Systematic Reviews for Occupational Safety and Health Questions

Resources for Evidence Synthesis

Susanne Hempel, Lea Xenakis, Marjorie Danz

For more information on this publication, visit www.rand.org/t/RR1463

Published by the RAND Corporation, Santa Monica, Calif.

Preface

The RAND Corporation's Justice, Infrastructure, and Environment research division and Infrastructure Resilience and Environmental Policy program compiled a collection of resources and considerations for the conduct of systematic reviews in occupational safety and health. This report was commissioned by the National Institute for Occupational Safety and Health (NIOSH). Systematic review methods, although already used within NIOSH, would benefit from clear delineation and standardization. In addition, the occupational safety and health community would benefit from NIOSH leadership in the development and application of systematic review methods. This report collates key aspects to be considered in systematic reviews addressing occupational safety and health questions and provides resources for conducting systematic reviews tailored to the needs of the occupational safety and health community.

This report was developed by systematic review experts at RAND and supported by expertise in fields relevant to occupational safety and health, such as patient safety and public health. RAND hosts an Evidence-based Practice Center (EPC). Under the EPC Program of the Agency for Healthcare Research and Quality, five-year contracts are awarded to institutions in the United States and Canada to serve as EPCs. The EPCs review scientific literature on a wide spectrum of clinical and health services topics to produce various types of evidence reports and conduct research on the methodology of evidence synthesis.

RAND Infrastructure Resilience and Environmental Policy

The research reported here was conducted in the RAND Infrastructure Resilience and Environmental Policy program, which performs analyses on urbanization and other stresses. This includes research on infrastructure development; infrastructure financing; energy policy; urban planning and the role of public-private partnerships; transportation policy; climate response, mitigation, and adaptation; environmental sustainability; and water resource management and coastal protection. Program research is supported by government agencies, foundations, and the private sector. This program is part of RAND Justice, Infrastructure, and Environment, a division of the RAND Corporation dedicated to improving policy- and decisionmaking in a wide range of policy domains, including civil and criminal justice, infrastructure protection and homeland security, transportation and energy policy, and environmental and natural resource policy.

Questions or comments about this report should be sent to the project leader, Susanne Hempel (susanne_hempel@rand.org). For more information about RAND Infrastructure Resilience and Environmental Policy, see www.rand.org/jie/irep or contact the director at irep@rand.org.

Contents

Figures and Tables

Figures

Tables

Summary

Evolving scientific standards and public policy increasingly require systematic reviews of the evidence base, with clear documentation and a transparent approach, in particular when developing guidance or recommendation documents. Government agencies targeting occupational safety and health questions, such as the National Institute for Occupational Safety and Health (NIOSH), need to be able to document how evidence was collected and evaluated, clearly describe the basis of recommendations, and document the underlying review methodology.

Systematic review is a research methodology that aims to summarize the existing evidence to answer a research or policy question with a transparent, reliable, and valid approach. Today's literature reviews are increasingly challenging and need a structured approach because of the sheer volume of available literature. Systematic reviews follow standardized, well-documented, and replicable steps. This includes a number of approaches to reduce literature-reviewer errors and bias. Systematic reviews also use standard documentation elements to enable a transparent, structured, and comprehensive overview of the available literature.

Occupational safety and health is an extensive multidisciplinary field and may include questions relevant to physiology, toxicology, epidemiology, industrial hygiene, and law. Systematic reviews supporting guidance and recommendations targeting occupational safety and health questions cover a wide range of content areas. Reviews typically consider diverse lines of evidence such as findings from mechanistic, human, and nonhuman research studies, and draw on a variety of sources. Existing guidance for systematic review, such as the *Cochrane Handbook for Systematic Reviews of Interventions*, is usually aimed at narrower types of bodies of evidence. While methods for systematic reviews in occupational safety and health can draw on the principles of existing systematic review approaches, these have to fit unique requirements, such as evaluating bodies of evidence that include few randomized controlled trials. In addition, the integration of different lines of evidence poses a particular challenge, typically exceeding the scope of traditional evidence-based medicine systematic review approaches. Furthermore, systematic reviews that support guidance and recommendation documents need to consider mechanisms to ensure transparency of the evidence synthesis as well as the recommendations supported by the evidence.

This report outlines the steps undertaken in a systematic review. It provides practical guidance to execute a systematic review as well as considerations and available resources specific to conducting systematic review for occupational safety and health questions. It draws both on our practical experiences as systematic reviewers as well as on key existing guidance documents for systematic reviews in health care. This was supplemented by a "snowball" search for resources specific to occupational safety and health systematic reviews. We performed a

review of the literature for publications on weight-of-evidence approaches to identify models and examples updating published resource collections. Key informants representing producers and consumers of systematic reviews provided input over the course of the project. NIOSH staff provided a draft systematic review framework for occupational safety and health and a large number of published resources and input on earlier drafts of the report to ensure relevance and applicability to occupational safety and health questions.

The chapters in this report outline the general steps to be undertaken to plan and execute a systematic review and provide practical resources:

- Chapter One: "Introduction and Report Methods" provides context on why systematic reviews are important for occupational safety and health recommendations, documents how resources were selected for this report, outlines the report structure, and describes the objective of the report.
- Chapter Two: "Systematic Review Step 1: Define the Question" addresses formulating systematic review questions and the scope of the review; introduces scoping reviews; and provides some guidance with regard to establishing a review team, key informants, technical expert panels, and stakeholders.
- Chapter Three: "Systematic Review Step 2: Create a Protocol" introduces the concept of systematic review protocols. It outlines the protocol elements, introduces analytic frameworks, highlights different approaches relevant to the methodological conduct of systematic reviews, and provides information on the function of systematic review protocols. This chapter discusses the scalability of systematic reviews; introduces other evidence synthesis and review products, such as rapid reviews; and highlights the importance of transparency of review methods.
- Chapter Four: "Systematic Review Step 3: Conduct a Literature Search and Screen for Inclusion" covers data sources, including electronic databases, grey literature, publication sets, and other sources. The chapter outlines the development of search strategies and the use of citation management software. It covers formulating eligibility criteria for the review, outlines the inclusion screening process, and introduces a literature flow diagram.
- Chapter Five: "Systematic Review Step 4: Document and Assess Included Studies" addresses the approach to data extraction in systematic reviews and available resources for systematic reviewers. The chapter introduces evidence tables as a standard documentation element of systematic reviews, provides an overview of critical appraisal addressing internal and external validity, and discusses data access requirements.
- Chapter Six: "Systematic Review Step 5: Evaluate and Interpret the Body of Evidence" is divided into sections on synthesizing evidence, grading evidence, and integrating evidence. Evidence synthesis addresses summarizing research studies across studies, introduces meta-analysis and forest plots, and promotes the use of tools such as summary of findings tables. Evidence grading introduces criteria to evaluate a body of evidence. In addition to the standard Grading of Recommendations Assessment, Development and Evaluation (GRADE) approach, several adaptations relevant to occupational safety and health are presented. The evidence-integration section is dedicated to combining evidence across lines of evidence. In addition, this chapter highlights the importance of transparency in the evaluation and interpretation of evidence.

- Chapter Seven: "Draw Conclusions and Develop Recommendations" describes aspects relevant to the conclusions and recommendations following the systematic review results. The chapter provides an overview of reporting standards for systematic reviews and guidance and recommendation documents. It describes the process of developing recommendations, discusses grading the strength of recommendations, and introduces translational products.
- Chapter Eight: "Discussion and Outlook" briefly summarizes the steps involved in conducting a systematic review that provided the structure for this report and discusses the resources identified for this report. It stresses the need for updating systematic review guidance given the evolving methodology. Finally, the chapter introduces the need for a mechanism to look for signals that indicate the need for updating of completed systematic reviews and issued recommendations.

The report collates resources tailored to systematic reviews in occupational safety and health to provide an overview of general steps in systematic reviews, highlight key considerations, and identify practical resources to support occupational safety and health evidence synthesis. Each chapter describes methods and resources in detail. All cited literature can be found in the References section, and the links to online-only resources are listed in Table S.1.

Table S.1. Online Resources

Resource	Link
Chapter One: "Introduction and Report Methods"	
Cochrane Handbook for Systematic Reviews of Interventions	http://handbook.cochrane.org
Grading of Recommendations Assessment, Development and Evaluation (GRADE) Working Group	http://www.gradeworkinggroup.org/
National Toxicology Program Office of Health Assessment and Translation (OHAT) systematic review framework	https://ntp.niehs.nih.gov/pubhealth/hat/noms/index-2.html
Guide to Community Preventive Services: Systematic Review Methods	http://www.thecommunityguide.org/about/methods.html
Chapter Two: "Systematic Review Step 1: Define the Question"	
Database of Abstracts of Reviews of Effects (DARE)	http://www.crd.york.ac.uk/CRDWeb/
PubMed Health database, U.S. National Library of Medicine	http://www.ncbi.nlm.nih.gov/pubmedhealth
Occupational Safety and Health Review Group of the Cochrane Collaboration review database	http://work.cochrane.org/cochrane-reviews-about-occupational-safety-and-health
Partnership for European Research in Occupational Safety and Health clearinghouse of systematic reviews	http://www.perosh.eu
PubMed database collection for biomedical literature	http://www.ncbi.nlm.nih.gov/pubmed
NIOSH publications database	https://www2a.cdc.gov/nioshtic-2/advsearch2.asp
International Labour Organization database	http://www.ilo.org/dyn/cisdoc/cismain.home
American College of Occupational and Environmental Medicine Practice Guidelines	http://www.acoem.org/PracticeGuidelines.aspx
European Agency for Safety and Health at Work database	https://osha.europa.eu
Web of Science database	https://apps.webofknowledge.com

National Institute for Health and Care Excellence, Appendix B: Methodology Checklist: Systematic Reviews and Meta-Analyses	http://publications.nice.org.uk/the-guidelines-manual-appendices-bi-pmg6b
U.S. Department of Health and Human Services Paperwork Reduction Act notice	http://www.hhs.gov/ocio/policy/collection/infocollectfaq.html
Canadian Institute for Work and Health web resources	http://www.iwh.on.ca
Chapter Three: "Systematic Review Step 2: Create a Protocol"	
PRISMA-P, reporting guidelines for systematic review protocols	http://www.prisma-statement.org/Extensions/Protocols.aspx
PROSPERO, registry for systematic review protocols	http://www.crd.york.ac.uk/PROSPERO
Campbell Collaboration Library of Systematic Reviews	http://www.campbellcollaboration.org/lib
Systematic Reviews, journal publishing systematic reviews, protocols, and methods for evidence synthesis	http://www.systematicreviewsjournal.com/
U.S. Preventive Services Task Force (USPSTF) analytic frameworks	http://www.uspreventiveservicestaskforce.org/Page/Name/methods-and-processes
USPSTF systematic review examples	http://www.uspreventiveservicestaskforce.org/BrowseRec/Search
Guide to Community Preventive Services: Systematic Review Resources	http://www.thecommunityguide.org/
Guide to Community Preventive Services, conceptual framework example	http://www.thecommunityguide.org/healthequity/education/supportingmaterials/LM-Health-Equity-Education.pdf
Guide to Community Preventive Services analytic framework example	http://www.thecommunityguide.org/violence/supportingmaterials/AFViolence-Trauma.pdf
American Psychological Association PsycINFO database of behavioral and social science research	http://www.apa.org/pubs/databases/psycinfo/index.aspx
Agency for Healthcare Research and Quality (AHRQ), evidence-based reports	http://www.ahrq.gov/research/findings/evidence-based-reports/search.html
Abstrackr, software for automated citation processing	http://abstrackr.cebm.brown.edu
Weka data mining algorithms	http://www.cs.waikato.ac.nz/ml/weka/index.html
USPSTF procedure manual	http://www.uspreventiveservicestaskforce.org/Home/GetFile/6/7/procedure-manual_2016/pdf
HLWIKI International	http://hlwiki.slais.ubc.ca/index.php/Rapid_reviews
Chapter Four: "Systematic Review Step 3: Conduct a Literature Search and Screen for Inclusion"	
GreenFILE database, environmental literature	https://www.ebscohost.com/academic/greenfile
TOXLINE database, effects of drugs and chemicals	http://toxnet.nlm.nih.gov/cgi-bin/sis/htmlgen?TOXLINE.htm
U.S. Environmental Protection Agency (EPA) risk assessment database collection	https://www.epa.gov/risk/risk-tools-and-databases
American Conference of Governmental Industrial Hygienists operations manuals	http://www.acgih.org/tlv-bei-guidelines/policies-procedures-presentations/tlv-bei-committee-operations-manuals
Grey Literature Report database	http://www.greylit.org/
WorldCat, library catalog, grey literature collection	https://www.worldcat.org/
EPA ChemView and Chemical Data Access Tool, health and safety data and regulatory actions	https://java.epa.gov/chemview https://java.epa.gov/oppt_chemical_search
U.S. Food and Drug Administration regulatory data	https://www.accessdata.fda.gov/scripts/cder/drugsatfda/
American Conference of Governmental Industrial Hygienists threshold limit value development	http://www.acgih.org/tlv-bei-guidelines/policies-procedures-presentations/tlv-bei-development-process

Cochrane Center Register of Controlled Trials (CENTRAL)	http://www.cochranelibrary.com/about/central-landing-page.html
ClinicalTrials.gov, U.S. registry of clinical trials	https://clinicaltrials.gov/
World Health Organization research database of trials	http://apps.who.int/trialsearch/
Ovid MEDLINE database	http://www.ovid.com/site/catalog/databases/901.jsp
Google Translate, online translation tool	https://translate.google.com
Chapter Five: "Systematic Review Step 4: Document and Assess Included Studies"	
Covidence, systematic review software	www.covidence.org
ICF International, Dragon, online tool for systematic reviews to support risk assessments	http://www.icfi.com/insights/products-and-tools/dragon-online-tool-systematic-review
Guide to Community Preventive Service, guidance for data extraction	http://www.thecommunityguide.org/methods/abstractionform.pdf
AHRQ Slide Training Module for data extraction	http://www.effectivehealthcare.ahrq.gov/tools-and-resources/slide-library/#slidetrainingmodules
Newcastle-Ottawa Scale for observational studies	http://www.ohri.ca/programs/clinical_epidemiology/oxford.asp
Enhancing the QUAlity and Transparency Of Health Research (EQUATOR) network of reporting guidelines	http://www.equator-network.org/reporting-guidelines/
Consolidated Standards of Reporting Trials (CONSORT), reporting guidelines for trials	http://www.consort-statement.org/
STrengthening the Reporting of OBservational Studies in Epidemiology (STROBE), reporting guidelines for observational studies	http://www.strobe-statement.org/
Comparative Effectiveness Reviews (CER) Methods Guide, e.g., critical appraisal resources	http://effectivehealthcare.ahrq.gov/index.cfm/search-for-guides-reviews-and-reports/?pageaction=displayproduct&productid=318
508 compliance resources	http://www.508checker.com/what-is-508-compliance
Navigation Guide, detailed documentation of a systematic review	http://ehp.niehs.nih.gov/wp-content/uploads/122/10/ehp.1307893.s001.508.pdf
NIOSH Data and Statistics Gateway	http://www.cdc.gov/niosh/data
AHRQ Systematic Review Data Repository	http://srdr.ahrq.gov
EPA data repository databases	https://www.epa.gov/chemical-research/distributed-structure-searchable-toxicity-dsstox-database, https://www.epa.gov/chemical-research/toxicity-forecaster-toxcasttm-data, https://hero.epa.gov/hero/
Chapter Six: "Systematic Review Step 5: Evaluate and Interpret the Body of Evidence"	
Handbook of the Cochrane Screening and Diagnostic Tests Methods Group	http://dta.cochrane.org/handbook-dta-reviews
Cochrane Review Manager (RevMan) software for meta-analysis	http://tech.cochrane.org/revman
OpenMeta[Analyst] resource platform	http://www.cebm.brown.edu/openmeta/
Open source meta-analysis R packages	https://cran.r-project.org/web/views/MetaAnalysis.html
Cochrane GRADEpro software supporting summary of findings tables	http://tech.cochrane.org/revman/gradepro
Navigation Guide Protocol for Rating the Quality and Strength of Human and Non-Human Evidence	http://prhe.ucsf.edu/prhe/pdfs/Instructions%20to%20Authors%20for%20GRADING%20QUALITY%20OF%20EVIDENCE.pdf
AHRQ National Guideline Clearinghouse	https://www.guideline.gov
Chapter Seven: "Draw Conclusions and Develop Recommendations"	
AHRQ Evidence-Based Practice Center report collection	http://www.ahrq.gov/research/findings/evidence-based-reports/search.html
Federal Register Request for Information	https://www.federalregister.gov/articles/search?conditions%5Bterm%5D=Request+for+information&commit=Go
AHRQ Effective Health Care Program portal for public commenting	https://effectivehealthcare.ahrq.gov/index.cfm/research-available-for-comment/

Portal for public commenting National Toxicology Program	https://ntp.niehs.nih.gov/help/contactus/input/index.cfm
Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) reporting guideline for systematic reviews	http://www.prisma-statement.org/
National Health and Medical Research Council of Australia, clinical practice guideline form to structure evidence	http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3053308/bin/1471-2288-11-23-S1.DOC
Centers for Disease Control and Prevention (CDC), trainings and modules	http://www.cdc.gov/od/science/quality/guidelines-and-recommendations/index.htm
Software to evaluate public health guidelines	http://www.openclinical.org/dld_gem.html, http://nutmeg.med.yale.edu/egliahome.php http://gem.med.yale.edu/BRIDGE-Wiz/
Portal to AHRQ Effective Health Care Program consumer summaries	https://effectivehealthcare.ahrq.gov/index.cfm/search-for-guides-reviews-and-reports
Canadian Institute for Work and Health online summaries	http://www.iwh.on.ca/sharing-best-evidence
Links to U.K. Institution of Occupational Safety and Health toolkit	https://www.iosh.co.uk/Books-and-resources/Our-OH-toolkit.aspx
Canadian Institute for Work and Health, slide deck summarizing a systematic review	http://www.iwh.on.ca/plenaries/2015-may-05
Links to NIOSH videos	http://www.cdc.gov/niosh/docs/video/
Links to U.S. Department of Labor Occupational Safety and Health Administration (OSHA) interactive tool	https://www.osha.gov/hazfinder/

NOTE: All links accessed on March 25, 2016.

Acknowledgments

We would like to thank John Piacentino, John Decker, Kathleen MacMahon, John Howard, Paul Schulte, and other National Institute for Occupational Safety and Health staff; Andrea Furlan of the Institute of Work and Health; Justin Timbie of the RAND Corporation; and John Mendeloff of the RAND Corporation and University of Pittsburgh for comments supporting the development of this report. In addition, we would like to thank Anita Chandra, Sean Grant, and Paul Shekelle at the RAND Corporation, Cameron Mustard at the Institute for Work and Health, Holger Schünemann at McMaster University, Kristina Thayer at the Office of Health Assessment and Translation, and Jani Ruotsalainen and Jos Verbeek at the Finnish Institute of Occupational Health for content input and suggested resources. We also would like to acknowledge RAND Corporation staff Jody Larkin for literature-review services, Sydne Newberry for editorial assistance, and Patty Smith for administrative assistance.

Abbreviations

AHRQ	Agency for Healthcare Research and Quality
CDSR	*Cochrane Database of Systematic Reviews*
DARE	Database of Abstracts of Reviews of Effects
EPA	U.S. Environmental Protection Agency
EPC	Evidence-based Practice Center
GRADE	Grading of Recommendations Assessment, Development and Evaluation
IOM	Institute of Medicine
NIOSH	National Institute for Occupational Safety and Health
NLM	U.S. National Library of Medicine
NTP	National Toxicology Program
OHAT	Office of Health Assessment and Translation
PECO	Population, Exposure, Comparator, Outcome
PICO	Population, Intervention, Comparator, Outcome
PICOTSS	Population, Intervention, Comparator, Outcome, Timing, Setting, Study design
PRECEPT	Project on Framework for Rating Evidence in Public Health
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses
USPSTF	U.S. Preventive Services Task Force
WHO	World Health Organization

1. Introduction and Report Methods

Background

Occupational safety and health encompasses a broad spectrum of issues that affect the health and safety of individuals in the workplace. It is an extensive multidisciplinary field, including scientific areas such as medicine, physiology, toxicology, epidemiology, industrial hygiene, ergonomics, physics, chemistry, technology, economics, law, and other areas specific to various industries and activities (Alli, 2008). Developing guidance documents and formal recommendations in occupational safety and health, in the context of evolving scientific standards and increasing public policy demands, requires systematic review of the evidence base, with clear documentation and a transparent approach. Professional bodies and government agencies such as the National Institute for Occupational Safety and Health (NIOSH) need to be able to document the methodology underlying how evidence was collected and evaluated and clearly describe the basis of recommendations and guidance. This report outlines methods and identifies resources to perform systematic reviews to address occupational safety and health questions.

