ausgangslevel der Vertrauenswürdigkeit der Evidenz angegeben. Eine weitere Veröffentlichung der Forschungsgruppe zur Prävalenz beruflicher Exposition gegenüber ergonomischen

[7] Die Anwendbarkeit dieser Empfehlung sollte in den nächsten 2–3 Jahren empirisch geprüft und dann ggf. angepasst werden.

[8] Steenland et al. (2020, S. 4) führen diesbezüglich aus: „*We believe there should be no a priori assumption that observational studies are weaker than RCTs for studying occupational and environmental exposures, and it should be acknowledged that they generally represent the best available evidence to assess causality… Thus, in our view, observational studies should be considered as the norm and then assigned a lower quality if significant substantial biases are likely that would affect the parameter estimates*". An dieser Stelle sei darauf hinzuweisen, dass die renommierte umweltepidemiologische Fachzeitschrift „Environmental Health Perspectives" in ihren Autorenrichtlinien (https://ehp.niehs.nih.gov/authors/reviews) zentrale Inhalte der Steenland et al. (2022)-Veröffentlichung aufnimmt und diese Veröffentlichung explizit unter „Recommended reading on systematic reviews and meta-analyses" aufführt.

[9] Arroyave et al. (2021, S. 9) konstatieren: „*Downgrading evidence based solely on study design limits the full potential and unique strengths of observational evidence in causal analysis and in resolving critical public health gaps*".

Risikofaktoren startet demgegenüber mit einem hohen Ausgangslevel (Hulshof et al. 2021b).

Die Entscheidung für ein hohes Ausgangslevel ist aus unserer Sicht vereinbar mit den Ausführungen des „Cochrane Handbook for Systematic Reviews of Interventions“[10] (Schünemann et al. 2022a). Die entsprechenden Ausführungen des „Cochrane Handbook for Systematic Reviews of Interventions“ beziehen sich auf nicht-randomisierte Interventionsstudien; die Angemessenheit eines hohen Ausgangslevels wird diesbezüglich damit begründet, dass das ROBINS I-Verfahren der Risk of Bias-Bewertung bei nicht-randomisierten Studien i. d. R. bereits eine Abstufung um zwei Levels (aufgrund von Confounding und Selektionsbias) ergibt. Daran angelehnt sieht auch das von unserer Arbeitsgruppe vorgeschlagene GRADE-Verfahren eine Abwertung von bis zu zwei Levels bei hohem Verzerrungsrisiko durch Risk of Bias vor. Von der "Sollte"-Empfehlung, mit einem hohen Ausgangslevel zu beginnen, kann in begründeten Fällen abgewichen werden.

Empfehlung 5.3

Führt das Verzerrungsrisiko der Einzelstudien zu einer Überschätzung des Effektschätzers, soll es zu einer Abwertung der Vertrauenswürdigkeit der Evidenz kommen (Domäne „Risiko für Bias“).

[10] Im Chapter 14.2. heißt es: „*In GRADE, a body of evidence from randomized trials begins with a high-certainty rating while a body of evidence from NRSI begins with a low-certainty rating. The lower rating with NRSI is the result of the potential bias induced by the lack of randomization (i.e. confounding and selection bias). However, when using the new Risk of Bias In Non-randomized Studies of Interventions (ROBINS-I) tool (Sterne et al. 2016), an assessment tool that covers the risk of bias due to lack of randomization, all studies may start as high certainty of the evidence (Schünemann et al. 2019). The approach of starting all study designs (including NRSI) as high certainty does not conflict with the initial GRADE approach of starting the rating of NRSI as low certainty evidence. This is because a body of evidence from NRSI should generally be downgraded by two levels due to the inherent risk of bias associated with the lack of randomization, namely confounding and selection bias. Not downgrading NRSI from high to low certainty needs transparent and detailed justification for what mitigates concerns about confounding and selection bias (Schünemann et al., 2018). Very few examples of where not rating down by two levels is appropriate currently exist*.“

Sterne, JA, Hernan MA, Reeves BV, Savovic J, Berkman ND, Viswanathan M, Henry D, Altman DG, Ansari MT, Boutron I, Carpenter JR, Chan AW, Churchill R, Deeks JJ, Hrobjartsson A, Kirkham J, Juni P, Loke YK, Pigott TD, Ramsay CR, Regidor D, Rothstein HR, Sandhu L, Santaguida PL, Schünemann HJ, Shea B, Shrier I, Tugwell P, Turner L, Valentine JC, Waddington H, Waters E, Wells GA, Whiting PF, Higgins JP (2016). "ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions." BMJ **355**: i4919.

Schünemann, HJ, Cuello C, Akl EA, Mustafa RA, Meerpohl JJ, Thayer K, Morgan RL, Gartlehner G, Kunz R, Katikireddi SV, Sterne J, Higgins JP, Guyatt G, GW Group (2019). "GRADE guidelines: 18. How ROBINS-I and other tools to assess risk of bias in nonrandomized studies should be used to rate the certainty of a body of evidence." J Clin Epidemiol **111**: 105-114.

Begründung

Nach Guyatt et al. (2011f) und Hulshof et al. (2019) kann die Vertrauenswürdigkeit in die Evidenz dann herabgestuft werden, wenn der Großteil der relevanten Evidenz aus Studien stammt, die ein hohes Verzerrungsrisiko (high risk of bias) aufweisen. Bei der Beurteilung der Qualität der Evidenz über mehrere Studien hinweg folgen sowohl Guyatt et al. (2011f) als auch Hulshof et al. (2019) den folgenden Prinzipien:

1. Bei der Entscheidung über die Gesamtqualität der Evidenz soll kein Durchschnittswert hinsichtlich des Risikos für Bias über alle Studien hinweg gebildet werden. Vielmehr soll der Fokus auf den Studien mit einem geringen Verzerrungsrisiko (low risk of bias) liegen.[11] [12]

2. Bei der Abwertung dieser Domäne sollte man konservativ vorgehen, das heißt, man sollte sicher sein, dass für den größten Teil der verfügbaren Evidenz ein erhebliches Verzerrungsrisiko besteht, bevor man eine Abwertung vornimmt.

3. Die Domäne „Risiko für Bias“ sollte im Kontext anderer Limitationen bewertet werden. Wenn sich die Reviewautor:innen beispielsweise bei der Bewertung von zwei Domänen (z. B. Risiko für Bias und Indirektheit) unsicher sind, sollte mindestens eine der beiden Domänen abgewertet werden.

Hulshof et al. (2019) und Guyatt et al. (2011f) geben zu bedenken, dass sich die Reviewautor:innen in einigen Fällen der Beurteilungen der Vertrauenswürdigkeit der Evidenz unsicher sein werden. Wenn dem so ist, sollten sie sowohl die Gründe für diese Unsicherheiten als auch für ihre endgültige Entscheidung angeben.