Systematic review is a research methodology that aims to identify, document, and summarize the existing evidence to answer a defined research or policy question with a transparent, reliable, and valid approach, often to support the development of guidelines or policies. Systematic reviews follow standardized, well-documented, and replicable steps. The objective is to reduce literature-reviewer errors and bias and to produce a transparent, structured, and comprehensive overview of the available literature. Organizations differ in their definitions and standards for systematic reviews (Eden et al., 2011, and Lau et al., 2013) and review authors need to balance resources and review methods. While it is important to clarify the specifications of a particular systematic review, several steps and principles are common to all systematic reviews. A number of relevant resources are available, including published guidance from established organizations describing their procedures (Higgins and Green, 2011, and Agency for Healthcare Research and Quality [AHRQ], 2014). However, while researchers conducting systematic reviews in occupational safety and health can draw on existing systematic review approaches, they will need to adapt this methodological guidance to the broad spectrum of disciplines that constitute occupational safety and health.

Guidelines, standards, guidance documents, and recommendations (referred to as *recommendations* in this report) of government agencies and professional bodies that pertain to occupational safety and health questions cover a wide range of content areas. In addition, those who set recommendations may need to consider diverse lines of evidence and a variety of sources, such as published research and "grey literature" (i.e., information or research that is not

published in scientific journals). Existing guidance for systematic reviews is usually developed for much-narrower questions and bodies of evidence than those in occupational safety and health. For example, the Cochrane Handbook for Systematic Reviews of Interventions addresses systematic reviews of clinical interventions to inform health care decisions that are predominantly based on randomized controlled trials assessing outcomes in patient populations. In contrast, summaries of available evidence in the area of occupational safety and health can include questions pertaining to multiple disciplines, all of which need to be summarized into a comprehensive report.

A particular feature of occupational safety and health recommendations is the need to integrate and interpret information across different types of evidence. This report refers to lines of evidence in the context of mechanistic, human, and nonhuman studies. Research disciplines often have their own unique sets of research standards, internal and external validity considerations, and standards for evidence evaluations. The Grading of Recommendations Assessment, Development and Evaluation (GRADE) Working Group has overhauled the approach to evaluating research evidence for clinical practice guidelines; however, researchers in other fields have repeatedly highlighted the need for criteria to evaluate the quality of evidence to be adapted to the research field and the type of research typically found in the content area of interest. For example, before conducting an updated evidence synthesis on patient safety, AHRQ commissioned the development of a system to rate the available evidence (Shekelle, Pronovost, and Wachter, 2010). The diverse types of available evidence that need to be integrated for systematic reviews of occupational safety and health issues likely exceed the scope of traditional evidence-based clinical systems.

Systematic reviews may be conducted for a variety of purposes, but those supporting recommendation and policy documents have unique requirements. In particular, they need to embrace mechanisms to ensure transparency of the evidence summary as well as the recommendations supported by the evidence. This report aims to facilitate systematic reviews in occupational safety and health that support formal recommendations.

Methods

In this section, we describe the methodology used to research and prepare this report.

Sources

The resources we present in this report were selected based both on our experiences as systematic reviewers and on existing general guidance for systematic reviews. These resources include, but are not limited to, the Evidence-based Practice Center (EPC) *Methods Guide for Effectiveness and Comparative Effectiveness Reviews* (AHRQ, 2014), the *Cochrane Handbook for Systematic Reviews* (Higgins and Green, 2011), *Systemic Reviews: CRD's [Centre for Reviews and Dissemination] Guidance for Undertaking Systematic Reviews in Health Care*

(2009) and the standards for systematic reviews published by the Institute of Medicine (Eden et al., 2011).

We supplemented these resources by a "snowball search" for systematic review guidance specific to occupational safety and health and public health. Additional guidance was sought on adaptations of systematic review methodology outside of medicine and health care, especially approaches that did not draw on clinical paradigms and trial methodology. This literature pool encompassed applications of systematic reviews for observational studies and safety research and included a systematic review framework published by the National Toxicology Program (NTP) Office of Health Assessment and Translation (OHAT), the Guide to Community Preventive Services approach to systematic reviews, a guide for environmental health science (Woodruff and Sutton, 2014), and a recently published framework for public health research (Harder et al., 2015).

The need to integrate different lines of evidence is an important consideration for occupational safety and health systematic reviews. Therefore, we performed a review of the literature in October 2015 for publications on "weight of evidence" and evidence-integration approaches. *Weight of evidence* refers to approaches to interpret and "weigh" different evidence elements, often in the context of risk assessment. The search was designed to identify pertinent models and examples, updating a key review on weight-of-evidence approaches (Weed, 2005). The search strategy used variations of the term *weight of evidence*, grading and integrating evidence for publications using the terms in the title ("ti") or in proximity in the abstract ("adj3"). The full strategy was as follows: [(weigh* adj3 evidence).ti. OR (hierarch* adj3 evidence).ti. OR (integrat* adj3 evidence).ti. OR (grade* adj3 evidence).ti. OR (grading adj3 evidence).ti. NOT "weight loss".ti; "body weight".ti; "birth weight".ti; "low weight".ti., integrative.ti., "high-grade".ti.] The search was performed in MEDLINE (Ovid) and identified 704 citations.

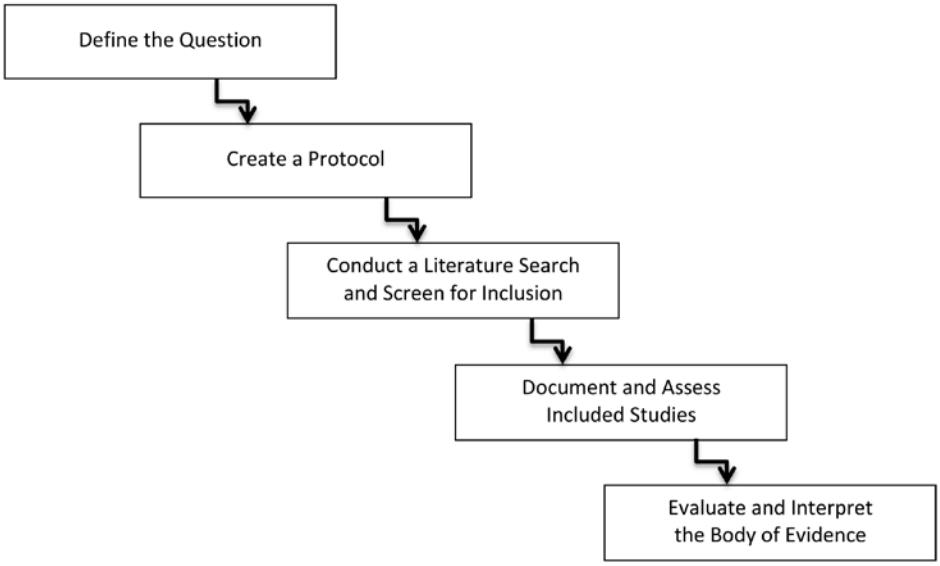
Furthermore, we elicited input from ten key informants over the course of the project (September 2015 to February 2016). Key informants were selected to represent users of systematic reviews in occupational safety and health and producers of systematic reviews in this area or related fields, such as patient safety. We asked them to share systematic reviews in occupational safety and health that would be of interest to the project team. We also requested references for systems to evaluate a body of evidence (i.e., all identified studies meeting inclusion criteria) in a systematic review on diverse types of evidence. In addition, we asked content experts to identify examples of grey literature that should not be included even in a systematic review of diverse types of evidence. Furthermore, we requested recommendations for frameworks to weigh different lines of evidence and approaches to determine when sufficient evidence exists to take protective action in the workplace. Finally, we asked key informants for published resources supporting systematic reviews in occupational safety and health and recommended literature on the topic that should be explored further.

NIOSH staff provided a draft framework for systematic review for occupational safety and health questions and a large number of published resources and literature relevant to occupational safety and health. Finally, NIOSH staff provided input throughout the project (Acknowledgments section) and reviewed the outline and earlier drafts of the report to ensure relevance and applicability to occupational safety and health systematic reviews.

Report Structure

Following discussions with NIOSH, we collated resources around five key systematic review steps shown in Figure 1.1 ("Define the Question," "Create a Protocol," "Conduct a Literature Search and Screen for Inclusion," "Document and Assess Included Studies," and "Evaluate and Interpret the Body of Evidence"). In each chapter, we describe the general systematic review step and its rationale, document relevant resources for occupational safety and health topics, and include operable resources for undertaking a systematic review as well as references to further in-depth discussion relevant to the topic of the chapter.

Figure 1.1. Systematic Review Steps



Chapter Seven of this report is dedicated to drawing conclusions and developing recommendations once the systematic review is complete. The report ends with a chapter titled "Discussion and Outlook." All cited literature can be found in the References section, and the links to online, publicly available documents other than journal publications are provided in the text and listed in Table S.1.

Objective

The report aims to provide resources for systematic reviews in occupational safety and health. It provides an overview of general steps in systematic reviews with a particular focus on occupational safety and health-relevant topics.

The remainder of this report comprises the following chapters: Chapter Two (Systematic Review Step 1: Define the Question), Chapter Three (Systematic Review Step 2: Create a Protocol), Chapter Four (Systematic Review Step 3: Conduct a Literature Search and Screen for Inclusion), Chapter Five (Systematic Review Step 4: Document and Assess Included Studies), Chapter Six (Systematic Review Step 5: Evaluate and Interpret the Body of Evidence), Chapter Seven (Draw Conclusions and Develop Recommendations), and Chapter Eight (Discussion and Outlook).

2. Systematic Review Step 1: Define the Question

The first step for systematic reviews is to *Define the Question*. This step results in an approach that requires thinking about the topic of interest not as a summary of available research in a given research or policy area, but as one or more questions that need to be answered. Furthermore, when systematic reviews are conducted to support guidance and recommendations issued by agencies such as NIOSH, this can be approached as

- question(s) the recommendation document tries to answer

and, more narrowly, as

- question(s) the systematic review tries to answer.

Hence, this chapter encourages those conducting systematic reviews to carefully think about what they want to find out.

Scope of the Review

Establishing the scope of the systematic review before it is executed is a key aspect of the systematic review process, and it determines the methods to be used. The scope of the review must be explicit so that the reader of the systematic review report has a clear understanding of what information was reviewed and what was outside the scope of the review. Establishing the scope also identifies exactly what should be determined from the existing literature through systematic literature-review methods. Some aspects of the topic of interest may not be included in the systematic review but may be important to understanding the review question. This information may include context, relevant theories, logic models, observations prompting the need for recommendations, or basic research knowledge. A recommendation document may include a range of contributing information, and determining the information that should be subjected to a systematic review of the literature needs to be clarified before the review is executed. Establishing that there is no relevant evidence about the topic of interest through systematic review methods is still very time consuming and resource intense (e.g., searching different databases, screening titles in duplicate, obtaining full text publications of potentially relevant citations, documenting the literature flow with reasons for exclusion). Hence, the aspects of the recommendations that should be based on a systematic literature search should be determined in advance with resources and utility in mind.

Ideally, a full systematic review should be conducted for those topics that have a body of research evidence and where there is a need to summarize results across studies in order to arrive at a complete and unbiased picture of the available information. The research literature is constantly growing, and we have increased access to it through better indexing in databases and

open-access publishing initiatives. This makes it likely that there will be numerous publications on any given topic. Systematic reviews ensure that all pertinent studies are considered, rather than citing selective evidence and thereby providing an incomplete or biased summary of the evidence. The methodology is widely accepted as a means to ensure that literature reviews are more objective. In addition, in some cases, meta-analysis may be employed to provide an objective approach to summarizing results across individual studies (Bushman and Wells, 2001). Meta-analysis has more statistical power than individual studies and can detect small effects that were not statistically significant in the original studies, and consequently ensures a more-precise estimate of the effect based on the cumulative evidence.

There are different ways to approach evidence in a systematic way. As alluded to in the introduction, different agencies have different definitions and standards for systematic reviews. However, it is undisputed that full systematic reviews of the literature require considerable resources, and there are some alternatives. Evidence-synthesis products such as rapid reviews may provide useful and resource-appropriate methods (see the section "Other Evidence Review Products" in Chapter Three). Rapid products such as evidence inventories and rapid reviews are less resource intensive, but use a systematic approach to evidence and a methodology that can be documented for transparency. In addition, all steps within the systematic review process can be scaled up or down (Woodruff and Sutton, 2014), as will be outlined in the next chapter.

Scoping Review

Before undertaking a systematic review of the literature, a scoping review is valuable. A scoping review is a brief literature search with a goal of obtaining an overview of the existing literature base before starting a detailed systematic review.

There are a few resources that should be routinely checked before undertaking a new systematic review. This includes databases for published systematic reviews. The Database of Abstracts of Reviews of Effects ([DARE]) summarizes more than 13,000 systematic reviews together with a critical appraisal of the review by trained systematic reviewers. However, no new reviews have been added to the database since March 2015. The entries are also indexed in PubMed Health, a publicly available database maintained by the U.S. National Library of Medicine (NLM). Another key resource is the collection of existing systematic reviews registered with the Occupational Safety and Health Review Group of the Cochrane Collaboration. Cochrane reviews cover a range of human health care and health policy questions and are indexed in the *Cochrane Database of Systematic Reviews* (CDSR). The Partnership for European Research in Occupational Safety and Health maintains a clearinghouse of systematic reviews. Some electronic databases, such as PubMed, offer a search filter for systematic reviews. The filter is more complex than simply searching for the term "systematic review," and identifies publications using deviations of the term, such as "systematic literature review." Identified systematic reviews may serve as an overview, or the scoping review may identify a suitable

systematic review that could be used as a starting point for the new review, which would allow the new review to focus on an updated search, see Chapter Four.

NIOSH maintains a searchable database of occupational safety and health publications, NIOSHTIC-2, which may identify related recommendation documents or other pertinent resources. The topic in question will determine which professional body or federal agency is likely to have published recommendations on the topic that may be used to inform the systematic review; for example, health promotion interventions may have been addressed in a practice guideline by the American College of Occupational and Environmental Medicine. Any scoping exercise should review ongoing federal and state evidence synthesis efforts to the extent possible. U.S. government agencies, including NIOSH, coordinate and communicate regarding current and upcoming systematic review topics to avoid a duplication of effort and to facilitate collaboration on overlapping topics. The International Labour Organization and the European Agency for Safety and Health at Work also include searchable publications databases.

It may also be useful to use sources of easily accessible knowledge, such as Wikipedia and Google Scholar, to identify issues of debate and open questions that consumers of recommendations may have. At this stage, the scoping review is not designed to identify specific sources to review; this step is meant to inform and to shape the systematic review question(s) and the scope of the review.

Another useful technique for scoping reviews is to create a citation report in the database Web of Science. The citation report will identify the most-cited scientific publications on the topic of interest. Other non–subscription-based databases also have sorting functions that select the most-relevant citations; the individual functionality depends on the database.

Review Questions

As outlined at the beginning of this chapter, for the systematic review, it is very useful to formulate specific review questions. The review questions will determine the purpose and refine the scope of the systematic review. A key process is the formulation of a review question that can potentially be answered using the literature. Features of answerable questions include that (at least theoretically) it is possible to give a concrete answer to the questions, such as, “What are the effects of a specific exposure on a specific outcome?” Determining review questions is very similar to hypothesis testing in an experiment where a hypothesis is formulated that can either be disproven or not with the empirical data.

To ensure that the question is sufficiently specific and relevant, it may address the elements participants/subjects, interventions/exposures, comparators, and outcomes; i.e., the basic Population, Intervention, Comparator, Outcome (PICO) framework covering the review inclusion criteria (see Chapter Four). A key consideration for occupational safety and health reviews is to determine for which population and under which circumstances answers are sought, such as the general population, employees who could be exposed during working hours, people

who are regularly in contact with a potential irritant, etc. Cochrane review authors do not use a question format, but articulate the review's objective using an equally specific format: "To assess the effects of [intervention or comparison] for [health problem] in [types of people, disease or problem and setting if specified]." Beyond the overarching review questions, there might be specific subquestions, such as, "Does the effect vary by duration or type of the exposure?" The UK National Institute for Health and Care Excellence has developed a checklist to evaluate whether a systematic review addresses an appropriate and clearly focused question relevant to the guideline review question; the tool is also helpful to systematic reviewers developing review questions.

The review team may discover, as a result of the systematic review, that there is insufficient evidence to answer the review question with confidence. However, it should be ensured that the review was asking an answerable question to begin with.

Review Team, Technical Expert Panels, Key Informants, and Stakeholders

The exact review questions are likely to influence the composition of the systematic review team. The review team should consist of content and methodological experts, with ideally two people who may serve as independent literature reviewers. The team also would benefit from including a librarian or information specialist to help design and execute the literature searches.

Systematic reviews in evidence-based medicine are often guided by a technical expert panel and key informants, and this is a useful practice for many topic areas. Key informant interviews should reflect the perspectives of those who would make a decision based on the subsequent recommendations as well as those who would be affected by those decisions (AHRQ, 2014). An expert panel supports the systematic review for the duration of the process and is expected to provide input on the review protocol and peer review the systematic review report. Public posting of the review protocol (see next chapter) can identify additional, pertinent questions the systematic review should address and ensures that topics are tied to real-world concerns and decisional dilemmas (AHRQ, 2014).

Technical expert panel members and key informants should represent diverse viewpoints and ensure that a number of sources have provided input. The panel composition should be guided by categories of stakeholders of the final product. An example is the 7Ps Framework for patient-centered outcomes research, which includes patients and the public (the consumers), providers, purchasers, payers, public policymakers, product makers, and principal investigators (other researchers) (Concannon et al., 2012). Individual stakeholders may have strong views and vested interests, and it is important to distinguish between the authors of the systematic review (the review team) and the advisors, such as key informants, expert panel members, and potential input from other stakeholders (e.g., comments received via public posting). Potential conflicts of interest need to be disclosed by all individuals providing input into the systematic review (Eden et al., 2011), and these should not be limited to financial conflicts of interest (Viswanathan et al.,

2013). Financial conflicts of interest should consider financial interests, employment, consulting, and individual and institutional research reports related to the systematic review topic. Nonfinancial conflicts of interest include professional interests, relationships, and activities. The conflict-of-interest determination should assess documented as well as perceived conflicts. Conflicts of interest can either be mitigated by excluding highly conflicted stakeholders or by balancing interest groups.

Practical strategies for eliciting input are teleconferences with a technical expert panel, interviews with individual key informants, written surveys, and online discussion boards. The modality needs to match the stakeholder: For example, eliciting input from patients and caregivers in evidence-based medicine reviews requires more preparation than consulting fellow researchers or policymakers. Teleconferences may be dominated by individuals, while an online survey ensures that all members are heard. Finding consensus can be challenging in larger groups. In addition, Paperwork Reduction Act clearance is required for federally sponsored data collections from ten or more respondents, often limiting the technical expert panel to nine members. To reduce the workload, panel members should be able to restrict their input to areas in which they feel qualified.

Identifying stakeholders that may help refine the systematic review question(s), and potentially the recommendation document questions, is a useful way to inform the review. Given that occupational safety and health recommendations are typically complex and may touch on many potential questions of interest, an expert panel may help prioritize the most-important questions that should be addressed in the systematic review. Other key input may be regarding selecting appropriate methods for different aspects of the planned recommendation document (i.e., which topic should be based on a systematic review) and scaling the level of effort ranging from rapid review to full systematic review methods. The Canadian Institute for Work and Health seeks stakeholder feedback on the relevance of the research question and clarity of final messages. Other authors have stressed the role of stakeholder engagement in disseminating the findings of systematic reviews (Keown, Van Eerd, and Irvin, 2008).

Stakeholder input can help to ensure that the systematic review team asks the right questions; can verify that all the relevant questions have been asked; may help rank the importance of individual questions; and may be able to determine which aspects of the topic are relevant but should not be subjected to a systematic review, i.e., shape the scope of the systematic review.

3. Systematic Review Step 2: Create a Protocol

This chapter introduces the concept, elements, and function of systematic review protocols. It is divided into protocol elements, analytic frameworks, conduct, function of systematic review protocols, and other review products. It lists standard protocol items (e.g., inclusion criteria), but also refers to other sections in this guidance document for more information. This chapter acknowledges that a full systematic review is not the only evidence synthesis product and outlines additional rapid products with fewer quality checks. This chapter primarily emphasizes the usefulness of an *a priori* road map for the planned evidence synthesis.

Systematic Review Protocol Elements

There are several standard elements of protocols for systematic reviews in health and evidence-based medicine. The Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) statement now includes a 17-item checklist for protocols called *Preferred Reporting Items for Systematic Review and Meta-Analysis Protocols (PRISMA-P)*. The checklist ensures that all elements that are considered standard reporting are covered in the protocol. The checklist is endorsed by a large number of scientific journals. Five items address administrative information. The remaining items are dedicated to the introduction section and the methods of the systematic review:

- Rationale
  - should describe the review in the context of what is already known.
- Objectives
  - should address explicit review questions.
- Eligibility criteria
  - should cover inclusion and exclusion criteria.
- Information sources
  - should include electronic databases searched and grey-literature sources where appropriate.
- Search strategy
  - draft search strings should be provided.
- Study records
  - data management

        - mechanism used to manage records and data, e.g., use of an electronic database for tracking and data extraction
    - selection process
        - process used for selecting studies, e.g., independent reviewers
    - data collection
        - process of extracting data, e.g., use of piloting, extraction in duplicate.
- Data items
    - include the data-extraction variables, definitions, and planned data manipulation.
- Outcomes and prioritization
    - the protocol should describe the eligible measures of effects and the primary outcomes of the review.
- Risk of bias in individual studies
    - criteria used to assess individual studies should be described.
- Data synthesis
    - the criteria for the decision to quantitatively synthesize study data; planned summary measures, methods of handling data, and methods of combining data from studies, including any planned exploration of consistency (e.g., heterogeneity estimate); any proposed additional analyses (e.g., sensitivity or subgroup analyses, meta-regression); if quantitative synthesis is not appropriate, describe the type of summary planned.
- Meta-bias(es)
    - any sources of bias across studies, e.g., publication bias that will be assessed.
- Confidence in cumulative evidence
    - criteria to assess the body of evidence should be made explicit.

The PRISMA group published a manuscript for further information (Shamseer et al., 2015).

Increasingly, systematic review protocols are registered, similar to registries for controlled trials. The database PROSPERO is a publicly accessible repository of systematic review protocols. One of the main goals of PROSPERO is to help reduce unplanned duplication and to increase transparency of systematic reviews. The PROSPERO record requires prospective specification of the objective and the primary outcome of the review to minimize selective reporting (e.g., peer reviewers can compare planned methods with the final review). Protocols for Cochrane systematic reviews are indexed in the CDSR and PubMed; Campbell Library review protocols are indexed in the Campbell Library. Of note, review authors are asked to specify the main review outcomes, i.e., all those that will be documented in a summary of findings table (see Chapter Six) *a priori* at the protocol level. Journals such as *Systematic Reviews* publish systematic review protocols after they have undergone peer review, a process

recommended by the Institute of Medicine (IOM) standards for systematic reviews (Eden et al., 2011).

While the PRISMA statement covers essential elements expected and often required by scientific journals and repositories, published (publicly available) protocols may include additional elements and serve as useful resources for new review protocols for occupational safety and health questions.

PRISMA elements refer to the systematic review of the literature. When systematic reviews are explicitly conducted to support formal recommendations and guidelines (see Chapter Seven), it can be useful to add steps to describe how the systematic review will feed into the recommendations. For example, the recently revised principles and processes for dealing with evidence in scientific assessments outlined by the European Food Safety Authority (EFSA) highlight that a key part of transparency and openness stems from systematically documenting and reporting all steps of the assessment and guaranteeing accessibility to data and results, as well as sharing the plan of the assessment with relevant parties in advance in order to seek feedback and input. Hence, the strategy for the assessment is determined *a priori*, i.e., before starting the assessment (EFSA, 2015).

Analytic Framework

Analytic frameworks may serve as analysis plans for the systematic review. This framework is not standard: It depends on the specific systematic review questions, which will be very different across reviews because issues addressed in occupational safety and health reviews vary widely.

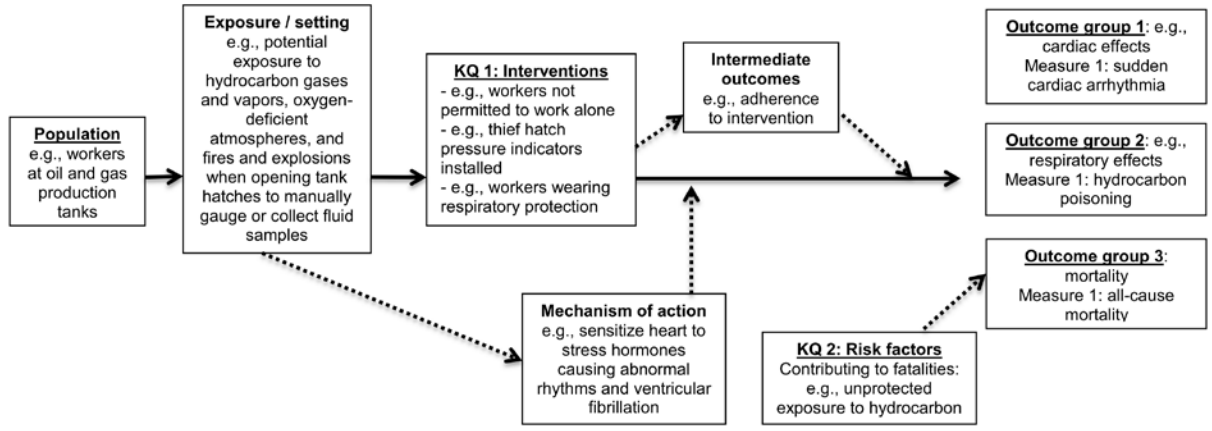
The analytic framework may depict the review question(s) and outline how the intervention or exposure of interest is linked to the outcomes of interest. While the exact mechanisms of action may not be fully understood at the time the protocol is written, analytic frameworks are helpful in differentiating the review question elements (e.g., population, outcomes) and their relationships (e.g., direct connections, assumption of effect-modifying factors).

The example in Figure 3.1 incorporates selective elements of a hazard alert on health and safety risks for workers involved in manual tank gauging and sampling at oil and gas extraction sites to illustrate analytic frameworks (the framework in Figure 3.1 did not guide the hazard alert and is only used here to demonstrate the general points of analytic frameworks). The framework documents the population (the target of the guidance) and the outcomes of interest grouped by cardiac and respiratory effects and mortality. The framework also incorporates intermediate outcomes, such as the adherence and uptake of protective interventions in practice. The framework shows two key questions addressed in the systematic review, one on the effects of interventions to protect workers and one on risk factors that may contribute to fatalities. In this hypothetical example, the harmful exposure is depicted as a setting characteristic and is not part of the systematic review (e.g., the review does not summarize the research on effects of exposure as such). The mechanism of action linking the exposure to outcomes is shown in the framework

to help the reader understand the conceptual framework of the review, but this element is background information; it is also not part of the systematic review (e.g., a systematic review aiming to identify published mechanisms of action for the exposure and associated effects). As shown, diagrams can get very complex very easily, and the review team will need to decide on the key information to be shown in the analytic framework.

Analytic frameworks may also serve as conceptual frameworks that differentiate underlying, theoretical, or broader concepts (e.g., safety) and the indicators that can be measured in research studies (e.g., specific adverse events, such as sudden cardiac death or increasing the risk of cancer as a long-term effect). Furthermore, as Figure 3.1 illustrates, the framework may function as a model that explains the logic underlying the relationship between exposure and effects or between interim and definitive outcomes (e.g., a documented increase in protective behavior [interim outcome] and a reported decrease in incidents of injuries [definitive outcome]). In addition to the review question, the specified inclusion criteria of the review may be incorporated into the framework.

Figure 3.1. Analytic Framework Example



Note: KQ = key question

Analytic frameworks are standard elements of U.S. Preventive Services Task Force (USPSTF) reviews. USPSTF statements provide a large number of examples for a range of systematic review questions. The Guide to Community Preventive Services recommends developing a conceptual framework and an analytic framework.

In the context of occupational safety and health reviews, the analytic framework also can be used to show how different types of evidence, including different arguments and lines of evidence, contribute to the systematic review question. In this case, the framework may incorporate direct evidence coming from human exposure studies and mechanistic evidence showing the assumed association between exposure and effects. The framework will help to differentiate the evidence elements and will clarify how different pieces of evidence are connected. This will help the reader understand why specific lines of evidence were chosen for

consideration for the systematic review and why they were selected to contribute to the evidence base for the review question. The framework may also showcase the analytic synthesis, i.e., the approach to synthesizing research, grading the body of evidence, and integrating different lines of evidence.

Furthermore, the systematic review framework may draw on published models, such as established risk assessment frameworks (Maier et al., 2014) and adverse outcome pathways (Ankley et al., 2010). It may also draw on the International Agency for Research on Cancer framework to assess the strength of the available evidence that an agent could alter the incidence of cancer in humans, which combines exposure data, cancer incidence in human studies, cancer incidence in experimental animals, and mechanistic and other relevant data in a standardized approach (World Health Organization [WHO], 2006).

Finally, in the context of systematic reviews undertaken to support recommendations by agencies such as NIOSH, the analytic framework may be limited to the specific systematic review questions and evidence eligible for inclusion in the review. Alternatively, as Figure 3.1 shows, it may include background material used to frame the review questions and considerations that contribute to the final recommendation, but that are not part of the systematic literature review.

Methodological Conduct of the Review

Systematic reviews aim to provide a reliable, valid, and transparent overview of the available evidence. Emphasis is placed on the transparency of methods that should be documented in detail to enable replication. Many steps are undertaken to reduce reviewer errors and bias in the review process. Conclusions are based on the presented data, and the reader should be able to follow the conclusions after reading the methods and results of the review.

A main aspect of the methodological conduct in systematic reviews is the safeguard against selection bias, i.e., bias in the material on which the systematic review is based. As discussed in an earlier section, standards for systematic reviews vary somewhat by the agencies producing the reviews. In addition, many standards are based on agreements in the scientific community and not on empirical evidence. Nonetheless, a guiding principle of systematic reviews is to ensure that all pertinent evidence has been considered and to ensure that the review is not based on a highly selective or biased selection of evidence. A comprehensive literature search is a key component to protect against selection bias. Apart from documenting a reproducible and systematic search strategy that captures the key literature, emphasis is placed on the extent of the search. The use of multiple databases to identify relevant research is a standard mechanism to ensure that pertinent studies will not be missed. Documenting at least two sources for identifying primary research was, for example, one of the inclusion criteria for reviews included in the database DARE. This standard is based on early research using a pool of known studies as the gold standard and comparing it with the proportion of studies that can be retrieved by searching

only one database (Dickersin, Scherer, and Lefebvre, 1994). Results vary by research type and database, but searching multiple databases is common practice in systematic reviews. Multiple sources may be different electronic databases (e.g., PubMed and PsycINFO) that cover the topic of interest. Other sources may be reference mining (e.g., systematically searching the bibliographic references of pertinent reviews and studies meeting review inclusion criteria).

A second key approach that specifically aims to minimize selection bias targets the procedure of selecting the review material. The use of independent reviewers to select studies for the review from the search output retrieved in the systematic searches is a key characteristic of systematic reviews. Using two independent reviewers serves different functions at different stages of the review and can guard against both random and systematic errors (bias). For example, screening a large search output in duplicate is a simple safeguard for not missing potentially relevant studies. At this stage, reviewers have access only to limited information (i.e., the title, abstract, citation details, and keywords). In many cases, reviewers will not have sufficient information to accurately determine whether it is worth pursuing a citation. As a general rule, the citation should be obtained as full text if at least one reviewer thinks it may be relevant. Of note, in some cases, it may be possible to rely on automated approaches to inclusion screening. For example, instead of using two independent reviewers, one may be replaced by a trained algorithm. Required resources to program the algorithm and its accuracy as well as available resources for dual review may determine whether this is a suitable option.

The next stage, the full text review of potentially relevant publications, should be performed in duplicate to ensure that at least two people agree that the study meets all inclusion criteria of the systematic review. At this stage, explicit inclusion and exclusion criteria have been formulated that should be applied systematically. Any disagreements between reviewers should be resolved through discussion by the reviewers or in the wider review team. Dual inclusion screening is one of the IOM systematic review standards (Eden et al., 2011).

Other aspects of the methodological conduct of systematic reviews are to report all review methods and decisions transparently and to describe the identified literature in a way that the reader can follow. Tools such as flow diagrams, evidence tables, and summary of findings tables described in the following chapters of this report are explicitly designed to enhance the transparency of systematic reviews of the literature. The reporting standards for systematic review protocols were highlighted in the previous section; reporting standards for completed reviews are presented in Chapter Seven.

Finally, A Measurement Tool to Assess Systematic Reviews (AMSTAR), a tool to assess the methodological quality of systematic reviews, may guide the review methods. The 11 items in the tool were derived by applying a large list of quality items to an empirical sample of reviews and then applying principal-component analysis to identify underlying factors. The items cover an *a priori* design (question and inclusion criteria), duplicate study selection and data extraction, comprehensive literature search (at least two electronic sources), and inclusion regardless of the status of the publication (unpublished grey literature, all languages). Furthermore, the items

address whether a list of included and excluded studies is provided, the characteristics of included studies are documented, the quality is assessed, the methodological rigor and scientific quality are considered in the analysis and conclusion(s) of the review, appropriate methods are used to combine results across studies, publication bias is assessed, and any conflict of interest is stated. The assessment tool has several modifications and extensions for specific research fields (Shea et al., 2007, and Pieper, Mathes, and Eikermann, 2014).

Systematic Review Protocol Function

This section suggests how to best use systematic review protocols and outlines how protocols serve a function for the reviewers, the research field, stakeholders, and the consumers of the systematic review.

First, as a communication tool, protocols may be used to elicit targeted input from experts and stakeholders. Public posting of systematic review protocols is another way to increase transparency, and many agencies have adopted this method. Practically, this means that those conducting systematic reviews have web space with a feedback function available to them, including developing a mechanism to alert stakeholders to the posting, determining the period of time the review should be up for posting, and determining for how long and to what extent to incorporate input into the review process. Review comments may be anchored in vested interests, and input should be critically reviewed. Highly contentious areas may benefit from a point-by-point reviewer disposition table documenting how the review team addressed comments.

Protocols of systematic reviews supporting recommendations can serve a very practical function. Protocols can be used as an accompanying document or repository that provides more details on the review methodology, sources, the electronic search strategy, or definitions. This allows the final recommendation document to be streamlined with fewer methodological details that could detract from the questions being examined. Other producers of systematic reviews, such as the AHRQ, publish the systematic review protocol together with the evidence reports. The protocol may also provide information on the sources used to identify evidence, explain search strategies such as reference mining, or provide a rationale for the search strategy (e.g., restricting to literature published after a certain benchmark study). Considering the different audiences and consumers of the subsequent guidelines document, not all users of the recommendation issued by NIOSH or other agencies may be interested in the full methodological detail of the systematic review and the document could refer to a detailed review protocol instead.

Other Evidence Review Products

The systematic review protocol determines the methods that will be used to conduct the review. Products other than systematic reviews may be used to address some or all of the aspects targeted

in a recommendation document. A review (Hartling et al., 2015b, and Hartling et al., 2016) of rapid products from 20 organizations differentiated four types of products:

- Rapid Reviews
- Evidence Inventories
- Rapid Responses
- Automated Approaches.

Rapid Review is the methodology most similar to systematic reviews, but with a narrower scope and fewer quality checks (e.g., using only a single reviewer). An *Evidence Inventory*, which primarily lists the available evidence without synthesis, is another potentially relevant methodology that may be employed. *Rapid Responses*, another type of evidence product, answer a review question, but only with the best-available evidence and no synthesis of the existing evidence. This type of approach does not strive to summarize all available evidence, but identifies any existing systematic reviews on the topic (or other strong research designs in the absence of systematic reviews). *Automated Approaches* use algorithms to generate a synthesis with the help of computer software. Machine-learning and citation-processing advances are increasingly employed to assist with steadily expanding research fields (Hempel et al., 2012b, and Dalal et al., 2013). Software, such as Abstrackr, as well as collections of text-mining algorithms, are available to support systematic reviews.

The USPSTF manual differentiates three types of evidence review products: *full systematic reviews, targeted systematic reviews*, *and staged reviews. Full systematic reviews* address every key question in the analytic framework. This is the default approach for all new topics. *Targeted systematic reviews* analyze selected questions in the analytic framework for which the evidence is not clearly established and is often used to update topics. Finally, *staged reviews* are systematic reviews for selected questions in the analytic framework. These are undertaken to establish or to revisit evidence insufficiency; serious gaps in evidence may preclude coming to a recommendation.

These evidence-synthesis products still use a systematic and transparent approach, but are less resource-intense because some components of systematic reviews are simplified or omitted (Tricco et al., 2015). Compared with full systematic reviews, the production process is accelerated or streamlined, which makes them particularly well suited to respond to questions regarding the evidence in a timely manner (Ganann, Ciliska, and Thomas, 2010). Timely products are especially important when developments require that advice is available quickly, for example, for emerging hazards or emergency response. The WHO has developed a process for Rapid Advice Guidelines that can be produced in fewer than six months (Schünemann et al., 2007b). The guidelines are based on literature searches and expert input. The literature review concentrates on systematic reviews of randomized controlled trials, recent trials, and selected nontrial evidence (case reports, animal and in vitro studies for specific interventions), and the evidence is summarized for predetermined, expert-selected outcomes using the GRADE approach. Examples of potential applications of rapid reviews in occupational safety and health

are NIOSH Alerts. The documents are developed in a relatively short time frame to identity emerging or priority health and safety risks and provide recommendations to prevent those hazards.

Rapid review methods allow a relatively quick review of the evidence when recommendations are needed because of the nature of the topic (e.g., emerging hazards), but there are other applications using alternatives to systematic reviews. A 2015 rapid review summit highlighted the need for up-to-date information as one of the reasons for the increased demand in rapid reviews (Polisena et al., 2015) given that the research base constantly develops. The Canadian Knowledge to Action research program has adopted rapid reviews to serve as an informative brief that prepares stakeholders for discussion on a policy issue, to support the direction and evidence-base for various health policy initiatives, and to support the development of clinical interventions and/or health services programs (Khangura et al., 2012). The methodology follows eight steps: (1) needs assessment, (2) question development and refinement, (3) proposal development and approval, (4) systematic literature search, (5) screening and selection of studies, (6) narrative synthesis of included studies, (7) report production, and (8) ongoing follow-ups and dialogue with knowledge users. In a project providing evidence-based support for clinical practice innovations, a type of rapid review was chosen in order to respond to specific questions posed by practitioners (Danz et al., 2013). The rapid reviews were produced for individual teams of practitioners adhering to short turnaround times; the reviews were conducted by an independent evidence-synthesis team in order to avoid a biased evidence syntheses for practitioner-selected innovations; and the questions addressed in the reviews were very specific (e.g., evidence for implementation strategies for a specific organizational intervention), which made them suitable for rapid reviews.

The criteria for using rapid-review methodology in the given examples included the need to synthesize evidence in a timely manner to inform emergent decisions faced by decisionmakers because of the nature of the review topic, limited resources that do not allow a systematic review but rather require an unbiased overview, responsive evidence summaries that are requested that target specific information needs, ongoing research developments that require up-to-date information, or a need to inform questions in a project-specific time frame and systematic reviewers needed to make a decision about what can be provided in the allotted time (Danz et al., 2013; Hartling et al., 2016; Khangura et al., 2012; and Polisena et al., 2015). Hartling et al. (2016) elicited AHRQ end-user perspectives of rapid reviews and found that evidence inventories and rapid responses were seen as useful for "hot" or timely topics; in an area with limited literature; to understand depth and breadth of evidence; to clarify whether a review is already available; or to ignite or catalyze change, or to challenge the status quo. End users also thought rapid reviews could be useful for guideline or recommendation development and updates, new issues subsequent to a guideline or recommendation, coverage decisions, organizational or policy change, implementation, quick decisions, and when no previous systematic review or guidance exits. In contrast, systematic reviews were seen as essential for

broad topic areas and population issues, to inform research agendas, and for in-depth understanding of a topic area.

The existing literature has repeatedly pointed to the lack of established rapid review methods (Featherstone et al., 2015; Khangura et al., 2012; and Polisena et al., 2015). More methodological guidance is expected to be published by the Cochrane Rapid Reviews Methods Group. HLWIKI International maintains a website dedicated to resources for rapid reviews.

Scalability of Evidence Review Methods

All evidence synthesis projects have constraints, whether they are resources, time, or money, and it is important to effectively manage the scale and scope of the systematic review or recommendation document appropriately. The importance of selecting an appropriate scope of a review—i.e., the content the review will cover—was discussed in Chapter Two. Other approaches to limit the required resources of the evidence review are the review methods.

The resource intensity of full systematic reviews is a recognized concern, and considerable interest in rapid review products has been generated among researchers and policymakers, particularly in rapidly evolving fields (Hartling et al., 2016). This has sparked the discussion about the scalability of review methods in combination with maintaining the systematic approach to literature summaries. Reviews of published rapid reviews (Ganann, Ciliska, and Thomas, 2010, and Tricco et al., 2015) outlined methods that allowed for faster review production. Methods focused on the amount of literature to be reviewed, while other rapid review teams applied strict inclusion criteria to the retrieved search output to restrict the included study pool. Other methods to accelerate the review targeted the process of reviewing the included studies. Some reviews used a single reviewer for data extraction. Some used single-reviewer critical appraisal for all studies or for studies with smaller impact on the conclusion (e.g., observational studies) and some did not undertake formal quality assessment. In some cases, single-reviewer methods were used for the inclusion screening process. The rapid review products reviewed by Hartling et al. (2015a) used methods differing from systematic reviews at all stages, affecting the scope (limited questions, limited number of questions, limited number of included studies); comprehensiveness (limited search strategy, study type, data extraction); rigor and quality control (no dual study selection or data extraction, no peer review); synthesis (limited or eliminated risk of bias assessment, quantitative analysis, quality of evidence grading); and conclusions (simplified or eliminated conclusive statements about evidence).