Es liegt im Ermessen der Reviewautor:innen, wann die Domäne „Risiko für Bias“ um eine (-1) oder zwei (-2) Stufen abgewertet wird, da der Navigation Guide diesbezüglich keine klare Empfehlung gibt (Hulshof et al. 2019). Guyatt et al. (2011f) erläutern dazu, dass eine Abwertung um -1 erfolgen soll, wenn die Verzerrungsrisiken der vorliegenden Evidenz das Vertrauen in den Effektschätzer beeinträchtigen. Eine Abwertung um -2 soll erfolgen, wenn die Verzerrungsrisiken das Vertrauen massiv beeinträchtigen.

[11] Hempel, S, Xenakis L, Danz M (2016). Systematic Reviews for Occupational Safety and Health Questions: Resources for Evidence Synthesis. Santa Monica, CA, RAND Corporation. führen dazu folgendes aus: *„A small number of studies with low risk of bias is more informative than a large number of studies with limited validity”*.

[12] Leave-one-out-Analysen können ergänzend sinnvoll sein, wenn beispielsweise eine Studie einen besonders hohen Einfluss auf den Gesamtschätzer hat.

Empfehlung 5.4

Beim Vorliegen einer nicht erklärbaren Variabilität (in Metaanalysen: einer nicht erklärbaren Heterogenität) kann eine Abstufung der Vertrauenswürdigkeit der Evidenz erfolgen (Domäne „Inkonsistenz“).[13]

Begründung

Die Domäne „Inkonsistenz“ bewertet die Variabilität der Effektschätzer zwischen den Studien und das Ausmaß der Heterogenität, also die *unerklärte Varianz* zwischen den Studien in Metaanalysen (Hempel et al. 2016). Die (statistische) Heterogenität kann sich dabei aus einer klinischen Heterogenität, also Unterschieden der Population, der Exposition oder der Endpunkte, und/oder aus einer methodischen Heterogenität, also der Variabilität des Studiendesigns, der Operationalisierung der Endpunkte oder des Verzerrungsrisikos, ergeben (Deeks et al. 2022).

Konkrete Ausführungen dazu, welche Verfahren genutzt werden sollten, um zu ermitteln, ob eine Heterogenität in Metaanalysen vorliegt, finden sich in der Empfehlung 4.5.

Sollte sich keine schlüssige Begründung finden lassen, worauf eine Heterogenität oder Variabilität der Studienergebnisse basiert, sinkt die Vertrauenswürdigkeit in diese Evidenz (Guyatt et al. 2011c, Hulshof et al. 2019, Schünemann et al. 2022a). Es liegt im Ermessen der Reviewautor:innen, wann die Domäne „Inkonsistenz“ um eine (-1) oder zwei (-2) Stufen herabgestuft wird, da der Navigation Guide diesbezüglich keine klare Empfehlung gibt (Hulshof et al. 2019). Ein Hochstufen ist bei Konsistenz der Studienergebnisse jedoch nicht möglich, da auch ein konsistenter Bias zu falschen Ergebnissen führen kann (Guyatt et al. 2011c).

Empfehlung 5.5

Eine Herabstufung der Vertrauenswürdigkeit der Evidenz soll dann erfolgen, wenn die Einzelstudien der Forschungsfrage des systematischen Reviews in Hinsicht auf Population, Exposition und Outcome nicht adäquat folgen (Domäne „Indirektheit“).

[13] Methodenpapiere, die sich mit der GRADE-Bewertung von systematischen Reviews aus dem Bereich der Arbeits- und Umweltepidemiologie befassen, beschreiben keine Spezifikationen zur Beurteilung der Domäne „Inkonsistenz“ (Arroyave et al. 2021, Hempel et al 2016, Morgan et al. 2016, Steenland et al. 2020, Woodruff und Sutton 2014).

Begründung

Die Domäne „Indirektheit“ stellt dar, ob die verfügbaren Studien der Forschungsfrage eines systematischen Reviews in Hinsicht auf die Population, die Exposition sowie das Outcome adäquat folgen (Guyatt et al. 2011b, Hempel et al. 2016). Damit spiegelt sie die externe Validität wider, also inwieweit das ermittelte Ergebnis tatsächlich ohne Einschränkungen auf die fragliche Zielpopulation übertragbar ist.

Nach Hulshof et al. (2019) kann die Evidenz auf drei Ebenen indirekt sein:

1. Die Population in den Einzelstudien unterscheidet sich von der interessierenden Population des systematischen Reviews. Nach Guyatt et al. (2011b) kommt es in der Regel nicht zu einer Abwertung aufgrund von Populationsunterschieden, es sei denn, es liegen zwingende Gründe für die Annahme vor, dass sich die interessierende Population des systematischen Reviews so sehr von der in den Studien untersuchten Population unterscheidet, so dass sich der Effektschätzer erheblich unterscheiden wird. Dies ist laut Guyatt et al. (2011b) in den meisten Fällen nicht zu erwarten.

2. Die in den Einzelstudien untersuchte Exposition kann von der interessierenden Exposition des systematischen Reviews abweichen. Es sollte nur dann eine Abwertung in Bezug auf die Population und Exposition erfolgen, wenn die Abweichungen so groß sind, dass sie einen Unterschied der Effektschätzer wahrscheinlich machen.

3. Die Outcomeparameter der untersuchten Studien können von denen abweichen, die in dem systematischen Review von primärem Interesse sind. Abweichungen können sich hier u. a. in Bezug auf den Zeitraum der Messung ergeben. Je nach Ausmaß der Diskrepanz zwischen dem Zeitraum der Messung in den Einzelstudien und dem interessierenden Zeitraum der Messung in dem systematischen Review erfolgt eine Herabstufung um eine oder zwei Stufen. Eine weitere Quelle der Indirektheit im Zusammenhang mit dem Outcome ist die Verwendung von Substitut- oder Surrogatparametern in den Einzelstudien zur Abbildung des interessierenden Endpunktes. Im Allgemeinen erfordert die Verwendung eines Surrogats eine Herabstufung der Vertrauenswürdigkeit der Evidenz um eine oder sogar um zwei Stufen. Dabei kann die Berücksichtigung der Biologie, des Wirkmechanismus und des natürlichen Verlaufs der Krankheit hilfreich sein. Surrogate, die auf dem Kausalpfad näher an den interessierenden Outcomes liegen, rechtfertigen eine Herabstufung um nur eine Stufe. In seltenen Fällen sind Surrogate sogar so gut etabliert, dass man die Vertrauenswürdigkeit der Evidenz nicht herabstufen sollte. In der Regel sollte Evidenz, die auf Surrogatparametern beruht, eine Herabstufung auslösen, während andere Arten der Indirektheit ein wohlüberlegtes Urteil erfordern.

Empfehlung 5.6

Bei fehlender Präzision der (in Metaanalysen gepoolten) Effektschätzer soll die Vertrauenswürdigkeit in die Evidenz um bis zu 3 Stufen herabgestuft werden. Eine Herabstufung erfolgt bei Überschreitung einer prädefinierten Schwelle durch das Konfidenzintervall (i. d. R. des Nulleffekts bei Studien zur „hazard identification“) und eventuell zusätzlich bei einem breiten Konfidenzintervall (Domäne „Präzision“).

Begründung

Die Beurteilung der Domäne „Präzision“ berücksichtigt in Anlehnung an Hulshof et al. (2019) die Breite des Konfidenzintervalls des Effektschätzers (die wiederum abhängig ist vom Stichprobenumfang ist), da dieses Aufschluss über die Auswirkungen zufälliger Fehler auf die Vertrauenswürdigkeit der Evidenz gibt. Die Einstufung erfolgt dabei bestenfalls auf Grundlage des gepoolten Effektschätzers einer Metaanalyse (Hulshof et al. 2019).

Bei der Beantwortung von ätiologischen Fragestellungen mit dem minimal kontextualisierten Ansatz wird die Überschreitung lediglich einer Schwelle (meist des Nulleffektes) beurteilt[14]. Eine Herabstufung für die Präzision um insgesamt bis zu drei Stufen (vgl. Schünemann et al. (2022b)) soll bei einem breiten Konfidenzintervall erfolgen. Mehrere Studien stufen herab, wenn bei einem positiven Risikoschätzer die Breite des Konfidenzintervalls größer 2 beträgt (u.a. Kuijer et al. (2018), Romero Starke et al. (2020), Seidler et al. (2022)). Bei kontinuierlichen Expositionen wird vorgeschlagen, die Breite des Konfidenzintervalls bei einem relativen Risikoschätzer von 2 zu beurteilen. Bei einem Risikoschätzer <1 kann die entsprechende Breite des Konfidenzintervalls auf der Grundlage der Kehrwerte der unteren und der oberen Grenze des Konfidenzintervalls beurteilt werden.

Bei Fragestellungen zu präventiven Interventionen erscheint die Anwendung des – im Rahmen dieser Empfehlungen nicht weiter ausgeführten – partiell kontextualisierten Ansatzes sinnvoll; für weitere Ausführungen zu diesem Ansatz wird auf Zeng et al. (2021) und Schünemann et al. (2022a) verwiesen.

[14] Zeng et al. (2022, Example 6) empfehlen in einer entsprechenden Situation die Herabstufung um zwei Stufen.

Empfehlung 5.7

Wenn ein Publikationsbias zu einer Überschätzung der Effekte führt, soll die Vertrauenswürdigkeit der Evidenz herabgestuft werden (Domäne „Publikationsbias“).

Begründung

Mit der Domäne „Publikationsbias“ wird überprüft, ob es Hinweise darauf gibt, dass in der ermittelten Evidenzbasis einschlägige Studien fehlen (Hempel et al. 2016), so dass die Auswahl der eingeschlossenen Studien nicht repräsentativ ist und somit die Effektschätzer verzerrt sein können, selbst wenn die identifizierten Studien perfekt angelegt und durchgeführt wurden (Guyatt et al. 2011d). In der Praxis tritt hauptsächlich das Problem auf, dass „negative“ Studien nicht berücksichtigt werden, woraus sich eine Überschätzung des Effektschätzers ergeben kann (Guyatt et al. 2011d). Dies resultiert beispielsweise daher, dass Studien gar nicht oder nicht in adäquater Form (z. B. als Abstract oder Dissertation) veröffentlicht werden, nicht in englischer Sprache publiziert werden (Guyatt et al. 2011d) oder dass wissenschaftliche Zeitschriften die Entscheidung zur Veröffentlichung nicht unabhängig vom Ergebnis der Studie treffen (Hempel et al. 2016).

Genaue Ausführungen dazu, welche statistischen Verfahren genutzt werden sollten, um zu ermitteln, ob ein Publikationsbias vorliegt, finden sich in der Empfehlung 4.6.

Es liegt im Ermessen der Reviewautor:innen, wann die Domäne „Publikationsbias“ um eine (-1) oder zwei (-2) Stufen abgewertet wird, da der Navigation Guide diesbezüglich keine klare Empfehlung gibt (Hulshof et al. 2019). Nach Guyatt et al. (2011d) sollte die Herabstufung in der Regel nur um eine (-1), nicht um zwei Stufen (-2) erfolgen.

Empfehlung 5.8

Im Falle von binären Expositionen, bei denen die Vergleichsgruppe (Komparator) nahe an einer fehlenden Exposition liegt, soll für Effektgrößen von >2 bis 5 (oder 0,2 bis <0,5) ein Heraufstufen von +1 und für Effektgrößen von >5 (oder <0,2) ein Heraufstufen von +2 vorgenommen werden.

Bei kontinuierlichen Expositionen ist entsprechend fachlicher Expertise zu prüfen, wie die Wirkung mit der Exposition zunimmt und ob die zu erwartende Exposition bei einem Effektschätzer von >2 oder >5 einer plausiblen Exposition am Arbeitsplatz über eine 8-Stunden-Schicht und rund 40 Arbeitsjahre entspricht (Domäne „Effektgröße“).

Begründung

Die Domäne „Effektgröße“ ist auch die erste der von Hill (1965) beschriebenen ergänzenden Kriterien zur Beurteilung der Kausalität von Zusammenhängen zwischen einer Exposition und einem Endpunkt. Hill nennt als Beispiele die in der Vergangenheit beobachtete auffallend hohe Mortalität von Skrotalkarzinomen bei Schornsteinfegern und Lungenkrebs bei Rauchern. Relative Risiken, die einem großen Effekt entsprechen, sprechen tendenziell gegen Confounding als alleinige Erklärung der Studienergebnisse (Guyatt et al. 2011e).

Laut GRADE können relative Risiken von über 2 bis 5 (oder von 0,2 bis <0,5) aus methodisch einwandfreien Beobachtungsstudien nicht bloß durch Bias erklärt werden (Guyatt et al. 2011e). Noch größere Effekte eines relativen Risikos über 5 (oder unter 0,2) sind wahrscheinlich nicht das Ergebnis von Confounding (Guyatt et al. 2011e). Aus diesem Grund empfehlt GRADE eine Aufwertung von +1 für Effektgrößen von über 2 bis 5 (oder 0,2 bis <0,5) und eine Aufwertung von +2 für Effektgrößen über 5 (oder <0,2).

Hulshof et al. (2019) erklären im Navigation Guide, dass diese Relative Risiko-Definitionen von GRADE ungeeignet sind für Fragen zu den gesundheitlichen Folgen von Umweltexpositionen, weil diese Effektgrößen eine Funktion der Vergleichsgruppe (Komparator) sind und sich nicht für kontinuierliche Variablen eignen. Die Vergleichsgruppe in umwelt- (und arbeits-)epidemiologischen Studien ist nicht immer vollständig ohne Exposition (counterfactual). Außerdem sind umwelt- und berufsbedingte Expositionen oft kontinuierlich zu betrachten, und relative Risiken von über 2 sind bei ausreichend hohen Expositionen einfach zu erreichen. Für diese Fälle schlagen Hulshof et al. (2019) eine Abweichung von der GRADE-Methode vor und empfehlen ein "auf Expertise basiertes Rating" für starke Effekte, ohne eine spezifische Definition für einen starken Effekt zu formulieren.