The following list shows a number of approaches that allow for a more-rapid and less resource-intense review production. The list is ordered by the systematic review steps that correspond to the chapters in this report:

- Define the question scalability options:
  - Restrict the topics that will be subjected to a systematic review.
  - Build on existing reviews and recommendations.

  - Limit the number of review questions.
  - Limit the complexity of review questions.

- Create a protocol for scalability options:
  - Use a single reviewer for title and abstract screening.
  - Use a single reviewer for full text screening.
  - Use a single reviewer for data extraction.
  - Use a single reviewer for critical appraisal of all or some of the studies.
  - Do not use a technical expert panel or key informant input.

- Conduct a literature search and screen for inclusion of scalability options:
  - Restrict the number of searched databases.
  - Apply date restrictions.
  - Use studies included in existing reviews with or without updating the search.
  - Apply language restrictions.
  - Restrict the search by study design or publication type.
  - Restrict the search by study outcome or setting.
  - Limit the search for or exclude grey literature and unpublished data.
  - Do not consult with experts to identify additional studies.
  - Use only readily available published research (open access).
  - Apply strict inclusion criteria to participants and subjects, intervention and exposure, study design and comparator, outcomes and type of data, follow-up period or intervention/exposure timing, and setting.
  - Apply strict inclusion criteria to the publication type (e.g., peer-reviewed research).

- Document and assess included studies of scalability options:
  - Minimize data extraction (number of variables).
  - Minimize coding/abstraction (e.g., using the authors' conclusion without standardized result extraction by reviewers).
  - Do not follow up with original authors for missing data.
  - Limit or do not perform a critical appraisal.
  - Perform critical appraisal only for specific study designs (e.g., studies that carry the most weight in the evidence synthesis).
  - Do not follow up with original authors for missing information relevant to assess study quality.

- Evaluate and interpret the body of evidence on scalability options:
  - Limit synthesis efforts (e.g., list studies rather than comparing and contrasting or summarizing).
  - Do not perform a meta-analysis but use the authors' conclusions from the individual studies.
  - Do not perform a quality-of-evidence assessment.
  - Do not integrate evidence (e.g., defer to expert panel).

Only after reviewing the individual chapters in this report will it be possible to fully appreciate these methods and to understand where they deviate from standard systematic review methods. The following chapters provide specific details of the review steps and highlight the rationale for the approach (e.g., for why a dual-reviewer process is recommended for inclusion screening, see Chapter Four). In addition, not all listed approaches to scale back the effort of systematic reviews may be feasible: For example, when restricting the literature search, it is important to determine whether the selected criteria are searchable (i.e., the features have to be indexed in databases). A review might decide to limit the search to studies in a particular population group, but this is only possible at the search level if the population characteristics were systematically coded in potentially relevant studies. As a general rule, restricting the number of included studies will affect all other steps in the review because it affects resources needed to complete steps such as data extraction (e.g., extracting data for ten or 100 studies). In addition, restricting the types of eligible studies will affect all other review steps and determines, for example, the necessity of finding different appropriate critical-appraisal tools and the complexity of the evidence-integration step. The type of potentially relevant studies should be critically reviewed with regard to the impact on the review's conclusions: Some studies will allow stronger evidence statements than others. For example, while there might be a number of case studies reporting on an outcome of interest in the literature, a synthesis of these studies will not be able to quantify the risk for participants. Other studies may report on an adverse event, but without a comparator group to establish the baseline risk in the study population, the relative frequency of the outcome cannot be determined. Determining which studies should be eligible for inclusion should be discussed in sufficient detail to ensure that the most-appropriate evidence is being reviewed.

In order to decide which review methods would be most appropriate, it is important to identify and discuss the potential constraints when designing the review protocol. The time needed for screening and documenting studies depends on the topic and the literature, but a general rule of thumb is that, per hour, a literature reviewer can screen 100 titles and abstracts from a database search output, screen six full-text publications, or data extract and critically appraise one or two studies. Scoping searches will help estimate the size of the literature, and pilot tests will allow a more-specific estimate of reviewer time.

Questions such as the following may help in the decision process of selecting the review scope and the methods: How much time is there to complete the review? Does the urgency of required recommendation suggest rapid review methods? Are there existing systematic reviews that could be updated, used as the research study pool, or to validate that all relevant studies have been identified? Are there parts of recommendations from other agencies that could be adopted or incorporated in the new recommendation document? Which recommendation aspects need examples rather than a complete collection of all available research (e.g., undisputed basic science explaining the mechanism of action)? What are the implications of missing studies in an abbreviated search (i.e., are there a large number of studies known that show consistent results or

can one study make a substantial difference in the evidence summary)? Are there sufficient human studies so that no other lines of evidence need to be considered? Are there studies that are key to answering the review questions, and can the review be limited to the best evidence? Can the systematic review be limited to studies in occupational settings or workers? Which databases can be accessed free of charge? Are there benchmarks or developments in the field that narrow the date range of studies that should be reviewed? Are content experts available that know the research area well to justify an abbreviated search? How many reviewers are available? How experienced and knowledgeable are the reviewers about the topic (e.g., a single-reviewer approach depends on the accuracy of the specific reviewer)? Did a pilot test suggest the inclusion criteria are easy to interpret, or are two reviewers needed to avoid missing pertinent studies? Are the data easy to extract from studies, or do they need to be transposed (e.g., converted to a common metric)? Does the data extraction require statistical training? Did a pilot test suggest the data extraction is prone to errors, or can one reviewer be trusted to do it? How difficult is it to assess the quality of the studies (do two independent reviewers come to the same conclusion)? Is the quality of evidence assessment likely to be controversial? Is the evidence integration across lines of evidence likely to be complex? What are the available funds to elicit expert panel support, and what is their availability to respond in time?

Transparency of Review Methods

As discussed previously, the steps undertaken in systematic reviews are designed to reduce reviewer errors and bias to arrive at a reliable and valid summary of the literature. However, it should also be kept in mind that there is only limited research available that has empirically determined the effect of any shortcuts or deviations from systematic review standards (Tsertsvadze et al., 2015). Only a few studies have investigated the incremental value of specific systematic review methods (Dickersin, Scherer, and Lefebvre, 1994; Moher et al., 2000; Bushman and Wells, 2001; Horsley, Dingwall, and Sampson, 2011; Giustini and Boulos, 2013; Selph, Ginsburg, and Chou, 2014; and Haddaway et al., 2015) or have tested the validity of the end product by comparing the conclusions reached in rapid reviews versus systematic reviews (Watt et al., 2008a; Watt et al., 2008b; and Hartling et al., 2015a). Hence, we know very little about the impact of deviating from full systematic review methodology.

At a minimum, review authors need to document the methodology and alert readers to potential limitations, such as the lack of duplicate screening (Oxman, Schünemann, and Fretheim, 2006). Any shortcuts and efforts to reduce the resource intensity of the review have to be clearly documented and should be determined at the protocol stage, not when the review is already under way.

4. Systematic Review Step 3: Conduct a Literature Search and Screen for Inclusion

This chapter is divided into the following sections: data sources, search strategy, eligibility criteria and inclusion screening, and flow diagram.

Data Sources

Data sources can refer to the method of identifying research literature (e.g., via electronic databases, by consulting with experts) or the type of information that is being sought (e.g., articles published in scientific journals, reports or other publications, conference abstracts, unpublished data). This section is divided into the following subsections: databases, grey literature, publication sets, and other sources representing key issues relevant to data sources.

Databases

The main source for any systematic review will be electronic databases of published scientific literature. One of the best-maintained databases of peer-reviewed scientific literature is PubMed. It can be accessed free of charge through the NLM National Institutes of Health and includes more than 25 million biomedical literature citations. The continuously expanding NLM platform includes a number of other smaller collections (e.g., PubChem Substance, a database for chemical structure information). Other large databases include the Web of Science, Scopus (broad spectrum of indexed journals), and PsycINFO (psychological literature). Specialist databases such as GreenFILE (environmental journals and books) or TOXLINE (biochemical, pharmacological, physiological, and toxicological effects of drugs and other chemicals) cover a much more defined scope. The NIOSH database NIOSHTIC-2 contains more than 50,000 occupational safety and health information resource citations. A WHO report on the use of research information to improve the quality of occupational health practice provides an annotated list of databases relevant to occupational health (Verbeek and Van Dijk, 2006). Appendix B of the operations manual of the American Conference of Governmental Industrial Hygienists provides a list of databases that are routinely searched when preparing threshold-limit values for chemical substances. The U.S. Environmental Protection Agency (EPA) maintains an online database collection relevant to risk assessments.

Databases vary in their range, accessibility, and user-friendliness. The range can be described in terms of the content and topics covered but also by the range of journals indexed in the database, whether it is limited to citations (title, abstract, keyword, full citation) or full-text searches, or whether journals and books are indexed. Before searching the database, it should be established if the indexing scope is compatible with and sufficient for the review question (e.g.,

by establishing whether a key journal for the review topic is indexed and since when). PsycINFO, for example, indexes articles going back 100 years, while other databases have a much-later date of inception. The Web of Science indexes conference abstracts more thoroughly than other databases. Some databases can be accessed free of charge and others need to be purchased for a period of time and charge per downloaded record. Some databases, such as MEDLINE, are offered on different platforms, such as PubMed and Ovid. The platforms influence the search functionality. In addition, the search syntax differs by database. Information specialists or librarians trained in the conduct of systematic review searches can inform, strengthen, and improve the quality of the literature search process.

Grey Literature

The decision to include grey literature is another key issue that needs careful consideration. Grey literature is information or research not published as articles in peer-reviewed scientific journals. This may include reports produced by government and nongovernment agencies or organizations and private companies that are in the public domain, including high-quality systematic reviews published by the Campbell Collaboration. This information would be missed when limiting the search to major electronic databases targeting traditional academic publishing channels. There are other aspects to be considered, but one is the ability to systematically identify grey literature. As an example, some AHRQ EPC reports are indexed in PubMed and others only in PubMed Health. Some databases specialize in grey literature, such as Grey Literature Report, or include large quantities of grey literature, such as WorldCat; however, these searches may retrieve a large number of irrelevant citations, and the exportability to citation management software is limited.

Conference abstracts constitute another type of grey literature. Including conference abstracts in systematic reviews is a source of ongoing debate. Not all research studies get published; this is known as the "file-drawer" problem. This would not be of much consequence for systematic reviews were there no relationship between study characteristics and publication status. However, publishing is associated with statistically significant study results. This seems to be primarily a function of the authors' diminished interest in the research, rather than journals not accepting nonsignificant results (Dickersin, Min, and Meinert, 1992). For systematic reviews, this means that it cannot be assumed that the published literature is a representative sample of all existing studies. The published literature is likely biased toward statistically significant results. Including conference abstracts that require less effort than full publications may reduce publication bias, given that many abstracts are never published as full articles (Von Elm et al., 2003) and publication is associated with a positive result (Scherer, Langenberg, and Von Elm, 2007). In addition, conference abstracts are often published before a full publication, thereby representing the latest research at the time of a review. On the other hand, peer review in scientific journal submissions serves as a quality check, and many times, problems with the data are discovered and corrected during this process. Furthermore, the limited information available

in these abbreviated formats may be seriously misleading when summarizing the results. Dissertations and theses are another source of research not (yet) published in scientific journal formats.

Contacting manufacturers for unpublished data is an additional step that some researchers conducting systematic reviews undertake. For some areas, regulatory data, such as risk notifications sent to the EPA (ChemView, Chemical Data Access Tool) or data submitted to the Food and Drug Administration, may be available. Other agencies, such as the American Conference of Governmental Industrial Hygienists, publish a list of substances "under study" and invite interested parties to submit data. Unpublished data have to be citable and released upon request. Contacting known researchers for additional, unpublished data may also be an option. Response success is limited and depends on several factors.

Caveats for including unpublished data and grey literature are based on quality considerations, the lack of systematic identification methods, and the fact that the data are not in the public domain and therefore not accessible to everyone.

Publication Sets

Many databases include special features that can help identify sets of publications that have been precollated or grouped by databases. The "methodology filter for empirical studies" in PsycINFO, for example, can help focus the literature search by restricting searches to publications reporting empirical data. The literature review may be limited to a specific set of publications and the use of a published filter will help define the selection (e.g., all publications on the topic that are indexed as randomized controlled trials).

Another example is publications found through a related article search. The "related article function" in PubMed is an algorithm that identifies publications similar to a seed article (i.e., a publication that is exactly on topic). Content experts are a great source for identifying seed articles.

The Web of Science offers the option of a "forward search" function. A forward search systematically identifies all studies that have cited a key article. The fact that the publication has cited the key article makes it likely that it deals with a similar topic, even if this connection would not have been detected through a standard keyword search.

Other Sources

Reference mining is a key resource for identifying relevant publications. Reference mining systematically screens the bibliographic citations of included studies and pertinent reviews on the topic. Often, the source publication can provide valuable information to determine whether the citation is likely to be relevant, making it a resource-effective search strategy.

Hand searching is less common in systematic reviews now. It describes the process of manually screening tables of contents of relevant journals to ensure that no relevant studies have

been missed. This approach is very time consuming and should perhaps be limited to topics for which no clear search terms can be developed.

Another way to identify research studies is to search funding agencies' websites. This approach does not search for the resulting publication. Searching for publications or contacting authors for unpublished data would be a follow-up step. Similarly, trial registries such as CENTRAL, ClinicalTrials.gov, or the WHO research database are additional sources that can be used to locate studies, given that most journals now require trials to be registered in advance. Some allow registration of observational studies, and some registries track publications associated with the study.

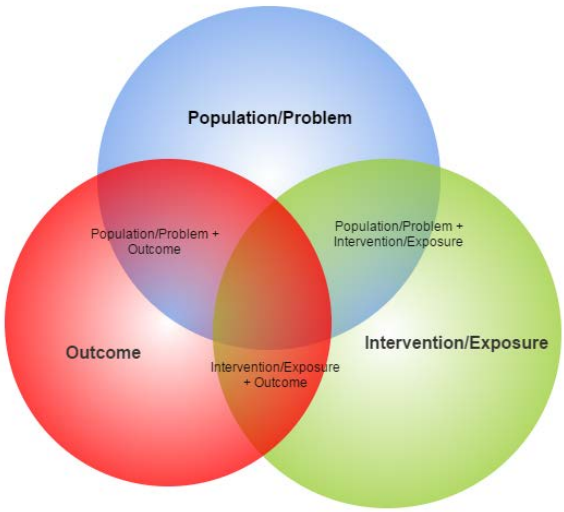
A way to quickly locate the most relevant studies is to ask content experts. This can be very useful when setting up the systematic review. The identified articles can be used to validate the search strategy (see next section) as well as to apply the inclusion criteria (which may reveal that more detail needs to be added). The articles also can be used to draft the data extraction using concrete examples, and will help to select critical appraisal tools as outlined in the following chapters. For complex topics in particular, it is also useful to ask experts to review the included and the excluded studies after completing the searches (e.g., during peer review) to ensure that no studies have been missed.

Finally, a scoping review may have discovered one or more existing systematic reviews that could be used as the main source for relevant studies up to the respective search date, allowing the new search to focus only on new research. Obviously, the inclusion criteria and methodological conduct need to be scrutinized for comparability. Note that there are different ways to use existing systematic reviews; in this context, they are used as a source of included studies. Another approach is to synthesize the existing review and any new studies that have been identified in a new search (to ensure the new review includes all currently available evidence). This poses a challenge of having to integrate primary (study-level) and secondary (review-level) evidence. In some cases, a systematic review of systematic reviews may be appropriate. Where there is more than one review available on a topic, the integration of existing reviews is particularly complicated because the reviews are likely to show some, but not perfect, overlap. Synthesis is hindered since the studies included in the reviews cannot be treated as independent units because individual studies may be included in more than one review. The EPC program recently published recommendations regarding the integration of reviews that may be helpful for individual cases. Recommendations target selecting reviews, assessing risk of bias, qualitative and quantitative synthesis, and summarizing and assessing the body of evidence (Robinson et al., 2015). While using systematic reviews only as a source for included studies avoids the synthesis issue, it may not make the most of the existing evidence. The best way to use existing systematic reviews will depend on the identified reviews for the area of interest; a scoping review will be instrumental in determining the approach.

Search Strategy

This section outlines developing search strategies for electronic databases and search engines. It is useful to structure the components of the review question in a framework and to represent the association of the components in a Venn diagram, as illustrated in Figure 4.1. This step is useful even if an information specialist executes the search.

Figure 4.1. Search Venn Diagram



For each component, key terms and synonyms (e.g., *harm*, *adverse event*, *adverse effect*, *safety risk*, *side effect*) need to be identified. The scoping review (see Chapter Two) will be helpful in anticipating how researchers and database indexers may have labeled the topic of interest. Term preferences often vary by discipline, so it is useful to be creative when brainstorming terms. However, some terms are not useful because they may be relevant in too many contexts (e.g., *fall*, as in *hospital fall prevention* or *numbers expected to fall*). In addition, each term needs to be assessed for its incremental validity (e.g., *exposure level* does not increase the yield when the term *exposure* alone is already part of the search strategy). The search should operationalize the review question. However, some aspects of the inclusion criteria cannot be used in the search (e.g., the review may be interested in long-term effects, but the timing of the outcome measurement is often not reported in the title or abstract of the publication). In most databases, search strategies rely on the full citation, abstract, and key words to select relevant studies and, if an aspect is unlikely to be mentioned, "ANDing" it will exclude many relevant studies (i.e., studies that only reported on the timing of the outcome assessment in the full text of the study, studies that did not use the term *long-term* but specified the assessment date or number of months).

Search strings combine terms with Boolean logic (*AND*, *OR*, *NOT*) and several sets of search strings are needed for effective search strategies. As a general rule, *AND* will reduce the number of search hits because two terms are required to be present in the citation, while *OR* will increase the number of hits because one or the other term or both may be present. Two different components of the topic should be combined with the operator *AND*, while synonyms within a component should be "ORed." In some cases, it will be possible to "NOT" out a term (i.e., to exclude some terms or term combinations). Apart from standard operators, databases usually have wildcard options (e.g., a signal to automatically search all variations of a term [improv* retrieves *improve*, *improved*, and *improving*]); let users search for an exact phrase (e.g., *fall prevention*); or offer additional functions such as the *near function*, which specifies word proximity (e.g., within three words, to retrieve *toxin exposure* and *exposure to toxins*). Some researchers have published search strategies to identify studies specific to occupational safety and health (Verbeek et al., 2005).

Many databases carefully index citations and use controlled vocabulary to assist searches (e.g., Medical Subject Headings [MeSH] terms). For example, PubMed introduced a MeSH term for *occupational health* in 1991. However, this tag has been shown to be insufficient in identifying specific occupational safety and health research such as occupational health intervention studies (Verbeek and Van Dijk, 2006). A key limitation of relying on indexing is that it takes databases a while to fully index citations; hence, if a search strategy entirely relies on assigned vocabulary, it will miss the newest research on the topic. It should also be ascertained when the controlled vocabulary term was introduced; most databases will not go back and re-index older citations. Furthermore, relying on the controlled vocabulary assumes that indexers reliably and accurately tagged pertinent citations. This can be tested by reviewing the assigned vocabulary of known studies relevant to the review.

Databases also allow a variety of other limiters: for example, restricting to a period of publication years, to selected languages, or to sets for which search strings have already been developed, such as search filters for specific study designs. The impact of limiters varies considerably. For example, restricting to a specific (searchable) study design will greatly reduce the search yield, while restricting the search date to newer literature can be easily achieved but its effects will be less pronounced because of the speed of publishing in recent years. Throughout, it will require some thought in order to translate the review's inclusion criteria into the search strategy. For example, a review that wants to limit results to U.S. settings needs to explore how this can be operationalized (e.g., by using the first author's affiliation as a proxy). It should also be noted that some aspects of research studies are better indexed than others. For example, a number of search filters exist for randomized controlled trials, while searches for other study designs are less reliable. The InterTASC Information Specialists' Sub-Group has collated a large number of search filters for different databases that can assist systematic review searches.

Some databases are menu-based, user-friendly, and designed to assist multiple stakeholders, while others require detailed knowledge of the syntax and functionality of the search platform and database. Hence, depending on the platform (e.g., MEDLINE accessed through OVID rather than PubMed), database, and complexity of search strings, some searches are better executed by trained information specialists, such as librarians. In addition, the search strategy (the exact search strings) for each database needs to be documented in full in systematic reviews. Many journals require publishing the search strategy together with the systematic review, and the exact search strings are needed to update searches. In addition, the dates of the searches need to be recorded. Research continues to be published, and systematic reviews need to state until which date existing literature has been reviewed.

The purpose of developing a search strategy is to identify all pertinent citations. Searches can be evaluated by recall and retrieval rates (e.g., how many relevant studies are included in the search output, how many studies known to be relevant are included in the search output). There is a trade off between sensitivity (recall, or successfully identifying relevant studies) and specificity (precision, or successfully excluding irrelevant citations). An inclusive search strategy will identify more-relevant studies but at the expense of also including many irrelevant citations. The total yield of the search has to be a consideration for the review team (e.g., the yield may be too large for the available resources to screen it). The evaluation of the search strategy may balance total yield, recall, recall-to-yield ratio, precision, and its face validity (Hempel et al., 2011a). In some cases, it may be possible to use a multitiered approach that prioritizes parts of the search output, guided by the relevance of databases.

Citation Management Software

Given the continuously growing research literature, it is difficult to keep track of searches without reference-manager software. Newer software allows storing not just citations but also other material, such as websites. Most electronic databases or web services, such as Google Scholar, allow importing records into a software manager. In some cases, this requires detailed knowledge of import-filter options or the help of a professional librarian.

Importing into citation-management software automatically discards duplicates of citations indexed in more than one database. Programs are also designed to facilitate title- and abstract-inclusion screening. In addition, citation managers not only import the publicly available information of the citations (e.g., abstract, keywords), they also allow annotating citations (e.g., to document the origin of the citation or to document the inclusion screening decision).

Not only are literature reviews now characterized by an extensive and overwhelming amount of accessible research, systematic reviews also have to be able to account for the entire literature flow. This includes documenting the number of all deduplicated (i.e., excluding duplicates) initially identified citations, the number of studies obtained as full text, the number and citations of excluded studies together with the reason for exclusion, and the number and citations of included studies. In addition, having the citations in the citation software will facilitate

documentation of citations in the systematic review report and the final recommendation document.

Eligibility Criteria and Inclusion Screening

Inclusion criteria are a core element of systematic reviews. Inclusion criteria are explicit criteria that document the type of publication that will be included in the review. However, technically, these are eligibility criteria. Eligibility criteria define what publications need to report in order to be eligible for inclusion in the review. It needs to be made transparent which studies have been included in the review and also which studies were eligible for inclusion in the systematic review. This will allow the reader to evaluate the presence and the absence of evidence (i.e., the absence of evidence despite explicit attempts to find it).

Inclusion screening is typically done in two stages: at the title and abstract level and at the full-text level. At the title and abstract stage, literature reviewers need to decide whether the publication is likely to meet inclusion criteria (based on limited information). At the full-text inclusion–screening stage, typically two independent reviewers screen the publication against prespecified and explicit inclusion and exclusion criteria. Any disagreements are resolved by the review team. Dual review and resolving disagreements are performed to reduce (random) errors and (systematic) reviewer bias. Inclusion criteria have already been formulated at the protocol stage. During the review process, more details and decision rules for reviewers may need to be added to minimize ambiguity.

It is common practice in systematic reviews to document the reasons for exclusion of publications that were retrieved as full text. The reasons for exclusion should be summarized in a literature flow diagram (see Figure 4.2). In order to allow peer reviewers to determine whether all studies have been considered, a list of excluded citations can be added to the appendix of a systematic review together with the reason for exclusion. Ideally, the reasons for exclusion are stored together with the citation in a citation-management program.

For systematic reviews, it is very useful to use a framework such as PICO or Study design, Participants, Interventions, Outcomes (SPIO) to organize the inclusion criteria. Many reviews use extended versions (i.e., Population, Intervention, Comparator, Outcome, Timing, Setting, Study design [PICOTSS]). The NTP OHAT handbook (OHAT, 2015) uses a Population, Exposure, Comparator, Outcome (PECO) framework. It is crucial that systematic reviews formulate explicit and detailed inclusion and exclusion criteria at the review protocol stage. If there is no restriction (e.g., by setting), this should be made explicit. All dimensions relevant to the topic should be addressed:

- Population
  - For complex reviews, this can be a complex question (e.g., human participants or in vitro samples; safety initiatives may be directed at a number of different targets such

as workers, managers, and organizations], which complicates the determination of the study population).

- Intervention/Exposure
  - This is the independent variable in experiments; it can be an intervention or an exposure. The eligible intervention(s) and exposure(s) need to be described in detail. It may be necessary to define eligible interventions/exposures or list components that define the interventions/exposure (e.g., interventions using checklists), rather than using terms that could be interpreted differently by different stakeholders (e.g., *safety intervention*).
- Comparator (Study Design)
  - This dimension originates from systematic reviews of randomized controlled trials that all have a comparator, which determines the comparative-effectiveness estimates. For broader reviews, this dimension is sometimes used to describe eligible study designs. It should be made explicit whether concurrent (e.g., a control group in a prospective study) and historic comparators (e.g., pre-post comparisons) are eligible. When studies will be included regardless of the presence or absence of a comparator, this should be made explicit.
- Outcome
  - This should be a list of eligible measures (e.g., occupational exposure, worker behavior, occupational health or injury outcomes). Some outcome measures are only useful together with a denominator. The list of outcomes should be limited to outcomes that decide whether the study is of interest to the review (i.e., if the outcome is present the study will be included; if it is not present, the study will be excluded).
- Timing
  - Timing can refer to the measurement in relation to the exposure, the duration of the intervention or exposure, or the duration of follow-up (e.g., short- or long-term follow-up after the end of the intervention or exposure).
- Setting
  - This dimension should address all settings eligible for inclusion (e.g., interventions in different types of environments).
- Study Design
  - PICOTSS uses an additional dimension to clearly state the eligible study designs (e.g., case-control studies). In some cases, this may include analysis features (e.g., studies reporting a multivariate analysis).

In addition, the review may include other limiters, such as English-language publications or publications within a set period of time; these should be clearly communicated. Of note, programs such as Google Translate, which translate documents instantly, have changed the resources required to include non-English studies in systematic reviews and have made this

process more feasible. Often, the abstract of the publication will be sufficient to determine eligibility (Moher et al., 2000). However, not all PDF documents can be read by the program, and copying and pasting sections manually works best for Latin-based languages. The decision to include non-English studies should be based on the consideration of whether the topic is well represented in English-language publications or whether the topic is particularly popular in a non-English language country. Limiting the review to recent literature can make the literature findings more applicable to current working conditions, even when the effective savings in reviewable literature is limited (see discussion in previous section). Ideally, the eligible data timing should be linked to a benchmark, such as a legislative change or technological advance, rather than arbitrarily selecting a date limit for the review.

Wherever possible, inclusion and exclusion criteria should be formulated. This will further clarify the scope of the review. Conceptually, eligibility criteria should be understood as necessary information a publication needs to report in order to be of interest for the review. The data extraction may capture a number of study details that are necessary to understand the study, however, the inclusion criteria are the key variables that decide whether a study will be included or excluded from the review. The review may be designed to support a particular health topic, such as determining the effects on employees working the majority of a day at a computer screen, but the review inclusion criteria need to determine which participant samples and which exposure levels are relevant, apart from studies reporting on this exact population and condition. As discussed, the criteria should be sufficiently detailed to leave little room for interpretation across literature reviewers. For example, the review needs to determine whether mentioning an outcome is sufficient (e.g., burns associated with equipment) or whether the incidence has to be reported together with a denominator (e.g., burns per procedure, per year). Detailed eligibility criteria also will ensure that a reader of the review can follow the inclusion decisions.

Often, it is necessary to determine inclusion criteria for each key question individually. In some cases, it may be easier to write inclusion criteria for individual lines of evidence (e.g., separately for human-participant studies and for the type of mechanistic study eligible for inclusion in the review). The overall goal is to use a framework such as PECO to document as clearly as possible what publications need to report in order to be eligible for inclusion in the review.

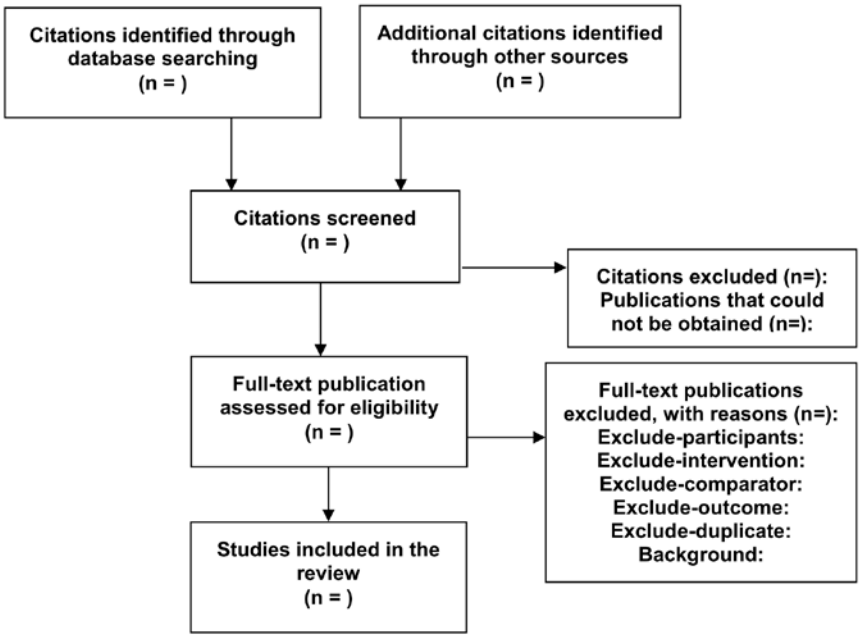
Furthermore, inclusion criteria are formulated at the protocol stage (i.e., before the review is conducted). While an iterative process is necessary to develop precise inclusion criteria, the process of deciding which study is relevant to answer the review question should be performed before the review is under way. Scoping searches or consultation with experts can provide example studies that will help clarify inclusion and exclusion criteria. While it is usually necessary to add more detail to the inclusion criteria while the review develops, explicit inclusion criteria are designed to be safeguards against including or excluding studies based on their results. Applying strict inclusion criteria aims to ensure that all relevant studies have been

reviewed, and what is relevant has been determined before the results of the included research studies were known to the reviewers.

Flow Diagram

Documentation of the application of eligibility criteria and the process of inclusion screening is shown in Figure 4.2. It documents the number of citations initially retrieved, the number of included publications at each step of the process, and the number of studies ultimately selected as relevant for the review.

Figure 4.2. Literature Flow Diagram Example



The last box shows the number of studies included in the review. Since study results may have been reported in more than one publication, studies should be identified by the included participants or subjects and not the publication. In some cases, this requires some detective work to determine which population is included in more than one study; original study authors do not always reference the parent study clearly.

As discussed earlier, it is important to disclose the information sources and the inclusion criteria so that the reader can put the presence and the absence of research in context (by knowing what was explicitly searched for and where). Cochrane reviews, for example, always publish a list of excluded studies so that the reader can see which publications have been reviewed as full text, even if they have not met inclusion criteria.

5. Systematic Review Step 4: Document and Assess Included Studies

This chapter addresses two steps in the documentation and assessment of individual studies: data extraction to populate an evidence table and critical appraisal of the included studies. The final section addresses data access requirements for systematic reviews.

Data Extraction

Data extraction (i.e., collecting key information from relevant studies) is a core step of systematic reviews. Data extraction should be performed systematically by collecting the same information from each study to facilitate a structured overview of the existing evidence. It should concentrate on key aspects of the included studies to facilitate presentation of a clear overview. The main variables should already have been listed in the protocol.

Data extraction (or data abstraction) should aim to abstract, rather than to copy information. This involves categorizing as much information as possible without losing key details necessary to describe the study. The data-extraction form should be designed with the evidence table (see Table 5.1) or another final product in mind. This will showcase how little information can be presented in evidence tables when trying to provide a concise overview. The data-extraction form may include some variables that will only be used for analytic purposes, in particular if the topic is amenable to meta-analysis. However, data extraction should only cover data that will actually be used, given that the extraction process is labor intensive and prone to errors (which requires additional steps to ensure accuracy). Data extraction in systematic reviews is a means to an end; it is not a final product. The data-extraction form and the evidence table should be designed with the review questions in mind to ensure that all relevant information is shown.

Data extraction is best performed by using a standardized form in a database. Many programs are available to do this, both specific and nonspecific to systematic reviews. Systematic review–specific software, such as Covidence, is set up to support the entire review process, not just the data extraction (e.g., managing literature flow). Online software is useful for geographically dispersed review teams and avoids having to keep track of master versions. Some data-extraction software allows the publication of data after completion of the review (see the section on data access in this chapter). A range of systematic review products is available; some are specific to systematic reviews of hazard and risk assessments, such as the Dragon software. Deciding which program to adopt should consider costs and functionality, including data input and export functionality. Data-export considerations have to do with how easy it is to move the extracted data from the data-extraction program into another program (e.g., a Microsoft Word document for the report), and how flexible the program is in supporting custom-specific formats for the

evidence table. The availability of technical support as well as data security and archive options are other considerations that will help determine the choice of a particular program.

It is useful for categorical data with known categories and information that can be translated into a predesigned rating scale to undergo dual review by two independent reviewers in order to avoid reviewer errors and bias. Information in published papers can be ambiguous and sometimes hard to find. Any disagreements should be resolved through discussion within the review team. For free-text information, the review team needs to determine an approach to resolving inconsequential disagreements (i.e., resulting from a different word order), since data-extraction software will mark all differences as discrepancies. For free-text information, it is often advisable to use an abstractor-checker model, with one reviewer abstracting the data and one reviewer checking the information and initiating discussion about disagreements if necessary.

It is essential to pilot the data-extraction tool with actual studies that meet inclusion criteria for the review. In addition, it is crucial for teams to have detailed instructions and operational definitions. These serve to keep reviewer differences in interpretation of the variables (e.g., number of participants included in the study versus included in the analysis), differences in extracted detail (e.g., point estimate together with confidence intervals or *p* values), and differences in extracted information (e.g., when deciding which result is important) to a minimum. A mock-up of an evidence table for the first extracted studies ensures that the data is represented correctly and will showcase any needed format changes (e.g., reviewers extracting lengthy text).

The data extraction is primarily based on published data; however, in some cases, information may be unclear or data of interest may exist but were not reported or only partially reported. The IOM systematic review recommendations include inviting study authors to clarify information about study eligibility, study characteristics, and risk of bias (Eden et al., 2011). In contrast, this is not commonly practiced among EPCs (Lau et al., 2013). Contacting authors can be a resource-intense and lengthy process, and review authors need to decide on the method (Young and Hopewell, 2011) and intensity of efforts. For example, an empirical study that investigated the response rate of authors of diagnostic-accuracy studies found that authors who provide data did so by the third emailed request, and further efforts were not productive (Selph, Ginsburg, and Chou, 2014).

The Guide to Community Preventive Services provides guidance for data extraction (Zaza et al., 2000). An example of a complex data extraction form is available online. The AHRQ Training Module for the Systematic Reviews Methods Guide on data extraction provides practical hints and tips. The NTP OHAT systematic review handbook provides a list of data-extraction elements for human, animal, and in vitro studies (OHAT, 2015).

Evidence Table

An evidence table that summarizes basic information about the included studies and allows a structured overview is a key component of systematic reviews. Evidence tables are organized by included studies with one row per study.

The function of evidence tables is to document study details and results so that the reader does not need to depend entirely on the review authors' conclusions but has more information on each of the included studies. For the review team, tabulated information in evidence tables helps to abstract information and to order and structure findings. Evidence tables force the information into a standardized overview. Having the data shown this way is very useful for objectively reviewing study results, regardless of how the authors of the original studies presented the data. Where possible, data should be presented to facilitate comparisons across studies (e.g., by converting response rates to percentages rather than reporting raw data). Evidence tables and summary of findings tables (see Chapter Six) should have the review question in mind and the recommendation it is aiming to support. Depending on the complexity of the risk of bias assessment, the results of the assessment may either be reported in a separate table or incorporated, at least in summary format, in the evidence table.

As discussed in the previous section of this chapter, there is a trade-off between presenting details and allowing a concise overview. Generally, there is limited space available. When designing the evidence table, Section 508 Standards compliance (i.e., U.S. federal agency requirements to provide website accessibility compatible with assistive technology) also should be kept in mind (e.g., using a basic grid and avoiding nested cells so that the systematic review report can be easily posted online). Any methodological characteristic that is key to evaluating the validity of the study should be shown, such as the reliability of the data source or the method of ascertainment of exposure. Similarly, variables necessary to understand the results of the study (e.g., reported power calculation indicating sufficient power to detect an effect) also should be incorporated in the table. The credibility of the conclusions of the literature review depends to a large extent on the transparency of the methods and whether the reader can follow how the conclusions were drawn from the identified evidence.

To enable a useful overview, the items that should be presented in the evidence table are topic-specific. However, PICOTSS should provide a general framework for selecting variables. Table 5.1 is an example of an evidence table.

Table 5.1. Example of an Evidence Table

Study Details	Participants	Intervention/Exposure	Results
Author, Year: Country: Study Design: Funding:	N: Characteristics: Inclusion Criteria: Exclusion Criteria:	Hazard: Concentrations: Critical effect levels: Duration of exposure: Type of exposure:	Outcome measure 1, follow up: Outcome measure 2, follow up:
Author, Year: Country: Study Design: Funding:	N: Characteristics: Inclusion Criteria: Exclusion Criteria:	Hazard: Concentrations: Critical effect levels: Duration of exposure: Type of exposure:	Outcome measure 1, follow up: Outcome measure 2, follow up:

Note: N = number of participants, subjects, or other study targets

Critical Appraisal

Another important component of systematic reviews is the critical appraisal of the individual studies included in the review. While all included studies are considered in the review, their validity is often assessed individually. Critical-appraisal results can be taken into account when evaluating the evidence base overall, can be explored as factors that may potentially explain differences in results across studies, can be used for sensitivity analyses (e.g., testing the robustness of the finding by omitting poor quality studies), or can be used to select the best evidence (i.e., an additional inclusion criterion). All but the first approach require a sufficient number of identified studies, and, often, this may not be feasible because of the lack of empirical studies of interest.

Critical-appraisal assessments are time consuming and require detailed scrutiny of the published papers. In addition, even well-known tools, such as the Cochrane risk of bias tool, have limited inter-rater reliability characteristics (Hartling et al., 2013). Furthermore, the reliability and validity of other tools, in particular tools newly developed as part of the review process, may not be known. Generally, the assessment should be limited to a key instrument or key dimensions, and scoring rules should be reported together with the results of the assessment to increase transparency.

Critical appraisal may focus on the internal and external validity of the study.

Internal Validity

Internal validity refers to the conduct, internal logic, and methodological rigor of a research study and assesses the extent to which an observed effect seen in the study can be attributed to the intervention or exposure being evaluated. This is especially important if we suspect that study characteristics may systematically affect the findings of the study (e.g., the size of effect). Bias is a systematic deviation from the true effect. Risk of bias assessment is a standard element of systematic reviews. A large number of potentially relevant study characteristics, checklists, and assessment scales have been suggested in the literature. Nonetheless, there are few empirical

associations between study characteristics and reported effect estimates that have been conclusively determined (Hempel, Suttorp, et al., 2011, and Hempel, Miles, et al., 2012).

Different study designs minimize different types of bias, and the different study designs of included studies already will provide a broad differentiation of studies. The Centre for Review and Dissemination's guidance for undertaking reviews in health care provides a collation of study designs (2009). However, study design alone should not be used to assess the published evidence. For example, high-quality observational studies may be much more useful than low-quality randomized controlled trials. Risk of bias assessments are usually study-design specific (e.g., the Cochrane Risk of Bias tool applies to randomized controlled trials). When different study designs are eligible for inclusion in a review, the review team may either use different tools or assess common sources of bias across all studies. The Cochrane handbook considers *selection bias*, *performance bias*, *attrition bias*, *detection bias*, and *reporting bias* to be core sources of bias in research publications (Higgins and Green, 2011):

- Selection bias
  - *Selection bias* addresses systematic differences in baseline characteristics between compared groups (e.g., through self-selection of the interventions or exposure). Random allocation to intervention groups and allocation concealment in controlled trials, as well as matching samples in key characteristics in observational studies aim to avoid this bias.
- Performance bias
  - This bias describes systematic differences between compared groups other than the intervention or exposure that originates in the knowledge of the study condition (e.g., unintended co-interventions or differential behavior through study personnel, Hawthorne effect in participants). Blinding of study personnel aims to reduce this bias; behavioral responses to the knowledge of being part of a research study can sometimes be addressed through covert observational research.
- Attrition bias
  - *Attrition bias* addresses systematic differences in the loss of participants (e.g., incomplete follow-up and differential attrition in the study and control groups); drop-offs may be systematically different from study completers. Intention to treat analyses aim to reduce the risk.
- Detection bias

  This bias describes systematic differences in outcomes assessment between compared groups; the knowledge of the prior exposure or participation in the intervention may influence the measurement. Blinding of outcome assessors aims to reduce the risk.
- Reporting bias
  - *Reporting bias* addresses systematic differences between reported and unreported findings (e.g., selective outcome reporting by selecting measures with positive

results). A priori–reported study protocols that can be reviewed together with the results of the study aim to avoid this bias.