Empfehlung 5.9

Liegt ein Hinweis auf das Vorhandensein einer monotonen oder epidemiologisch plausiblen Expositions-Wirkungs-Beziehung vor, soll eine Heraufstufung der Vertrauenswürdigkeit der Evidenz auf Basis der Konsistenz und der Stärke der Expositions-Wirkungs-Beziehung vorgenommen werden (Domäne „Expositions-Wirkungs-Beziehung“).

Begründung

Das Kriterium der „Expositions-Wirkungs-Beziehung“ wird von Hill (1965) für die Bewertung der Kausalität von Zusammenhängen erwähnt, aber von ihm als "biologischer Gradient" bezeichnet. Ein Dosis-Wirkungs-Gradient bietet nach Hill (1965) eine

zusätzliche Dimension, mit der eine mögliche kausale Beziehung untersucht werden kann. Gibt es einen eindeutigen Trend, dass sich das Risiko mit zunehmender Exposition verändert, deutet dies auf eine Expositions-Wirkungs-Beziehung hin.

Eine Expositions-Wirkungs-Beziehung lässt sich am besten in einer Studie überprüfen, in der mehrere Expositionsstufen untersucht wurden. Es können auch Effektschätzer aus mehreren Studien in Betracht gezogen werden, die verschiedenen Expositionsniveaus untersucht haben. Eine Expositions-Wirkungs-Beziehung ist am einfachsten zu erkennen, wenn das Risiko in einem linearen Verhältnis zur Exposition stetig ansteigt oder absinkt. Kompliziertere Beziehungen, wie beispielsweise eine U-Kurve oder eine logistische Kurve, können ebenfalls bestehen, erfordern aber möglicherweise mehr Daten, um sie sicher zu erkennen.

Der Navigation Guide gibt keine klare Empfehlung, wann man die Domäne „Expositions-Wirkungs-Beziehung“ um eine (+1) oder zwei (+2) Stufen höher bewerten sollte. Es wird empfohlen zu prüfen, ob die Expositions-Wirkungs-Beziehungen in einer oder mehreren Studien konsistent waren und/oder ob die Expositions-Wirkungs-Beziehung in unterschiedlichen Studien beobachtet wurde (Hulshof et al. 2019). Es liegt im Ermessen der Reviewautor:innen, die Konsistenz und die Stärke der Expositions-Wirkungs-Beziehungen zu prüfen und zu entscheiden, ob dies die Vertrauenswürdigkeit der Evidenz um eine oder zwei Stufen erhöht.

Empfehlung 5.10

Die Vertrauenswürdigkeit der Evidenz soll heraufgestuft werden, wenn trotz einer verbleibenden Verzerrung oder eines verbleibenden Confoundings, welches zu einer Unterschätzung des Effektes führen würde, ein Effekt beobachtet wurde (Domäne „Residual Confounding“).

Begründung

Während die Auswirkungen bekannter und messbarer Confounder (Störfaktoren) und die angemessene Berücksichtigung dieser Faktoren in den einzelnen Studien bei der Bewertung des Risikos von Bias berücksichtigt werden (Empfehlung 5.3), ist eine vollständige Korrektur für Confounding nicht immer möglich. Störfaktoren, die nicht messbar sind, nicht genau gemessen oder nicht vollständig berücksichtigt werden können, können ein Rest-Confounding verursachen. In einigen Fällen führt dieses Rest-Confounding zu einer Unterschätzung des Effektes (Guyatt et al. 2011e).

Obwohl diese Domäne als „residuales Confounding“ bezeichnet wird, ist hiermit nicht nur residuales Confounding, sondern jede verbleibende Verzerrung gemeint, die eine Unterschätzung des Effekts verursachen könnte (Guyatt et al. 2011e). Wenn trotz dieser

verbleibenden Verzerrung oder des verbleibenden Confoundings, welches zu einer Unterschätzung des Effekts führen würde, ein Effekt beobachtet wurde, kann die Vertrauenswürdigkeit der Evidenz erhöht werden. Guyatt et al. (2011e) nennen als Beispiel, wenn Patient:innen mit schweren Erkrankungen eher die experimentelle Intervention oder die Exposition erhalten, aber dennoch eine bessere Prognose haben.

Es liegt im Ermessen der Reviewautor:innen, wann die Domäne „residuales Confounding“ um eine (+1) oder zwei (+2) Stufen höher bewertet wird, da der Navigation Guide diesbezüglich keine klare Empfehlung gibt (Hulshof et al. 2019).[15]

[15] Im Bereich der Arbeitsepidemiologie kann die Wirkung von berufsbedingtem Lärm auf die Schwerhörigkeit als theoretisches Beispiel dienen. Die Verwendung von Gehörschutz sollte die Arbeitnehmer vor den Auswirkungen von berufsbedingtem Lärm schützen und werden wahrscheinlich bei höherer Lärmexposition konsistenter getragen. Wenn nur wenige Studien für die Verwendung von Gehörschutz adjustieren, dürfte dieser Mangel an Adjustierung zu einer Unterschätzung der Auswirkungen von berufsbedingter Lärmbelastung führen.

Referenzen

Armstrong BG (1998). "Effect of measurement error on epidemiological studies of environmental and occupational exposures." Occup Environ Med **55**(10): 651–656.

Arroyave WD, Mehta SS, Guha N, Schwingl P, Taylor KW, Glenn B, Radke EG, Vilahur N, Carreón T, Nachman RM, Lunn RM (2021). "Challenges and recommendations on the conduct of systematic reviews of observational epidemiologic studies in environmental and occupational health." J Expo Sci Environ Epidemiol **31**(1): 21–30.

Assink M, Wibbelink CJM. (2016). Fitting Three-Level Models in Meta-Analysis. Research Synthesis Methods **7**(3): 232–240.

Borenstein M, Hedges LV, Higgins, JPT, Rothstein HR (2009). Chapter 7: Converting Among Effect Sizes. Introduction to Meta-Analysis, Wiley.

Boutron I, Page M, Higgins J, Altman D, Lundh A, Hróbjartsson A (2023). Chapter 7: Considering bias and conflicts of interest among the included studies. Cochrane Handbook for Systematic Reviews of Interventions J Higgins, J Thomas, J Chandler, M Cumpston, T Li, M Page und V Welch, Cochrane.

Bramer WM, Rethlefsen ML, Kleijnen J, Franco OH (2017). "Optimal database combinations for literature searches in systematic reviews: a prospective exploratory study." Syst Rev **6**(1): 245.

Buscemi N, Hartling L, Vandermeer B, Tjosvold L, Klassen TP (2006). "Single data extraction generated more errors than double data extraction in systematic reviews." J Clin Epidemiol **59**(7): 697–703.

Checkoway H, Pearce N, Kriebel D (2004). Research Methods in Occupational Epidemiology, Oxford University Press.

Cheung MWL (2014). “Modeling Dependent Effect Sizes with Three-Level Meta-Analyses: A Structural Equation Modeling Approach.” Psychological Methods **19**(2): 211–229.