The outlined sources of risk of bias have most frequently been addressed in randomized controlled trial research, and review teams need to decide which study characteristics are likely to affect the study results for each individual review topic. In observational studies (i.e., studies in which the intervention or exposure was not assigned by the study investigator but observed in self-selected cohorts and case-control studies), the risk of confounders needs to be critically reviewed (e.g., were known confounding variables controlled in the analysis, can the study results be explained by confounders). Critical appraisal tools for observational studies such as the Newcastle-Ottawa Scale focus on the selection of cohorts and cases and controls (*selection bias*); the comparability of cohorts, cases, and controls (*selection bias*, *performance bias*); and the assessment of outcomes in cohorts and the ascertainment of exposure in case-control studies (*detection bias*). Principles, such as outcome-assessor blinding (*detection bias*), are universally applicable and almost always possible even when participant blinding is not.

Risk of bias assessment is performed using the published information in the original research article unless original authors are contacted and a response is received. While the quality of the reporting and the quality of the methodology are not identical, reporting guidelines can help identify key study characteristics, meaning those that should have been reported in detail and that are key to evaluate the study and its findings. The EQUATOR network is a resource for reporting guidelines. Reporting guidelines have been published for a large number of research study designs (e.g., STROBE for observational studies). Reporting guidelines help to identify key elements for the particular study design or topic areas and may therefore be instrumental in designing the data-extraction form and evidence tables (see Table 5.1).

Living documents such as *The Cochrane Handbook for Systematic Reviews of Interventions* and the Comparative Effective Rules Methods Guide provide resources for critical appraisal for a broad range of study types.

External Validity

External validity concerns the generalizability of results shown in individual study samples. It addresses the conclusions that can be drawn from the study and how applicable the findings are to other settings and conditions. External validity can be a particular concern for experimental studies, for example, due to strict inclusion criteria restricting the study population and the artificial and highly controlled environment in which the intervention or exposure was tested. Reporting guidelines for randomized controlled trials such as the CONsolidated Standards of Reporting Trials (CONSORT) Statement recommend that authors discuss the generalizability (referring to external validity and applicability) of the findings.

In the context of occupational safety and health systematic reviews, external validity may also refer to the degree of applicability of study results to the systematic review questions. The

included studies have already been selected as relevant to the systematic review questions outlined in the analytic framework and eligibility criteria. Nonetheless, studies may differ in their extent of external validity. The external validity is particularly important when reviewing different lines of evidence. For example, the external validity assessment may differentiate the relevance of human participant health data and results of in vitro test data, and their applicability and contribution to answering the occupational-exposure review questions. Furthermore, studies may vary in how easily results can be extrapolated from participants in the research study to workers in occupational settings.

Some published assessment tools that can be used to assess the external validity of research studies are available (Burchett, Umoquit, Dobrow, 2011, and Dyrvig et al., 2014). However, in most cases, this validity assessment has to be determined by the review team, and tailored toward the review questions. The external validity assessment may follow a population, intervention/exposure, comparator, outcome, timing, setting (PI/ECOTS) framework, addressing all or selected criteria. Considerations regarding the *population* should be focused on the representativeness of the study sample as well as the similarity of the study population to the population of interest in the review or recommendation document (e.g., people versus rats; study participants versus members of the workforce). The *intervention or exposure* can be assessed with regard to applicability and similarity to conditions observed outside of the research study (e.g., taking realistic occupational-exposure concentration and the route of exposure into account) and the *comparator* can be similarly evaluated. The external validity of the *outcome* focuses on how closely linked the study outcome is to the outcome of interest (see discussion about intermediate outcomes in the "Analytic Framework" section in Chapter Three). The assessment of the external validity of the *timing* may assess the follow-up points in research studies (e.g., reporting on short-term effects while the review or recommendation is concerned with long-term exposure). The *setting* refers to the comparability of context conditions, context-specificity of the results, and transferability of findings to other settings (e.g., transferability of results across different countries and labor markets). The context of the study may also be critically reviewed, if not already addressed, in the other criteria. For example, assessing whether observational studies published 30 years ago are still relevant and applicable to the current work environment.

The external validity does not need to be assessed within individual study design categories, but the selected criteria can be applied to all identified studies. Furthermore, where there is little variation in the assessment across individual studies (e.g., all in vitro studies may have the same rating), it may be more useful to apply the external validity assessment to the body of evidence. In that case, the external validity will not be judged for each individual study but for groups of studies or lines of evidence and their contributions to answering the review question (see "Evidence Integration" in Chapter Six). The body of evidence evaluation typically addresses the external validity of included studies in the domain *Indirectness* (see Chapter Six). Furthermore, one aspect of external validity, the robustness of findings across studies, is always assessed at the

body of evidence level (see "Consistency" in section "Criteria to Evaluate a Body of Evidence" in Chapter Six). This measure of generalizability assesses the independence of study results from the individual study context (Green and Glasgow, 2006).

Other Assessment Criteria

Other considerations for critical appraisal of individual studies are the credibility of the source, conflicts of interest of the study authors, and the role of the funding source that supported the research. In many instances, there may be a vested interest associated with the authors or sponsors of the research study. While industry sponsorship is not a source of bias per se, the funding source information should be reported together with the study.

Critical Appraisal Adaptations for Different Lines of Evidence

The Navigation Guide systematic review methodology for risk of bias assessment is based on the Cochrane Risk of Bias tool, but modifications were made to be appropriate for human observational studies and animal toxicological studies (Johnson et al., 2014). It differentiates *low*, *probably low*, *probably high*, and *high risk of bias* and applies risk of bias criteria to human and nonhuman studies separately. For human studies, these include (1) recruitment strategy, (2) blinding, (3) exposure assessment, (4) confounding, (5) incomplete outcome data, (6) selective outcome reporting, (7) conflict of interest, and (8) other bias. For nonhuman studies, these include (1) sequence generation, (2) allocation concealment, (3) blinding, (4) incomplete outcome data, (5) selective reporting, (6) conflict of interest, and (7) other bias. Guidance is provided for each level of risk for each criterion.

The NTP OHAT systematic review handbook (OHAT, 2015) differentiates *definitely low*, *probably low*, *probably high*, and *definitely high risk of bias*. It provides a table with different types of bias relevant to experimental studies (human or animal) and observational studies. Finally, the handbook includes a risk-of-bias tool that lists the bias criteria and specific assessment questions together with information of the applicability to experimental animal trials, human-controlled trials, cohort studies, case-control studies, cross-sectional studies, and case series.

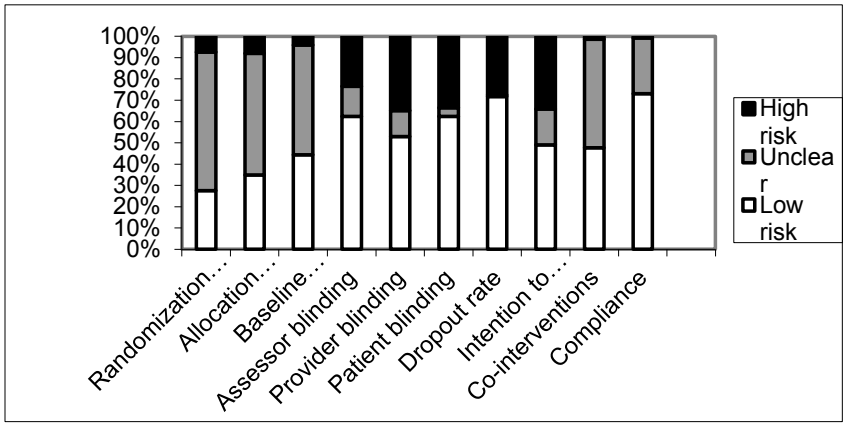
Documentation of Critical Appraisal

Critical appraisal tools have different formats, and there is little consensus about how to summarize the critical appraisal results across studies. Some tools use the format of a checklist, others a scale, and the criteria as well as the conceptual approach to critical appraisal differs (Hempel et al., 2011a). In a checklist, each criterion is scored independently and conceptually; the criteria are not necessarily related (e.g., whether matching of the study cohorts was successful may not be related to the outcome ascertainment quality). Scales assume that the individual criteria (items) assess and contribute to an underlying concept (e.g., study quality), and the total score-scale score is the key unit of critical appraisal judgment. *The Cochrane*

Handbook for Systematic Reviews of Interventions advocates deriving a summary judgment not based on mean or sum of the individual assessed criteria, but based on an overall judgment of the study for a particular outcome (Higgins and Green, 2011). The differentiation in the answer mode also varies, with some tools assessing a continuum (e.g., a Likert scale to grade the extent of suspected bias for the criterion) and others using a dichotomous answer mode (e.g., validity criterion met or not met). Tools also differ in how they handle uncertainty, for example, whether they allow reviewers to indicate that the risk is unknown. The critical appraisal results should be documented in the format in which they were analyzed and contributed to the systematic review (e.g., an overall internal validity score per study may have been used to differentiate studies).

Documenting critical appraisal may use the format of reporting results for individual studies in a table. Alternatively, the critical appraisal assessment can be documented across all included studies. Figure 5.1 is an example showing the proportion of studies with a *high risk of bias*, *unclear risk of bias*, or *low risk of bias* for ten selected risk-of-bias criteria.

Figure 5.1. Study-Level Critical Appraisal Summary Example



Results of key critical appraisal criteria (or a summary across criteria) may be added to the evidence table for each study. However, the results for all assessed and contributing criteria (all criteria that influence the decision about the internal or external validity of the study) should be presented either in a separate table or in a graph to increase transparency in the review process.

Critical appraisal tables may either simply report the result of the reviewer's rating (e.g., *low risk of bias*) or present the rating together with support of the individual assessment (e.g., criterion *confounding* rated *low risk of bias*, with the supporting evidence: "The authors adjusted for age and gender in the analysis"). The Navigation Guide systematic review methodology group has published an example of a detailed documentation as an appendix to Johnson et al., 2014.

Increasingly, authors use color or shading to provide an overview of the assessment results, for example, green to indicate a low risk of bias for the criterion and red indicating a high risk of bias. The risk-of-bias chapter of *The Cochrane Handbook for Systematic Reviews of*

Interventions illustrates different ways to show data, such as an example of a summary of the risk-of-bias assessment across studies that presents the percentage of studies judged to have *high*, *unclear*, or *low risk of bias* (Higgins and Green, 2011). An example of the Navigation Guide approach to color-coded risk-of-bias display is given in a case study on effects of exposure to perfluorooctanoic acid on fetal growth (Johnson et al., 2014).

Generally, the presentation should match the available information. A simple color-coding system is an easily accessible format for stakeholders; however, critical appraisal ratings tend to oversimplify the complexity of the research study and its appraisal. In addition, there is an element of subjectivity in the rating as well as the interpretation of the presented information in the original study publications, and the reliability of critical appraisal assessments is limited (Hartling et al., 2013, and Hempel et al., 2015b). While systematic reviews can report the agreement between two independent raters to communicate subjectivity, the standard approach is to minimize errors and bias through dual rating and reconciling disagreements between reviewers. As outlined at the beginning of the section, the criteria and the scoring instructions should be presented to increase transparency for the reader of the systematic review.

Data Access

Increasingly, scientific journals and government agencies require that the data used in research studies are made accessible to the reader. This requirement also applies to systematic reviews, and readers can request access to the data that were collected about the included studies. Data may be made available through detailed online appendices, access to databases, or data access by request to the author. This trend is not limited to academic research published in journal articles. NIOSH has established a gateway to access data produced and used by them. EFSA now enables access to data, methods, and results, as well as relevant supplementary information (EFSA, 2015). The EPA maintains a searchable database that includes documents that have been cited in EPA risk assessments and which can be viewed publically. AHRQ has established a data repository for systematic reviews to improve access to data by consumers of evidence reviews and to promote transparency and reliability in the systematic review process. The repository includes study-level data for all included studies included in a systematic review.

Another function of data repositories is to enable other researchers to access the data for future projects. The ARHQ-maintained repository can be used for data extraction and, after project completion, the data are made publically available. The EPA also encourages standardized data extractions that are meant to be application-independent (see examples here: database 1, database 2, database 3). The time savings from using already-extracted data depend on the similarity of the review question being addressed: Different systematic reviews may need different information from the available research.

6. Systematic Review Step 5: Evaluate and Interpret the Body of Evidence

Given the multidisciplinary nature of occupational safety and health research and the complexity of recommendations issued by agencies such as NIOSH, this chapter distinguishes among synthesizing evidence, evidence grading, and evidence integration. These steps refer to the assessment of the body of evidence (i.e., all studies meeting inclusion criteria of the systematic review). There is disagreement in the definition and use of terms in the existing literature. In this chapter, synthesizing evidence refers to organizing evidence into homogeneous categories and summarizing results across identified studies within these categories. Grading evidence refers to assessing the quality of the body of evidence to communicate the confidence in the evidence summary. Integrating evidence and weighing of evidence refers to evaluating evidence across categories and lines of evidence.

Synthesizing Evidence

Evidence synthesis involves stratifying the types of evidence identified in the review. The Project on Framework for Rating Evidence in Public Health (PRECEPT) suggests grouping evidence first based on four general domains (disease burden/level of contamination, risk factors, diagnostics, and interventions questions) (Harder et al., 2015). Within domains, systematic guidance specific to the domains or individual study designs will be helpful to guide the synthesis. Examples are guidance for systematic reviews targeting observational epidemiological studies reporting prevalence and incidence data (Munn et al., 2015), the handbook of *The Cochrane Screening and Diagnostic Tests Methods Group*, or the EPC methods guide for comparative effectiveness for reviews for interventions (AHRQ, 2014).

Within domains, summarizing evidence within individual lines of evidence (e.g., in vitro, human, and nonhuman studies or lines of arguments) is a further key organizational option. Evaluating identified research evidence within lines (*realms*) of evidence is consistent with best practices for weight of evidence analyses (Rhomberg et al., 2013).

In the synthesis step, study results are summarized across identified studies, within general domains (e.g., diagnostic accuracy), and within lines of evidence (e.g., in vitro studies). This process assumes that there are multiple research studies on the topic of interest within the domain and within the line of evidence. Furthermore, the individual results of studies are (typically) not identical and instead may vary, potentially contradict one another, and are in need of a summary across studies.

There are different ways to summarize research studies in a literature review. Rather than describing one identified study after the other in a narrative, a useful approach is to summarize

results by outcome, i.e., data on key measures of interest. The summary should compare and contrast research findings for individual outcomes (or groups of outcomes) across studies.

Meta-Analysis

The summary of results reported in individual studies does not need to be based on a meta-analysis, although that is a key approach to summarizing results across studies. Meta-analysis is a statistical technique to pool study results across studies in order to derive a reliable summary estimate. Meta-analysis does not provide a simple mean across study results for an outcome. Instead, study results are weighted by study precision, i.e., studies with more-precise estimates (usually larger studies) contribute more to the summary. In addition, aggregating data across studies increases statistical power. Meta-analysis can find patterns and evidence of small effects across studies, even when individual studies are too small to show statistically significant associations.

Several meta-analysis programs are available for free and do not require specialist statistical training. The Cochrane Software Review Manager (RevMan) is available online for a range of basic meta-analytic techniques, and the user will be prompted by a program wizard. It can be used, for example, to compute rates or the mean and standard deviation for two groups across studies to arrive at a summary estimate. It computes a range of effect sizes from count data or continuous data such as the relative risk, risk difference, odds ratio, hazard ratio, weighted mean difference, or standardized mean difference to compare the size of effects across different scales. Studies will be weighted automatically by study size, but weights also can be assigned on the basis of other dimensions, such as the risk of bias associated with each study. Finally, resulting forest plots provide a visual overview of individual studies and their summaries.

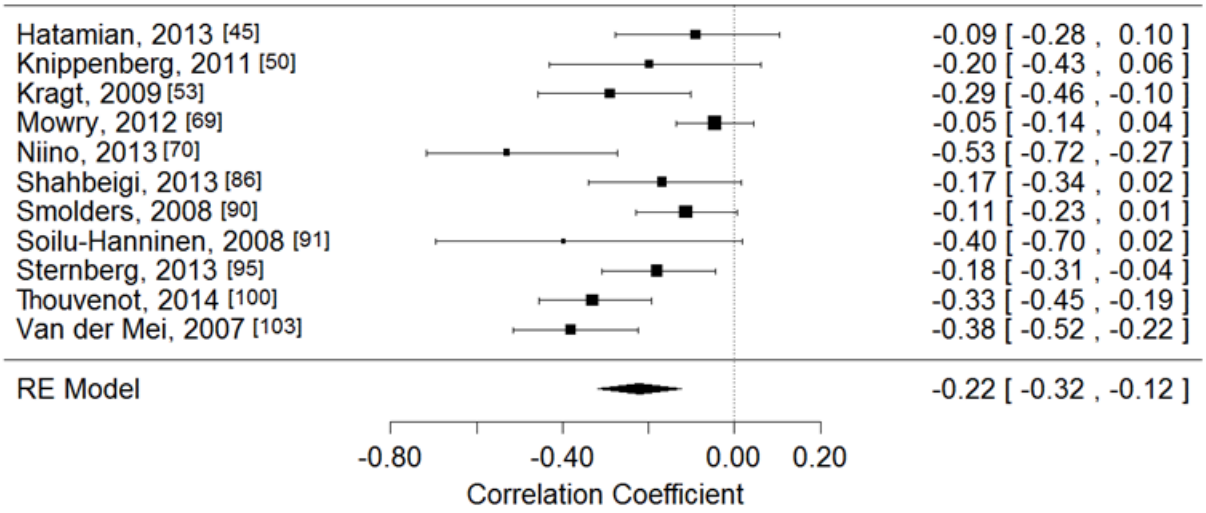
Online platforms are available that provide a variety of meta-analytic tools, and open-source programs such as R allow a wide range of complex and innovative analyses in packages such as meta and metafor in addition to standard commercially available statistics packages.

Forest Plots

Forest plots visualize a range of key aspects of studies and the available evidence base. Of note, forest plots can be useful even when no summary estimate has been computed across studies; in some instances, it may be inappropriate to pool across study results because the studies are too diverse and a summary estimate is void of meaning (i.e., comparing apples and oranges) or the average effect of the exposure or the intervention is not informative. Forest plots graphically show the included studies by illustrating a study identifier together with the point estimate and a measure of dispersion (e.g., the confidence interval around the effect), depict the size of the study and its weight in the summary estimate, visualize study effects on a common metric (e.g., the x-axis showing the mean difference between two groups), and may include numerical study results allowing a clear comparison across studies. Forest plots graphically show the direction of effects, show the consistency of results across studies, visualize outliers, and depict the

individual and summary-effect estimates relative to the line of no effect. Figure 6.1 is an example of a forest plot that shows correlations (i.e., a measure of the strength of association of two variables) reported in individual studies (the example includes 11 individual studies) and pooled across studies (e.g., the flat diamond at the bottom of the figure).

Figure 6.1. Forest Plot Example



Visualizing study results in figures can make data more accessible to different stakeholders, in particular nonresearchers. Graphic displays are not limited to forest plots and may encompass simpler data displays, such as a histogram plotting the number of adverse events in exposure studies. Studies ordered by exposure level and the associated response shown in exposure-response arrays is another method to structure data across studies.

Other Approaches to Summarize Results Across Studies

Other ways to numerically summarize results across studies include the reporting of the range of study results. Computing the mean or the median of the metric reported in the individual studies will determine the average effect across studies.

Where the metric differs across studies, studies may need to be stratified by overall result. Vote counting summarizes the direction of effect across studies independent from the individual metrics reported in the studies. This may, for example, count what proportion of studies show an association across all identified studies or how many studies favor the intervention for an outcome of interest. This approach will provide a general overview and show the consistency of study results across all identified studies. Vote counting is complicated by the fact that many studies may not have had sufficient statistical power to detect an effect, so it is difficult to determine the threshold for a "positive" study result. A sign test—a nonparametric test to determine whether two groups are equally sized—can determine the statistical significance of the

direction of effects across studies (e.g., statistically significantly more positive than neutral or negative results). However, in addition to the difficulty of determining what counts as a positive study result, each study contributes an equal amount to a simple sign test, which may not be appropriate if studies differ in size, methodological rigor, or external validity and applicability.

Summary of Findings Table

A useful tool for evidence synthesis is a summary of findings table (Table 6.1). *The Cochrane Handbook of Systematic Reviews* introduced summary of findings tables in 2008 (Higgins and Green, 2008). The table is organized by outcomes and stratified by review question and intervention or exposure. Summary of findings tables are based on evidence profiles suggested by the GRADE group, and the GRADEpro or the GRADEpro Guideline Development Tool (GDT) program may be used to generate summary of findings tables in the Cochrane and GRADE format. However, the tables are easily constructed by hand, which allows more flexibility, for example, for incorporating the preferred format for summary tables advocated by the AHRQ EPC program (Berkman et al., 2013). The EPC Methods Guide suggests paraphrasing the findings in a sentence and indicating the direction of effect, in addition to describing any numerical results or simply stating the magnitude of effect, to help readers understand the findings of the review.

Generally, the table should include the study design contributing to the evidence base and the number of identified studies reporting on the specific outcome of interest. It is often also useful to report the number of participants (or subjects) the studies are based on, in particular when sample sizes are small and results are only based on a handful of participants. The table should summarize the results, ideally both in terms of the direction of effect and the size of the effect. The table is organized by outcomes. Cochrane and Campbell Collaboration review authors are asked to select a short list of outcomes that will be documented in the table at the review-protocol stage, adding emphasis to selecting key outcomes that determine whether an intervention is effective or an exposure should be considered harmful.

Table 6.1 shows an example of a summary of findings table. It also includes dimensions used to grade the evidence, which will be discussed in the next section.

Table 6.1. Example of a Summary-of-Findings Table

Outcome	Study Design: Number of Studies, Number of Subjects	Findings: Direction and Magnitude of Effect	Study Limitations	Inconsistency	Indirectness	Imprecision	Publication Bias	GRADE
Key question #1: line of evidence (e.g., exposure, human data studies)								
Outcome 1 (e.g., days absent)	e.g., X before-after studies (XX participants)	e.g., mean reduction X (CI X, X) post-intervention compared with pre-intervention						
Outcome 2 (e.g., incidence rate of X)	e.g., X cohort studies (XX participants)	e.g., pooled RR XX (CI X, X), risk increased with exposure						
Outcome 3 (e.g., risk of developing X)	e.g., X cross-sectional studies (XX participants)	e.g., X studies report an association, X do not						
Interventions studies, human participants								
Outcome 1 (e.g., mean stress scores)	e.g., X controlled trials (XX participants)	e.g., difference in mean stress score X (CI X, X), favors intervention compared with control						

NOTE: CI = confidence interval, RR = relative risk.

This approach to summarizing studies is different from describing one included study after another in a narrative synthesis. It keeps one or more variables constant, most importantly, the outcome that is being evaluated. The approach aims to facilitate comparing and contrasting findings across studies. Highlighting the consistency or conflicting results across studies and documenting the number of available studies to inform the question of interest helps to provide a clear overview of the existing evidence and also will be useful in developing recommendations (see Chapter Seven).

Grading Evidence

Grading evidence is the process by which criteria are used to evaluate the quality of a body of evidence. This should take all identified studies into account, hence, it is an assessment across studies and different from the critical appraisal of individual studies described in the "Critical Appraisal" section in Chapter Five. This is usually done at the outcome level (i.e., for a particular outcome of interest that was measured in multiple studies) so that concrete evidence statements can be made and the confidence in the statement can be graded. This section differentiates criteria to evaluate a body of evidence, the starting point for evaluations, and the level of quality of evidence.

Criteria to Evaluate a Body of Evidence

A 2002 systematic review of systems to rate the strength of scientific evidence (West et al., 2002) found a small number of published systems available at the time but also helped to identify dimensions of rating systems. Systems often consider the *quality* of the included research—such as the study design, conduct, analysis, or methodological rigor that the majority of included studies is based on. The included studies may be judged based on level of evidence hierarchies that primarily take the study design into account (Briss et al., 2000, and Baker et al., 2010) or reflect the methodological conduct of the individual study independent from the study design. A second common dimension is the *quantity* applied to the relationship between the exposure and the outcome, as well as the amount of research available; this may cover the magnitude of effect, the number of studies available to contribute to the research question, and the number of individuals studied in the evaluations. A third dimension, *consistency*, is the degree to which a body of scientific evidence is in agreement across studies and how much conflicting evidence has been identified.

GRADE Criteria

The GRADE working group published a series of publications and online resources on grading evidence. The primary achievement of the group was to move away from simple classifications of evidence by study design. Instead, the GRADE working group uses eight specific criteria to assess a body of evidence. Five are used to downgrade the quality of evidence if serious concerns are detected: (1) risk of bias (study limitations), (2) inconsistency, (3) indirectness, (4) imprecision, and (5) publication bias. Three criteria can be used to upgrade the quality of evidence: (1) large effect, (2) dose-response relationship, and (3) all plausible residual confounding (Balshem et al., 2011). The criteria are as follows:

- Risk of bias (study limitations)
  - The *risk of bias* domain takes the study limitations of all identified studies contributing to the result into account (see "Critical Appraisal" section in Chapter Five).
- Inconsistency
  - The dimension evaluates the variation in effect estimates across studies and the amount of heterogeneity (unexplained variance across studies in meta-analyses). In summaries without meta-analysis, the consistency of the direction of effect and differences in effect estimates need to be judged without quantifying the effect. Where confidence intervals can be calculated for individual studies, minimal or no overlap of confidence intervals will indicate *inconsistency*. In some cases, it may be possible to identify the source of inconsistency (e.g., the effect size varies by exposure dose). In that case, separate evidence statements for the different scenarios may be warranted.

- Indirectness
  - Evidence can be indirect because study participants may differ from those of interest in the review, the intervention/exposure tested may differ from the intervention/exposure of interest, outcomes may differ from those of primary interest (e.g., surrogate measures), and interventions/exposure types may have not been tested in head-to-head comparisons (but were compared indirectly across studies). This domain requires judgment to determine whether the degree of *indirectness* across relevant aspects warrants lowering the confidence of an effect (Schünemann et al., 2013).
- Imprecision
  - *Imprecision* takes the confidence interval or standard error around the point estimate reported in a study into account. Small studies will report large confidence intervals, i.e., a large range or interval in which the true value may fall. Meta-analysis across studies increases precision of homogenous studies, and the width of the confidence interval can show the extent of imprecision. Precision is often judged in combination with the proximity of the effect estimate to the point of no effect (e.g., if the upper and lower confidence interval limits include a suggested benefit as well as no benefit of an intervention). Judging imprecision across studies without meta-analysis is challenging and needs to take into account the precision reported in individual studies and the sample sizes of individual studies. In some cases, it may be possible to calculate effect sizes and confidence intervals, even when they were not reported in the original studies. Unfortunately, many publications outside of trial literature do not report measures of dispersion together with point estimates.
- Publication bias
  - This reporting bias assesses whether there is evidence that pertinent studies are missing from the identified evidence base. *Publication bias* occurs when the decision to publish in a scientific journal is not independent from the result of the study. Without formal tests in meta-analytic models, the presence of or absence of publication bias will be difficult to determine. However, one protection against publication bias is to consider unpublished data or grey literature (see Chapter Four), in addition to data published in peer-reviewed journals.
- Large effect
  - The quality of a body of evidence can be upgraded for a *large effect*, i.e., in cases where studies show a substantial association between the intervention/exposure and the outcome.
- Dose-response relationship
  - This domain assesses whether there is evidence of a *dose-response relationship* between intervention or exposure and the corresponding result of the study. Evidence of a correlation between the two variables increases the confidence in the evidence base.

- All plausible residual confounding
  - This domain considers whether we can be confident that confounding would reduce, rather than explain, the reported effect. Or when no effect was observed, that *all plausible residual confounding* would suggest a spurious effect.

Of note, the quantity of research—such as the number of identified studies reporting on an outcome—is reported in the summary of findings table, but there is no general rule for downgrading the quality of evidence based on the quantity of research. However, the number of identified studies is likely to affect the overall confidence in the evidence base. In particular, research findings that have not been replicated by other, independent research groups have to be considered with caution when they are the only effect estimate and form the basis of conclusion of the review. The quantity of research needs to be evaluated together with the study limitations. A small number of studies with low risk of bias is more informative than a large number of studies with limited validity.

Some published approaches have opted to revise the interpretation of or to add guidance to individual criteria, such as defining imprecise or sparse data (*imprecision*). A position statement for grading evidence and recommendations for clinical practice guidelines in nephrology, for example, agreed to not downgrade evidence when there are adequately powered, relevant randomized controlled trials with more than 1,000 participants, even if only one study has been published (Uhlig et al., 2006). The current guidelines for systematic reviews in the Cochrane Back and Neck Group explicitly recommend that data should be judged imprecise when only one small study reports an outcome (Furlan et al., 2015).

Indirectness addresses several aspects, including the external validity of the included studies (see previous chapter). In particular where nonhuman data are considered, the *indirectness* domain needs to take the applicability of the research on the occupational safety and health review question into account. The NTP OHAT systematic review handbook (OHAT, 2015) considers the relevance of the animal model to the outcome of interest, the directness of the endpoints to the primary health outcomes, the nature of the exposure in human studies and route of administration in animal studies, and the duration of treatment in animal studies and length of time between exposure and outcome assessment in animal and prospective human studies. In addition, the handbook provides guidance to determine the relevance of different animal studies (e.g., invertebrate model systems) to the human-hazard systematic review question. As outlined, the applicability of studies conducted outside of occupational settings needs to be critically reviewed.

Grading the quality of evidence becomes more complex with increasing diversity of evidence that is being considered in the systematic review, and the criteria may need to be adapted according to the evaluated evidence. Apart from adaptations of individual grading criteria, there are also body-of-evidence evaluation adaptations that use different sets of criteria, or additional criteria. Three examples are outlined in the next subsection.

Body-of-Evidence Criteria Set Adaptations

The GRADE system has been adapted for several nonintervention studies, including prognostic factor studies (Huguet et al., 2013). The prognostic factor framework differentiates eight criteria: the *phase of investigation* determines whether the risk-factor evidence is primarily based on a study that aimed to identify potential prognostic factors (moderate quality) rather than based on studies aiming to confirm identified associations or explanatory research aiming to understand prognostic pathways (high quality). *Indirectness* takes into account whether the available research studies accurately reflect the review question. Evidence can be downgraded for *imprecision* if the sample size is insufficient, the confidence interval is wide and overlaps the value of no effect, there are fewer than ten outcome events for each prognostic variable, or there are fewer than 100 cases reaching endpoints. *Publication bias* is assessed because researchers may fail to report the lack of relationship between the potential prognostic factor and outcomes, and this bias should be considered unless the value of the risk or protective factor in predicting the outcome has been repetitively investigated. Consistent with GRADE, evidence may also be downgraded for *study limitations* and *inconsistency.* Evidence may be upgraded if effects are *moderate or large* or there is evidence of an *exposure-gradient* response for factors measured at different doses.

To update the evidence for patient safety practices, AHRQ commissioned two reports: one to determine criteria for assessing the effectiveness and safety and then the analysis of the evidence for patient safety (Shekelle et al., 2010; and Shekelle et al., 2013). This was deemed necessary because of the nature of patient-safety research (often assessing rare events, in organizational settings, and not using trial methodology) that does not necessarily fit traditional quality-of-evidence assessments. The team determined that the *strength of evidence* should be decreased for *important inconsistency across studies*; *serious imprecision*; *high probability of reporting bias*; *no explanation in any of the studies of why the patient safety practice might work* (either in terms of theory, logic models, or prior success in other fields or in pilot studies); and the *patient safety practice is not described in sufficient detail to permit replication*. The strength of evidence can be increased for a *very strong effect in the majority of studies*; *all plausible residual confounding would reduce a demonstrated effect or would suggest a spurious effect if no effect was observed*; and *use of theory/logic models, assessment of contexts, reporting of implementation process, and fidelity of implementation*. In addition, for observational studies, the assessment score should be increased for the *use of observational study designs of stronger internal validity* (controlled before and after, time series, statistical process control).

The PRECEPT framework (Harder et al., 2015) suggests upgrading evidence for *consistency of findings across settings and study designs* for risk factor and intervention studies. Upgrading for consistency of findings differs from downgrading for *inconsistency* as defined in the standard GRADE approach.

Quality-of-Evidence Starting Point

Systems to grade evidence set starting points from which the quality of evidence is either upgraded or downgraded (using the criteria to evaluate the quality of evidence). The study design of identified research usually initially determines a hierarchical grade of the evidence (Baker et al., 2010). The basic GRADE system is designed for clinical-intervention studies and starts with randomized controlled trials as high-quality evidence; other study designs are initially classified as low-quality evidence (Guyatt et al., 2011). The key advantage of randomized controlled trials over other study designs is that the random assignment to intervention and control groups functions as a mechanism to generate equivalent baseline groups. This ensures that differences among groups after the intervention exposure can be attributed to the intervention and not to differences in sample characteristics and other differences between groups.

Types of employed study designs differ by research field. For example, randomized controlled trials are unethical when assessing risk of toxicity in humans. The quality of the body of evidence in such a research field could only be high if other GRADE criteria (*large effects, dose-response relationship,* or *all plausible residual confounding*) apply that upgrade the quality of evidence. Applying a standard clinical-research evidence hierarchy to research fields where key study designs are missing reduces the ability to differentiate the quality of evidence in these fields. It can create an artificial floor effect because initially *high* quality of evidence does not exist and evidence is likely to be systematically rated as *low* or *very low quality*. Because of the absence of *high*-quality evidence, conclusions would necessarily be considered unsubstantiated (Concato, 2004). The quality-of-evidence starting point needs to be critically reviewed and, in research fields where study designs are missing due to conceptual reasons (e.g., randomized controlled trials cannot be performed), the starting point and the study-design hierarchy may need to be revised to ensure applicability of the grading system.

Evidence Grading Starting-Point Adaptations

Several research and policy areas have opted to revise the trial-centered starting point in body-of-evidence evaluations. Van Staa and colleagues (2008) make a strong case that, when evaluating drug-toxicity signals, randomized controlled trials should not be considered the highest research standard. The PRECEPT framework determined that nonrandomized designs that are less prone to bias (such as interrupted time series) should initially be judged as *moderate* quality rather than *low* quality, and risk-factor studies (e.g., step-function changes) should initially be rated as *high* quality (Harder et al., 2015). Systems for prognostic studies may start with longitudinal cohort studies as *high* quality of evidence (Iorio et al., 2015) or use the phase of investigation (exploratory or confirmatory study) to determine the initial level (Huguet et al., 2013). One key difference between the standard GRADE approach and the Navigation Guide systematic review methodology is that observational human studies are initially rated as *moderate* quality rather than as *low* quality (Woodruff and Sutton, 2011; Johnson et al., 2014; Koustas et al., 2014; Lam et al., 2014; and Woodruff and Sutton, 2014).

Other evidence hierarchies have been published to differentiate nontreatment studies (Merlin, Weston, and Tooher, 2009); systems that give more guidance for grading other than randomized controlled trial designs (Gugiu, 2015); and pertinent adaptations of the GRADE approach for areas such as diagnostic tests (Singh et al., 2012) that do not use randomized controlled trial research methodology.

Body-of-Evidence Quality Levels

The quality of evidence describes the extent to which we can be confident that an effect estimate is correct. As outlined, the effect estimate in a systematic review is based on the identified research, e.g., the summary of effects across relevant research studies. The GRADE system differentiates four levels of confidence. Literature reviewers arrive at these levels by applying criteria to evaluate the body of evidence (e.g., *inconsistency*) that was described in the previous section. The levels are assigned on outcome level so they apply to specific evidence statements (e.g., effect of the preventive workplace initiatives on the number of accidental falls). The GRADE group revised the interpretation of the levels in 2011 (Balshem et al., 2011) as follows:

- High quality of evidence
  - We are very confident that the true effect lies close to that of the estimate of the effect.
- Moderate quality of evidence
  - We are moderately confident in the effect estimate: The true effect is likely to be close to the estimate of the effect, but there is a possibility that it is substantially different.
- Low quality of evidence
  - Our confidence in the effect estimate is limited: The true effect may be substantially different from the estimate of the effect.
- Very low quality of evidence
  - We have very little confidence in the effect estimate: The true effect is likely to be substantially different from the estimate of effect.

In contrast, the AHRQ EPC program differentiates *high*, *moderate*, *low*, and *insufficient evidence* (AHRQ, 2014). *Insufficient* evidence indicates that no concrete evidence statements can be derived from the literature.

While the GRADE approach and the AHRQ EPC program categories are meant to be applicable to a broad range of evidence, there also are published grading suggestions for specific applications that have modified GRADE for their purposes. The Cochrane Back and Neck Group has added a fifth category, *no evidence*, to the four GRADE categories to communicate that no randomized controlled trial was identified that addressed the outcome of interest (Furlan et al., 2015). A framework for grading the evidence of effectiveness of health-promotion interventions

differentiates *Grade 1* (strong: what works is known, a high association is shown, and repeatability is widely demonstrated); *Grade 2* (weak: what works is known, but a low association is shown, and/or repeatability is limited); and *Grade 3* (insufficient: what works is not known). The first two grades differentiate two levels to describe whether it is known how the intervention works or not (Tang, Choi, and Beaglehole, 2008). The NTP OHAT approach to health assessments concerning environmental chemicals and physical substances does not describe the level of the quality of evidence but describes the level of confidence in the evidence directly. The approach differentiates four levels of confidence: *high*, *moderate*, *low*, and *very low confidence*. *High confidence* indicates that the review authors have high confidence in the association between exposure to the substance and the outcome; the true effect is highly likely to be reflected by the apparent relationship. *Moderate confidence* in the association indicates that the true effect may be reflected in the apparent relationship. *Low confidence* in the association indicates that the outcome with the true effect may be different than the apparent relationship. Finally, *very low confidence* indicates that the association indicates that the true effect is highly likely to be different than the apparent relationship (OHAT, 2015).

Quality-of-Evidence Summary

A summary of findings table (see Table 6.1) helps to provide a clear overview of the findings and the quality of the evidence the findings are based on. The table should show the number of studies and the results for individual outcomes together with criteria used to evaluate the reviewers' confidence in the summary statement and an overall quality assessment (e.g., "high quality of evidence"). The table should also indicate where evidence was upgraded or downgraded, and by what factor, to arrive at the summary. In the standard GRADE approach, quality of evidence can be downgraded by one (e.g., from *high* to *moderate*) or by two (e.g., from *high* to *low*), depending on the extent to which the issue affects the confidence in the finding.

Grading the quality of evidence becomes more complex as the diversity of the evidence being considered in the systematic review increases. The Navigation Guide methodology (Woodruff and Sutton, 2011; Johnson et al., 2014; Koustas et al., 2014; Lam et al., 2014; and Woodruff and Sutton, 2014) applies quality-of-evidence criteria separately for human and nonhuman studies. Detailed instructions can be found here.

Integrating Evidence

A challenging step when reviewing evidence is to integrate the individual pieces of research in the systematic review. This may include results across different outcomes (e.g., number of days absent, staff turnover measures) that report on the same conceptual domain (e.g., all contributing to the determination of the risk of harms to humans in occupational settings). In addition, evidence may need to be integrated across different research domains, such as risk-factor

research results and effects of interventions. In occupational safety and health reviews, evidence from different lines of evidence (e.g., human research and in vitro findings) will often need to be integrated.

Evidence may also need to be weighed across different review questions and outcomes, such as benefits versus harms; however, this balance is likely to be context- and application-specific and is often deferred to the stage of drawing conclusions for practice from the review and formulating recommendations. (Evidence-based) recommendations are based on the available research, but they will often draw on additional considerations than the state of the evidence base alone. Chapter Seven addresses balancing issues, while this chapter explores criteria and decision rules to integrate elements of evidence that may be diverse but that still contribute to the same general review question (e.g., determining safety).

First, different evidence elements may report similar findings. For example, there may be similar results for different outcomes that can be summarized in individual evidence statements (e.g., for specific adverse events) or combined to more-general evidence statements (e.g., general safety risk). Risk-factor studies may indicate a preventive effect, supported by emerging evidence from intervention studies. Different lines of evidence may come to similar conclusions, in which case the different lines are complementary or, in some cases, even redundant (e.g., where sufficient human studies exist, no other research evidence is needed). However, evidence integration is challenging when the systematic review identifies a mismatch between the elements, either in findings across lines of evidence or in the presence or lack of research in some lines of evidence. An example is in vitro studies indicating the potential for serious adverse effects on humans, while no human-participant study exists that has empirically demonstrated a risk to humans. Where different pieces of evidence differ in their conclusions, it is crucial that the systematic review is transparent regarding how elements, such as results from different lines of evidence, were combined and weighed. Ideally, the overall framework should be described in the analytic framework (see Chapter Three).

Weight of Evidence

Evidence integration is a key aspect of weight-of-evidence approaches. Weight-of-evidence analyses are routinely used in ecological and human health risk assessments to collate heterogeneous information and to justify regulatory benchmarks and recommendations (Occupational Safety and Health Administration, 2012). However, the term is defined and used in a variety of ways, it often describes a methodological approach to interpret and evaluate data without further explanation, and there is no agreed-upon approach across scientific communities using the term (Krimsky, 2005; Weed, 2005; Lutter et al., 2015). Stakeholders have pointed out that, while a weight-of-evidence evaluation generally appeals to logic and the scientific process, how it is implemented in practice is less clear (Cormier, 2011). Published approaches vary considerably, and how weights are assigned to specific evidence is not always reported. Systematic reviews are not a standard component in the process, and individual examples have

been criticized for subjectivity in the selection as well as the assignment of strength and relevance of the evidence supporting a hypothesis or statement.