Christensen R, Schünemann HJ, Guyatt GH (2021). "GRADE guidelines 32: GRADE offers guidance on choosing targets of GRADE certainty of evidence ratings." J Clin Epidemiol **137**: 163–175.

Cornfield J (1951). "A Method of Estimating Comparative Rates from Clinical Data. Applications to Cancer of the Lung, Breast, and Cervix." JNCI: Journal of the National Cancer Institute **11**(6): 1269–1275.

Crippa A, Discacciati A, Bottai M, Spiegelman D, Orsini N (2019). "One-stage dose–response meta-analysis for aggregated data." Stat Methods Med Res **28**(5): 1579–1596.

Cummings P (2009). "The relative merits of risk ratios and odds ratios." Arch Pediatr Adolesc Med **163**(5): 438–445.

Deeks JJ, Higgins J, Altman DG (2022). Chapter 10: Analysing data and undertaking meta-analyses. Cochrane Handbook for Systematic Reviews of Interventions. Version 6.3. Higgins J und Thomas J.

Egger M, Smith GD, Schneider M, Minder C (1997). "Bias in meta-analysis detected by a simple, graphical test." BMJ **315**(7109): 629–634.

Goossen K, Rombey T, Kugler CM, De Santis KK, Pieper D (2021). ”Author queries via email text elicited high response and took less reviewer time than data forms - a randomised study within a review.” J Clin Epidemiol **135**: 1-9.

Greenland S, Longnecker MP (1992). "Methods for trend estimation from summarized dose-response data, with applications to meta-analysis." Am J Epidemiol **135**(11): 1301–1309.

Greenland S, Thomas DC (1982). “On the Need for the Rare Disease Assumption in Case-Control Studies.” Am J Epidemiol **116**(3): 547–553. doi:https://doi.org/10.1093/oxfordjournals.aje.a113439.

Guyatt G, Oxman AD, Akl EA, Kunz R, Vist G, Brozek J, Norris S, Falck-Ytter Y, Glasziou P, DeBeer H, Jaeschke R, Rind D, Meerpohl J, Dahm P, Schünemann HJ (2011a). "GRADE guidelines: 1. Introduction-GRADE evidence profiles and summary of findings tables." J Clin Epidemiol **64**(4): 383–394.

Guyatt GH, Oxman AD, Kunz R, Woodcock J, Brozek J, Helfand M, Alonso-Coello P, Falck-Ytter Y, Jaeschke R, Vist G, Akl EA, Post PN, Norris S, Meerpohl J, Shukla VK, Nasser M, Schünemann HJ (2011b). "GRADE guidelines: 8. Rating the quality of evidence--indirectness." J Clin Epidemiol **64**(12): 1303–1310.

Guyatt GH, Oxman AD, Kunz R, Woodcock J, Brozek J, Helfand M, Alonso-Coello P, Glasziou P, Jaeschke R, Akl EA, Norris S, Vist G, Dahm P, Shukla VK, Higgins J, Falck-Ytter Y, Schünemann HJ (2011c). "GRADE guidelines: 7. Rating the quality of evidence--inconsistency." J Clin Epidemiol **64**(12): 1294–1302.

Guyatt GH, Oxman AD, Montori V, Vist G, Kunz R, Brozek J, Alonso-Coello P, Djulbegovic B, Atkins D, Falck-Ytter Y, Williams, Jr. JW, Meerpohl J, Norris SL, Akl EA, Schünemann HJ (2011d). "GRADE guidelines: 5. Rating the quality of evidence--publication bias." J Clin Epidemiol **64**(12): 1277–1282.

Guyatt GH, Oxman AD, Sultan S, Glasziou P, Akl EA, Alonso-Coello P, Atkins D, Kunz R, Brozek J, Montori V, Jaeschke R, Rind D, Dahm P, Meerpohl J, Vist G, Berliner E, Norris

S, Falck-Ytter Y, Murad MH, Schünemann HJ (2011e). "GRADE guidelines: 9. Rating up the quality of evidence." J Clin Epidemiol **64**(12): 1311–1316.

Guyatt GH, Oxman AD, Vist G, Kunz R, Brozek J, Alonso-Coello P, Montori V, Akl EA, Djulbegovic B, Falck-Ytter Y, Norris SL, Williams, Jr. JW, Atkins D, Meerpohl J, Schünemann HJ (2011f). "GRADE guidelines: 4. Rating the quality of evidence – study limitations (risk of bias)." J Clin Epidemiol **64**(4): 407–415.

Guyatt GH, Oxman AD, Vist GE, Kunz R, Falck-Ytter Y, Alonso-Coello P, Schünemann HJ (2008). "GRADE: an emerging consensus on rating quality of evidence and strength of recommendations." BMJ **336**(7650): 924–926.

Harbord RM, Egger M, Sterne JA (2006). "A modified test for small-study effects in meta-analyses of controlled trials with binary endpoints." Stat Med **25**(20): 3443–3457.

Hedges LV, Tipton E, Johnson MC (2010). “Robust Variance Estimation in Meta-Regression with Dependent Effect Size Estimates.” Research Synthesis Methods **1**(1): 39–65.

Hempel S, Xenakis L, Danz M (2016). Systematic Reviews for Occupational Safety and Health Questions: Resources for Evidence Synthesis. Santa Monica, CA, RAND Corporation.

Higgins J, Morgan R, Rooney A, Taylor K, Thayer K, Silva R, Lemeris C, Akl A, Bateson TF, Berkman ND, Glenn BS, Hrobjartsson A, LaKind JS, McAleenan A, Meerpohl JJ, Nachman RM, Obbagy JE, O´Connor A, Radke EG, Savovic J, Schünemann HJ, Shea B, Tilling K, Verbeek J, Viswanathan M, Sterne JAC (2024a). A tool to assess risk of bias in non-randomized follow-up studies of exposure effects (ROBINS-E). Environ Int 186:108602Higgins J, Savovic J, Page M, Elbers R, Sterne J (2023b). Chapter 8: Assessing risk of bias in a randomized trial. Cochrane Handbook for Systematic Reviews of Interventions version 6.3 (updated February 2022). Higgins J, Thomas J, Chandler J, Cumpstan M, Li T, Page M und Welch V, Cochrane.

Higgins JP, Green S (2008). "Cochrane Handbook for Systematic Reviews of Interventions."

Higgins JPT, Eldridge S Tianjing L (2024). Chapter 23.3.4: How to include multiple groups from one study. Cochrane Handbook for Systematic Reviews of Interventions version 6.5 (updated August 2024). Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, Welch VA (editors). Cochrane, 2024. Zugriff 26.10.2024. www.training.cochrane.org/handbook.

Hill AB (1965). "The environment and disease: association or causation?" Proc R Soc Med **58**(5): 295-300.

Howard J, Piacentino J, MacMahon K, Schulte P (2017). "Using systematic review in occupational safety and health." Am J Ind Med **60**(11): 921–929.

Hulshof CTJ, Colosio C, Daams JG, Ivanov ID, Prakash KC, Kuijer PPFM, Leppink N, Mandic-Rajcevic S, Masci F, van der Molen HF, Neupane S, Nygård C-H, Oakman J, Pega F, Proper K, Prüss-Üstün AM, Ujita Y, Frings-Dresen MHW (2019). "WHO/ILO work-related burden of disease and injury: Protocol for systematic reviews of exposure to occupational ergonomic risk factors and of the effect of exposure to occupational ergonomic risk factors on osteoarthritis of hip or knee and selected other musculoskeletal diseases." Environ Int **125**: 554–566.

Hulshof CTJ, Pega F, Neupane S, Colosio C, Daams JG, Kc P, Kuijer P, Mandic-Rajcevic S, Masci F, van der Molen HF, Nygård CH, Oakman J, Proper KI, Frings-Dresen MHW (2021a). "The effect of occupational exposure to ergonomic risk factors on osteoarthritis of hip or knee and selected other musculoskeletal diseases: A systematic review and meta-analysis from the WHO/ILO Joint Estimates of the Work-related Burden of Disease and Injury." Environ Int **150**: 106349.

Hulshof CTJ, Pega F, Neupane S, van der Molen HF, Colosio C, Daams JG, Descatha A, Kc P, Kuijer P, Mandic-Rajcevic S, Masci F, Morgan RL, Nygård CH, Oakman J, Proper KI, Solovieva S, Frings-Dresen MHW (2021b). "The prevalence of occupational exposure to ergonomic risk factors: A systematic review and meta-analysis from the WHO/ILO Joint Estimates of the Work-related Burden of Disease and Injury." Environ Int **146**: 106157.

Hultcrantz M, Rind D, Akl EA, Treweek S, Mustafa RA, Iorio A, Alper BS, Meerpohl JJ, Murad MH, Ansari MT, Katikireddi SV, Ostlund P, Tranaeus S, Christensen R, Gartlehner G, Brozek J, Izcovich A, Schünemann H, Guyatt G (2017). "The GRADE Working Group clarifies the construct of certainty of evidence." J Clin Epidemiol **87**: 4-13.

Joanna Briggs Institute (2020). Joanna Briggs Institute (JBI) Appraisal Tools.

Johnson PI, Sutton P, Atchley DS, Koustas E, Lam J, Sen S, Robinson KA, Axelrad DA, Woodruff TJ (2014). "The Navigation Guide - evidence-based medicine meets environmental health: systematic review of human evidence for PFOA effects on fetal growth." Environ Health Perspect **122**(10): 1028–1039.

Kale A, Kay M, Hullman J (2019). Decision-Making Under Uncertainty in Research Synthesis: Designing for the Garden of Forking Paths. Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. Glasgow, Scotland, Association for Computing Machinery**:** Paper 202.

Konstantopoulos S. (2011). “Fixed Effects and Variance Components Estimation in Three-Level Meta-Analysis.” Research Synthesis Methods **2**(1): 61–76.

Kuijer P, Verbeek JH, Seidler A, Ellegast R, Hulshof CTJ, Frings-Dresen MHW, Van der Molen HF (2018). "Work-relatedness of lumbosacral radiculopathy syndrome: Review and dose-response meta-analysis." Neurology **91**(12): 558–564.

Lam J, Koustas E, Sutton P, Padula AM, Cabana MD, Vesterinen H, Griffiths C, Dickie M, Daniels N, Whitaker E, Woodruff TJ (2021). "Exposure to formaldehyde and asthma outcomes: A systematic review, meta-analysis, and economic assessment." PLoS One **16**(3): e0248258.

Lam J, Sutton P, Padula A, Cabana M, Koustas E, Vesterinen H, Whitaker E, Skalla L, Daniels N, Woodruff T (2016). Applying the Navigation Guide Systematic Review Methodology Case Study# 6: Association Between Formaldehyde Exposure and Asthma: A Systematic Review of the Evidence: Protocol. San Francisco, CA, University of California at San Francisco.

Langan D, Higgins JPT, Jackson D, Bowden J, Veroniki AA, Kontopantelis E, Viechtbauer W, Simmonds M (2019). "A comparison of heterogeneity variance estimators in simulated random-effects meta-analyses." Res Synth Methods **10**(1): 83–98.

Lash T, Rothman K (2021). Selection Bias and Generalizability. Modern Epidemiology T Lash, T VanderWeele, S Haneuse und K Rothman, Wolters Kluwer**:** pp. 315–331.

Lash T, VanderWeele T, Rothman KJ (2021). Measurement and Measurement Error. Modern Epidemiology. Lash T, VanderWeele T, Haneuse S und Rothman KJ. Wolters Kluwer: pp. 287-314.

Lash TL, Fox MP, MacLehose RF, Maldonado G, McCandless LC, Greenland S (2014). "Good practices for quantitative bias analysis." Int J Epidemiol **43**(6): 1969–1985.

Last J, Ed. (2000). A Dictionary of Epidemiology. Oxford, Oxford University Press.

Lefebvre C, Glanville J, Briscoe S, Featherstone R, Littlewood A, Marshall C, Metzendorf M-I, Noel-Storr A, Paynter R, Rader T, Thomas J, Wieland L (2022). Chapter 4: Searching for and selecting studies. Cochrane Handbook for Systematic Reviews of Interventions version 6.3 (updated February 2022). Higgins JP, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, Welch VA. Cochrane.

Li T, Higgins JP, Deeks JJ (2022). Chapter 5: Collecting data. Cochrane Handbook for Systematic Reviews of Interventions. Version 6.3.

Mattioli S, Gori D, Di Gregori V, Ricotta L, Baldasseroni A, Farioli A, Zanardi F, Galletti S, Colosio C, Curti S, Violante FS (2013). "PubMed search strings for the study of agricultural workers' diseases." Am J Ind Med **56**(12): 1473–1481.

Mattioli S, Zanardi F, Baldasseroni A, Schaafsma F, Cooke RM, Mancini G, Fierro M, Santangelo C, Farioli A, Fucksia S, Curti S, Violante FS, Verbeek J (2010). "Search strings for the study of putative occupational determinants of disease." Occup Environ Med **67**(7): 436-443.

McGowan J, Sampson M, Salzwedel DM, Cogo E, Foerster V, Lefebvre C (2016). "PRESS Peer Review of Electronic Search Strategies: 2015 Guideline Statement." J Clin Epidemiol **75**: 40–46.

McNutt L, Hafner J, Xue X (1999). Correcting the Odds Ratio in Cohort Studies of Common Outcomes. JAMA **282**(6): 529. doi:10-1001/pubs.JAMA-ISSN-0098-7484-282-6-jbk0811.

McNutt L, Wu C, Xue X, Hafner JP (2003). Estimating the Relative Risk in Cohort Studies and Clinical Trials of Common Outcomes. Am J Epidemiol **157**(10): 940–943. doi:https://doi.org/10.1093/aje/kwg074.

Meursinge Reynders R, Ladu L, Di Girolamo N (2019). "Contacting of authors modified crucial outcomes of systematic reviews but was poorly reported, not systematic, and produced conflicting results." J Clin Epidemiol **115**: 64-76.

Morgan RL, Thayer KA, Bero L, Bruce N, Falck-Ytter Y, Ghersi D, Guyatt G, Hooijmans C, Langendam M, Mandrioli D, Mustafa RA, Rehfuess EA, Rooney AA, Shea B, Silbergeld EK, Sutton P, Wolfe MS, Woodruff TJ, Verbeek JH, Holloway AC, Santesso N, Schünemann HJ (2016). "GRADE: Assessing the quality of evidence in environmental and occupational health." Environ Int **92-93**: 611–616.

Morgan RL, Whaley P, Thayer KA, Schünemann HJ (2018). “Identifying the PECO: A framework for formulating good questions to explore the association of environmental and other exposures with health outcomes.“ Environ Int **121**(Pt 1):1027-1031.

Muka T, Glisic M, Milic J, Verhoog S, Bohlius J, Bramer W, Chowdhury R, Franco OH (2020). "A 24-step guide on how to design, conduct, and successfully publish a systematic review and meta-analysis in medical research." Eur J Epidemiol **35**(1): 49–60.

Office of Health Assessment and Translation (2015). Handbook for Conducting a Literature-Based Health Assessment Using OHAT Approach for Systematic Review and Evidence Integration. Washington D.C., National Toxicology Program, US Department of Health and Human Services.

Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, Shamseer L, Tetzlaff JM, Akl EA, Brennan SE, Chou R, Glanville J, Grimshaw JM, Hrobjartsson A, Lalu MM, Li T, Loder EW, Mayo-Wilson E, McDonald S, McGuinness LA, Stewart LA, Thomas J, Tricco AC, Welch VA, Whiting P, Moher D (2021). "The PRISMA 2020 statement: an updated guideline for reporting systematic reviews." BMJ **372**: n71.

Pastor DA, Lazowski RA (2018). “On the Multilevel Nature of Meta-Analysis: A Tutorial, Comparison, and Critique.” Psychological Methods **23**(4), 583–599.

Peters JL, Sutton AJ, Jones DR, Abrams KR, Rushton L (2006). "Comparison of two methods to detect publication bias in meta-analysis." JAMA **295**(6): 676–680.

Pieper D, Rombey T (2022). “Where to prospectively register a systematic review.” Syst Rev **11**(8).

Relevo, R (2012). "Chapter 4: effective search strategies for systematic reviews of medical tests." J Gen Intern Med **27**(Suppl 1): S28–32.

Rethlefsen ML, Page MJ (2022). "PRISMA 2020 and PRISMA-S: common questions on tracking records and the flow diagram." J Med Libr Assoc **110**(2): 253–257.

Romero Starke K, Kofahl M, Freiberg A, Schubert M, Gross ML, Schmauder S, Hegewald J, Kampf D, Stranzinger J, Nienhaus A, Seidler A (2020). "The risk of cytomegalovirus infection in daycare workers: a systematic review and meta-analysis." Int Arch Occup Environ Health **93**(1): 11–28.

Rothman KJ, Greenland S, Lash TL (2008). Modern Epidemiology, Wolters Kluwer Health/Lippincott Williams & Wilkins Philadelphia.

Rücker G, Schwarzer G, Carpenter J (2008). "Arcsine test for publication bias in meta-analyses with binary outcomes." Stat Med **27**(5): 746–763.

Savitz DA, Wellenius GA, Trikalinos TA (2019). "The Problem With Mechanistic Risk of Bias Assessments in Evidence Synthesis of Observational Studies and a Practical Alternative: Assessing the Impact of Specific Sources of Potential Bias." Am J Epidemiol **188**(9): 1581–1585.

Schmidt L, Finnerty Mutlu AN, Elmore R, Olorisade BK, Thomas J, Higgins JPT (2021). "Data extraction methods for systematic review (semi)automation: Update of a living systematic review." F1000Res **10**: 401.

Schünemann HJ, Cuello C, Akl EA, Mustafa RA, Meerpohl JJ, Thayer K, Morgan RL, Gartlehner G, Kunz R, Katikireddi SV, Sterne J, Higgins JP, Guyatt G, Grade Working Group (2019). "GRADE guidelines: 18. How ROBINS-I and other tools to assess risk of bias in nonrandomized studies should be used to rate the certainty of a body of evidence." J Clin Epidemiol **111**: 105–114.

Schünemann HJ, Higgins J, Vist G, Glasziou P, Akl EA, Skoetz N, Guyatt GH (2022a). Chapter 14: Completing ‘Summary of findings’ tables and grading the certainty of the evidence. Cochrane Handbook for Systematic Reviews of Interventions. Version 6.3. Higgins J und Thomas J. Cochrane.

Schünemann HJ, Neumann I, Hultcrantz M, Brignardello-Petersen R, Zeng L, Murad MH, Izcovich A, Morgano GP, Baldeh T, Santesso N, Cuello CG, Mbuagbaw L, Guyatt G, Wiercioch W, Piggott T, De Beer H, Vinceti M, Mathioudakis AG, Mayer MG, Mustafa R, Filippini T, Iorio A, Nieuwlaat R, Marcucci M, Coello PA, Bonovas S, Piovani D, Tomlinson G, Akl EA, Grade Working Group (2022b). "GRADE guidance 35: update on rating imprecision for assessing contextualized certainty of evidence and making decisions." J Clin Epidemiol **150**: 225–242.

Seidler A, Nußbaumer-Streit B, Apfelbacher C, Zeeb H (2021). "[Rapid Reviews in the Time of COVID-19 - Experiences of the Competence Network Public Health COVID-19 and Proposal for a Standardized Procedure]." Gesundheitswesen **83**(3): 173–179.

Seidler A, Schubert M, Freiberg A, Drossler S, Hussenoeder FS, Conrad I, Riedel-Heller S, Starke KR (2022). "Psychosocial Occupational Exposures and Mental Illness." Dtsch Arztebl Int **119**(42): 709–715.

Shea BJ, Reeves BC, Wells G, Thuku M, Hamel C, Moran J, Moher D, Tugwell P, Welch V, Kristjansson E, Henry DA (2017). "AMSTAR 2: a critical appraisal tool for systematic reviews that include randomised or non-randomised studies of healthcare interventions, or both." BMJ **358**: j4008.

Stang A (2010). “Critical evaluation of the Newcastle-Ottawa scale for the assessment of the quality of nonrandomized studies in meta-analyses.” Eur J Epidemiol **25:** 603–605.

Steenland K, Schubauer-Berigan MK, Vermeulen R, Lunn RM, Straif K, Zahm S, Stewart P, Arroyave WD, Mehta SS, Pearce N (2020). "Risk of Bias Assessments and Evidence Syntheses for Observational Epidemiologic Studies of Environmental and Occupational Exposures: Strengths and Limitations." Environ Health Perspect **128**(9): 95002.

Sterne JA, Hernan MA, Reeves BC, Savovic J, Berkman ND, Viswanathan M, Henry D, Altman DG, Ansari MT, Boutron I, Carpenter JR, Chan AW, Churchill R, Deeks JJ, Hrobjartsson A, Kirkham J, Juni P, Loke YK, Pigott TD, Ramsay CR, Regidor D, Rothstein HR, Sandhu L, Santaguida PL, Schünemann HJ, Shea B, Shrier I, Tugwell P, Turner L, Valentine JC, Waddington H, Waters E, Wells GA, Whiting PF, Higgins JP (2016). "ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions." BMJ **355**: i4919.

Sterne JAC, Gavaghan D, Egger M (2000). "Publication and related bias in meta-analysis: Power of statistical tests and prevalence in the literature." Journal of Clinical Epidemiology **53**(11): 1119–1129.

Sterne JAC, Sutton AJ, Ioannidis JPA, Terrin N, Jones DR, Lau J, Carpenter J, Rücker G, Harbord RM, Schmid CH, Tetzlaff J, Deeks JJ, Peters J, Macaskill P, Schwarzer G, Duval S, Altman DG, Moher D, Higgins JPT (2011). "Recommendations for examining and

interpreting funnel plot asymmetry in meta-analyses of randomised controlled trials." BMJ **343**: d4002.

Thomas J, Kneale D, McKenzie J, Brennan S, Bhaumik S (2022). Chapter 2: Determining the scope of the review and the questions it will address. Cochrane Handbook for Systematic Reviews of Interventions version 6.3 (updated February 2022). J Higgins, J Thomas, J Chandler, M Cumpston, T Li, M Page und V Welch, Cochrane

Thompson SG, Higgins JP (2002). "How should meta-regression analyses be undertaken and interpreted?" Stat Med **21**(11): 1559–1573.

Valentine JC, Pigott TD, Rothstein HR (2010). "How Many Studies Do You Need?: A Primer on Statistical Power for Meta-Analysis." Journal of Educational and Behavioral Statistics **35**(2): 215–247.

VanderWeele TJ (2020). "Optimal approximate conversions of odds ratios and hazard ratios to risk ratios." Biometrics **76**(3): 746–752.

Verbeek J, Salmi J, Pasternack I, Jauhiainen M, Laamanen I, Schaafsma F, Hulshof C, van Dijk F (2005). "A search strategy for occupational health intervention studies." Occup Environ Med **62**(10): 682–687.

Wells GA, Shea B, O´Connell D, Peterson J, Welch V, Losos M (2009). "The Newcastle-Ottawa Scale (NOS) for assessing the quality of nonrandomized studies in meta-analyses." Zugriff 26.10.2024, https://www.ohri.ca/programs/clinical_epidemiology/oxford.asp

WHO (2020). Risk of bias assessment instrument for systematic reviews informing WHO global air quality guidelines. Risk of bias assessment instrument for systematic reviews informing WHO Global Air Quality Guidelines, World Health Organization (WHO).

Woodruff TJ, Sutton P (2014). "The Navigation Guide systematic review methodology: a rigorous and transparent method for translating environmental health science into better health outcomes." Environ Health Perspect **122**(10): 1007–1014.

Yland JJ, Wesselink AK, Lash TL, Fox MP (2022). "Misconceptions About the Direction of Bias From Nondifferential Misclassification." Am J Epidemiol **191**(8): 1485–1495.

Young T, Hopewell S (2011). "Methods for obtaining unpublished data." Cochrane Database Syst Rev **2011**(11): MR000027.

Zeng L, Brignardello-Petersen R, Hultcrantz M, Siemieniuk RAC, Santesso N, Traversy G, Izcovich A, Sadeghirad B, Alexander PE, Devji T, Rochwerg B, Murad MH, Morgan R, Christensen R, Schünemann HJ, Guyatt GH (2021). GRADE guidelines 32: GRADE offers

guidance on choosing targets of GRADE certainty of evidence ratings. J Clin Epidemiol **137**: 163-175.

Zeng L, Brignardello-Petersen R, Hultcrantz M, Mustafa RA, Murad MH, Iorio A, Traversy G, Akl EA, Mayer M, Schünemann HJ, Guyatt GH (2022). GRADE Guidance 34: update on rating imprecision using a minimally contextualized approach. J Clin Epidemiol **150**: 216-224. doi: 10.1016/j.jclinepi.2022.07.014. Epub 2022 Aug 4. PMID: 35934265.

Zhang J, Yu KF (1998). What's the Relative Risk? A Method of Correcting the Odds Ratio in Cohort Studies of Common Outcomes. JAMA **280**(19): 1690–1691. doi:10.1001/jama.280.19.1690.

Anhang A

Template data extraction form:

General	**First Author:** *(name of first author)* **Year:** *(year of publication)* **Extracted by:** *(initials of person that did the extraction of the study)* (For cases when data extraction is checked by another person: **Checked by** initials of this reviewer should also be entered)
Study	**Study name:** *(name of the study; not the name of the publication)* **Country & geographic location:** *(e.g., state, federal state, …)* **Study design:** *(e.g., cross-sectional, cohort, case-cohort, case-control, experimental study, lab study)* **Time of Study:** *(month, year(s); time range)* # **Waves:** (*number of study waves*) **Follow-up:** *(e.g., mean, range, minimum, maximum)*
Population	***(If necessary, differentiation by study groups)*** **Recruitment method:** (*e.g., census-based; population-based, i.e. individuals taken from the general population that share common characteristics; snowball technique; convenience sampling; judgment sampling; quota sampling)* **Eligibility criteria:** *(sample population, setting, inclusion/ exclusion criteria)* **Job characteristics:** *(description of job and labour activities; full-time/ part-time)*

	Matching criteria: *(only case-cohort, case control designsonly)* # **invited:** (*number of individuals invited for participating in the study*) # **baseline:** *(all individuals participating at baseline)* **Response:** *(number of invited/ number participated at baseline in %)* # **follow-up:** *(all individuals participating at follow-up)* **Loss-to-follow-up:** *(e.g., number of participated at baseline / number participated at follow-up X in %)* **Age at baseline:** *(e.g., mean SD, median, range)* **F : M ratio at baseline:** *(absolute values)*
Exposure	*Report missing data.* **Description of exposure:** (*parameters, units)* **Study groups:** *(exposure groups, reference groups)* **Methods used for measuring exposure:** *(detailed description of how exposure was measured, validity of measurements used, standardisation)* **Time of measurement(s):** *(time when exposure measurement(s) were done)*
Outcome	*Report missing data.* **Outcome name:** **Outcome definition and assessment:** *(detailed description of how outcome was measured, validity of measurements used, standardisation)*

	Time of measurement(s): *(time when outcome measurement(s) were done)*
Results	*Description as stated in the paper. Report if results are missing, or if only significant results are reported. Calculation performed by the review team should be noted. Unpublished results obtained by personal communication should be noted.* **Statistical methods used:** (*Models used; adjustment sets*. *Report the confounders considered in the analysis.)* **Unadjusted/ adjusted estimates with precision (e.g., 95% confidence interval) for each outcome:** *(if available, report per exposure category: total number of cases and controls (or exposed + unexposed), number of cases (or number exposed, depending on study design))*
Comments	**Funding:** *(source)* **Conflict of interest stated:**