Researchers have proposed more-formalized approaches to weight-of-evidence approaches, such as quantitative frameworks using multi-criteria decision-analysis methodology, incorporating decision-analysis tools, using hypothesis-based assessments to better communicate uncertainties and inconsistencies, applying quantitative structure-activity relationship modeling, performing sequential analyses of lines of evidence to rule out risks, or using discriminant analyses to assign empirically based weights (Hull and Swanson, 2006; Matthews et al., 2007; Hujoel, 2009; Swaen and Van Amelsvoort, 2009; Rhomberg, Bailey, and Goodman, 2010; Linkov et al., 2011; Rhomberg et al., 2011; Pemberton, Bailey, and Rhomberg, 2013; Prueitt, Rhomberg, and Goodman, 2013; Rorije et al., 2013; Jiang et al., 2015; and Linkov et al., 2015). Others have advocated paradigm shifts, such as the use of Bayesian methods to integrate evidence from reviewed research and prior knowledge (Linkov et al., 2015). Suter and Cormier have compared strengths and weaknesses of a number of methods for weighing evidence (2011).

A review of 50 weight-of-evidence frameworks resulted in the formulation of 17 best practices for integrating and evaluating evidence (Rhomberg et al., 2013). Best practices relevant to different lines of evidence include:

- Assess all human and nonhuman data relevant to mode of action, their human relevance, and dose-response.
- Evaluate the types of data that have been considered.
- Trace the reasoning by which the data bear on the evaluation-of-the-assessment question.
- Consider alternative modes of action and develop a biological story for each plausible mode of action/outcome combination.
- Consider the relevance, response, and predictivity of the outcomes and use other knowledge (e.g., biological pathways) to inform the relevance determinations.
- Integrate data across all lines of evidence so that the interpretation of one will inform the interpretation of others (e.g., if the proposed mode of action were true, determine what observable consequences it should have across lines of evidence).
- Clearly present the weight-of-evidence findings and explore ways to measure and communicate different magnitudes of *weight* of evidence and different degrees of plausibility of explanations and their risk-assessment consequences.

The authors highlight that integrating evidence is central to the success of a sound weight-of-evidence approach but note that further discussion and methodological research on evidence integration is warranted. The authors also suggest applying the Bradford Hill Criteria to all evidence to serve as a guide to identify patterns in the data that can be linked to hypothesized underlying explanations (Rhomberg et al., 2013).

Bradford Hill Criteria

The Bradford Hill Criteria (Hill, 1965) identifies nine aspects of an observed association between an event and an environmental feature that could be considered in exploring cause and effect and in determining when it is appropriate to act to protect safety and health. These are:

1. Strength: ratio of the occurrence of the event in individuals with exposure compared with individuals without exposure to the environmental feature.
2. Consistency: repeated observation of the association by different individuals, at different times, in different places, and under different circumstances.
3. Specificity: observance of the association only in specific workers at specific sites.
4. Temporality: temporal relationship between the event and the environmental feature.
5. Biological gradient: presence of a biological gradient or dose-response curve
6. Plausibility: suspected cause and effect interpretation is biologically plausible
7. coherence: suspected cause and effect interpretation should not conflict with generally known facts of natural history and disease
8. Experiment: whether a preventive action does prevent the event
9. Analogy: similar evidence of event and environmental associations.

Bradford Hill believed none of the nine factors should be considered a definitive interpretation of cause and effect, and none were required to be present for cause and effect to be determined. However, he believed that together, they could be used to help decide whether any other interpretation was more likely than cause and effect. Many approaches to evidence integration incorporate some or all of the Bradford Hill criteria. Howick, Glasziou, and Aronson published a revision of the criteria to make them easier to use (2009). The authors suggested consolidating the criteria into the following three categories:

- Direct evidence
  - studies demonstrating that the effect size is greater than the combined influence of possible confounders
- Mechanistic evidence
  - information on the alleged causal process that connects the intervention and the outcome
- Parallel evidence
  - research that supports the causal hypothesis suggested in a study with related studies that have similar results (replicability, similarity).

Integrating Lines of Evidence in Systematic Reviews

The general approach to integrating evidence will depend on which lines of evidence were considered and whether the systematic review addresses the effectiveness of prevention initiatives, explores safety standards, conducts risk assessments, aims to establish guidelines for exposure levels, or a combination of the above.

Different lines of evidence are routinely considered by the International Agency for Research on Cancer monographs on the evaluation of carcinogenic risks to humans (WHO, 2006). While not explicitly based on a systematic review of the literature, the monographs set out to consider all available published research studies of cancer in humans, studies of cancer in experimental animals, and mechanistic and other relevant data. The studies are summarized and evaluated for carcinogenicity within lines of evidence. Human and animal data can be categorized as containing *sufficient evidence*, *limited evidence*, *inadequate evidence*, and *evidence suggesting lack of carcinogenicity*. Mechanistic data judged to affect the overall evaluation are also highlighted. In an overall evaluation, integrating evidence across lines of evidence, five categories are differentiated:

- Group 1: The agent is carcinogenic to humans.
  - The *Group 1* category is used when there is *sufficient* evidence of carcinogenicity in humans (although exceptions can be made).
- Group 2A: The agent is probably carcinogenic to humans.
  - *Group 2A* includes cases of *limited* evidence in humans and *sufficient* evidence in experimental animals.
- Group 2B: The agent is possibly carcinogenic to humans.
  - *Group 2B* is used for agents for which there is *limited* evidence in humans and *less than sufficient* evidence in experimental animals or *inadequate* evidence in humans but *sufficient* evidence in animals.
- Group 3: The agent is not classifiable as to its carcinogenicity to humans.
  - *Group 3* applies to bodies of evidence where evidence is *inadequate* in humans and *inadequate* or *limited* in experimental animals.
- Group 4: The agent is probably not carcinogenic to humans.
  - *Group 4* is used for agents for which there is evidence suggesting *lack of* carcinogenicity in humans and in experimental animals. The category can also be used for *inadequate* evidence in humans, but evidence suggesting *lack of* carcinogenicity in experimental animals supported by a broad range of mechanistic and other relevant data.

The Navigation Guide systematic review methodology (Woodruff and Sutton, 2014) also explicitly combines diverse lines ("streams") of evidence including *in vitro*, *in vivo*, *in silico*, and human observational studies to inform decisionmaking on environmental chemical exposures. The methodology describes four steps: (1) specify the study questions; (2) select evidence through a systematic search; (3) rate the risk of bias for individual studies and the quality and strength of evidence across studies (separately for human and nonhuman studies) and integrate the strength ratings; and (4) grade the strength of the recommendations. Published examples (Johnson et al., 2014; Koustas et al., 2014; and Lam et al., 2014) report evidence ratings that

reflect the level of certainty of toxicity: *sufficient evidence*, *limited evidence*, *inadequate evidence*, and *evidence of lack of toxicity* across all research studies. Considerations for determining strength of evidence include *quality of the body of evidence*, *direction of effect estimates*, *confidence in effect estimates*, and *other aspects of the data that may affect certainty*. The integration of the strength of evidence in human and nonhuman studies is determined based on the intersection of the ratings (Lam et al., 2014). It incorporates the International Agency for Research on Cancer's approach and EPA strength-of-evidence descriptions. The approach weighs human and animal data, with emphasis on human data. Where *sufficient* evidence from human data exists, other evidence does not need to be considered. This results in one of five strength-of-evidence conclusions: *known to be toxic* (*sufficient* toxicity evidence from studies in humans, regardless of nonhuman evidence); *probably toxic* (*limited* evidence of toxicity in humans but *sufficient* evidence of toxicity in animals); *possibly toxic* (*limited* evidence of toxicity in humans and *limited*, *inadequate*, or *evidence of lack of toxicity* in animals); *not classifiable* (*inadequate* evidence because of the limited number or size of studies, low quality of individual studies, or inconsistency of findings across individual studies); and *probably not toxic* (*evidence of lack of toxicity* on an adequate array of endpoints in more than one study from at least two species).

The NTP OHAT systematic review handbook (OHAT, 2015) also describes an approach to integrate results across outcomes and lines of evidence. The methodology involves seven steps: (1) formulate the problem and develop protocol, (2) search for and select studies for inclusion, (3) extract data from studies, (4) assess internal validity of individual studies, (5) synthesize evidence and rate confidence in body of evidence, (6) translate confidence ratings into level of evidence for health effect, and (7) integrate evidence to develop hazard identification conclusions. The fifth step of synthesizing evidence and rating confidence in the body of evidence combines confidence conclusions for all study types and multiple outcomes. While the initial confidence assessments were applicable to individual outcomes, this step develops conclusions for multiple outcomes. Biologically related outcomes may contribute to the overall health outcome conclusions. The body of evidence is reevaluated across the outcomes of interest from the initial evaluation within outcome categories and potentially upgraded to higher confidence because of consistency across outcomes. NTP OHAT guidance also addresses different lines of evidence. First, cross–species/population/study consistency is one of the criteria that can increase confidence in the body of evidence (Rooney et al., 2014). This includes consistent results across multiple models or species, consistent results across populations (human or wildlife), and consistent results reported in studies with different design features (e.g., experimental and observational studies). Furthermore, systematic review step seven is based on the integration of human and animal evidence. The NTP OHAT approach uses five hazard identification conclusion categories: *known to be a hazard to humans*, *presumed to be a hazard to humans*, *suspected to be a hazard to humans*, *not classifiable as a hazard to humans*, and *not identified as a hazard to humans*. The rating is primarily the result of the intersection between

two dimensions: level of evidence for health effects in human studies (*high*, *moderate*, or *low/inadequate*) and the level of evidence for health effects in nonhuman studies (*high*, *moderate*, or *low/inadequate*). Mechanistic data, such as *in vitro* and *in vivo* laboratory tests directed at cellular, biochemical, or molecular mechanisms, can also be used to upgrade or to downgrade evidence, but the approach is less structured and project dependent (OHAT, 2015).

Of note, all three presented approaches to integrating lines of evidence determine specific risks (i.e., carcinogenicity, toxicity, or hazards), not how these determinations translate into recommendations for prevention. Recommendation documents may include a risk assessment but primarily aim to determine how to best prevent the risk. In that case, a faceted risk determination may not be feasible or required, and the detailed quality of evidence assessment may instead concentrate on the evidence for preventative mechanisms. The appropriate level of differentiation will depend on the scope of the systematic review and the recommendations. The level of differentiation and the level of detail that will be used to answer the systematic review questions should be determined at the protocol state and added to the analytic framework where appropriate.

Expert Input in Assessing and Interpreting the Evidence

Integration of evidence often involves content expert input. Content experts may evaluate the identified evidence or may even fill data gaps in the available evidence base (Hristozov et al., 2014). The AHRQ National Guideline Clearinghouse differentiates expert consensus via a committee, expert consensus via Delphi method, expert consensus not further specified, weighting according to a rating scheme with the scheme given, weighting according to a rating scheme but no scheme given, and subjective review. The Delphi method uses two or more rounds of expert ratings and, after each round, panelists receive feedback on the rating as well as reasons provided for the judgments (Fitch et al., 2001). Experts are encouraged to revise earlier answers in light of the input from other members of the panel.

In addition to stating the criteria or algorithms used to summarize, evaluate, and integrate evidence, the format of the expert input also should be documented in guidance documents, particularly regarding how consensus was achieved across stakeholders.

Body-of-Evidence Evaluation and Interpretation Transparency

Synthesizing, grading, and integrating evidence is a complex process, in particular when different lines of evidence contribute to the systematic review. Synthesizing evidence across studies in homogenous categories is a first step, and the synthesis should be clearly documented because it is the basis of two further instrumental steps in summarizing evidence.

As outlined, published examples of evidence grading show that adaptations of the GRADE approach are possible and defendable. One system may not fit all occupational safety and health applications given the multidisciplinary nature; the diversity of questions (effectiveness, safety,

or risk assessment); lines of evidence; and topic areas. It might be more useful to start with the standard GRADE system, or another published system, and to adapt the system as needed. Given the popularity of GRADE, a good approach is to document any deviations from the standard process, thereby using GRADE as a point of reference that readers are familiar with. Documenting which aspect of grading the evidence was adapted should be made explicit, such as the starting point of the evaluation system (e.g., moderate quality of evidence for interrupted time series); the criteria to up- and downgrade evidence (e.g., *consistency* of results across studies); and the quality-of-evidence levels (e.g., *high, moderate, low, insufficient quality of evidence*). The approach should be outlined in the systematic review protocol and not be decided after research results have been reviewed. Furthermore, given the subjectivity involved in the process, steps to reduce errors and bias should be taken, such as having independent reviewers rate the evidence and peer review by technical expert panel members.

Methods to integrate findings across lines of evidence continue to develop. Where the systematic review is based on a range of elements and lines of evidence, the review needs to be transparent regarding how the different information was combined and how the elements were weighted. There is agreement that the process should be transparent, specified in detail, structured, and defensible. Algorithms developed for integrating research along lines of evidence will need to be able to allow case-by-case judgment and interpretation. Regardless of the method used, expert judgment is still required, and the procedure should be documented together with the criteria used to weigh the different evidence elements.

7. Draw Conclusions and Develop Recommendations

This chapter assumes that the systematic review has been completed. The first part of this chapter, "Systematic Review Reporting," addresses the reporting of the methods and results of the completed systematic review. The second part, "Conclusions and Recommendations," addresses the conclusions that follow from the systematic review, in particular when it informs recommendations or guidance documents. A final section addresses translational products.

An additional step that is increasingly used by producers of systematic reviews is to peer review the systematic review and recommendations and to publically post draft documents. AHRQ EPC reports, for example, are extensively peer reviewed and responses to comments are included in the final reports. Public posting of the draft report promotes transparency and can elicit responses that help to clarify sections in the report. The NTP OHAT manual lists as one of the functions of public posting that any data in the data extraction can be corrected (OHAT, 2015). U.S. government agencies can seek input through a "Request for Information" in the Federal Register. Some agencies have established dedicated portals for public commenting.

Systematic Review Reporting

This section addresses the elements that should be reported for systematic reviews. The international PRISMA initiative has developed guidelines for authors of systematic reviews. PRISMA is endorsed by many scientific journals and represents the current reporting standard for systematic reviews. The basic checklist covers the following elements (Moher et al., 2000):

- Title
  - should identify the report as a systematic review.
- Abstract
  - structured, addressing background, objective, data sources, study eligibility criteria, appraisal and synthesis methods, results, limitations, conclusions, implications, and the PROSPERO registration number.
- Introduction
  - rationale (in the context of what is already known)
  - objective (using the PICO framework).
- Methods
  - protocol and registration
  - eligibility criteria
  - information sources
  - search (the full search strategy should be presented for at least one database)

  - study selection process
  - data collection process
  - data items (items for which information was sought)
  - risk of bias in individual studies
  - summary measures (e.g., risk ratio, difference in means)
  - synthesis of results (e.g., meta-analysis method)
  - risk of bias across studies (criteria such as publication bias)
  - additional analyses (e.g., pre-specified subgroup analysis).
- Results
  - study selection (flow diagram with reasons for exclusion, see Figure 4.2)
  - study characteristics (study details for all included studies, see Table 5.1)
  - risk of bias within studies (see Figure 5.1)
  - results of individual studies (ideally with a forest plot, see Figure 6.1)
  - synthesis of results (see Table 6.1)
  - risk of bias across studies (see Table 6.1)
  - additional analysis.
- Discussion
  - summary of evidence
  - limitations (at study and outcome as well as at review level)
  - conclusions (interpretation of results and implications for future research).
- Funding
  - sources as well as the role of funders in the systematic review.

As shown in the previous chapters, there are standard tables and figures that document the literature field, the review processes, and the content of the identified literature. The literature flow diagram is a key element of the documentation of a systematic review. Evidence tables provide an overview of the existing research studies and allow the reader to review the evidence without further interpretation. The results of the critical appraisal can either be shown in a table or a figure. The summary of findings table summarizes the results of the study across studies and incorporates the body of evidence assessment.

As mentioned in Chapter Three, not all elements need to be reported in the main document if there is a publicly available review protocol to which the authors can refer. This could, for example, include the full search strategy or risk of bias criteria definitions and scoring rules. Other options include creating an online appendix accompanying the main document.

It is particularly important to alert the reader to any action undertaken to reduce the resource intensity of the review, such as single-reviewer data abstraction that may have introduced errors and bias into the review. Furthermore, given the rapidly evolving research fields, the date of the literature search needs to be clearly documented.

The full report of a systematic review will need to contain many details to ensure transparency. Elements such as the flow diagram, evidence table, and summary of findings table are designed to provide a clear overview to summarize the available evidence. A concise synthesis is of particular importance when the systematic review is supporting guidelines and recommendations.

Conclusions and Recommendations

This section describes the step from reviewing the evidence to formulating recommendations and implications for practice that follow from the evidence. The results of the systematic review are included in a summary of the existing evidence base. What to conclude from the evidence, in particular what the implications are for current practice, is a further step. The Cochrane handbook explicitly discourages review authors from making recommendations (Higgins and Green, 2011). Different stakeholders may make different decisions based on the same evidence, and the purpose of the systematic review should be to present information and aid interpretation, rather than to offer recommendations. However, a key role of systematic reviews is to inform health and safety guidelines for practice.

Developing Recommendations

First, evidence-based recommendations are more easily made when the evidence base is clearly documented. This may be achieved by presenting a summary of the findings and quality of evidence evaluation for each of the different pieces of evidence or a summary of the integrated evidence across individual outcomes, domains, and lines of evidence as discussed in Chapter Six.

Furthermore, criteria to consider when formulating recommendations should be structured. In particular, clinical practice guideline developers have highlighted the importance of clear presentations of the evidence and a structured approach to formulating recommendations. The National Health and Medical Research Council of Australia uses a standardized form to structure information that was developed after reviewing several existing frameworks for developing practice guidelines (Hillier et al., 2011). The form is available in an online appendix to a journal article outlining the agency's current process. The form documents a rating (*excellent, good, satisfactory*, or *poor*) of five independent factors (*evidence base*, *consistency*, *clinical impact*, *generalizability*, and *applicability*). The system has four resulting levels of recommendations, ranging from "the body of evidence can be trusted to guide practice" to "the body of evidence is weak and recommendations must be applied with caution."

The evidence summary may list different outcomes representing both benefits and harms (or downsides) associated with a preventive intervention or safety recommendation. Balancing these opposing dimensions is a key challenge for authors of recommendations. For occupational safety and health, recommended courses of action may not be associated with "harms" equivalent to adverse events or side effects known in clinical intervention research. However, side effects of

recommendations could be social or socioeconomic in nature. For example, recommended safety procedures will be associated with required time and resources and may have clear monetary consequences. In other cases, these undesirable consequences of recommendations may not have been part of the evidence synthesis but are nonetheless apparent to stakeholders (e.g., resulting economic burden or the need to restructure current practices in order to comply with the new standards). In this case, other factors than the evidence base will be considered in the process of formulating recommendations.

The original GRADE system (Atkins et al., 2004), developed to support clinical practice guidelines, recommended four factors be considered when making a recommendation: (1) the trade-offs, taking into account the estimated size of the effect for the main outcomes, the confidence limits around those estimates, and the relative value placed on each outcome; (2) the quality of the evidence; (3) translation of the evidence into practice in a specific setting, taking into consideration important factors that could be expected to modify the size of the expected effects, such as availability of necessary expertise; and (4) uncertainty about baseline risk for the population of interest (to accurately balance benefits and harms). Examples of factors relevant to occupational safety and health recommendations that may or may not have been part of the evidence synthesis, are projected costs of the preventive action, current industry standards, context-dependent values and practices, stakeholder preferences, and technical feasibility. Standards and training modules of the Centers for Disease Control and Prevention (CDC) for guidelines and recommendations require that recommendations are clear, transparent, and *implementable*. Software and critical appraisal tools, such as GEMCutter, eGLIA, and BridgeWiz, may help identify barriers to implementation. Similarly, the Community Guide Systematic Review Methods for public health questions describes identifying issues of applicability and barriers to implementation for recommended interventions as one of 14 steps in the review process. All factors contributing to recommendations require careful and context-dependent judgment. The considerations may either influence whether a recommendation is issued for a particular aspect or influence the strength of the recommendation as outlined below.

While undoubtedly additional factors than the research evidence base have to be considered to make recommendations, some guideline groups have also included safeguards for considering too much, or inappropriate, input. A WHO Rapid Advice Guidelines (Schünemann et al., 2007a) outlined a set of rules: Additional evidence would only be allowed at the guideline meeting if it had been omitted from the evidence summaries or was new and critical for decisionmaking, the GRADE approach was to be used to grade the quality of evidence and the strength of recommendations, recommendations were to be based on a consensus of the panel and voting would be used if agreement could not be reached, all panel members would be asked to consider their conflicts of interest and to abstain from voting if necessary, and subsequent interaction and discussion would take place through email but recommendations would not be changed after the meetings except for minor wording changes or corrections of factual errors.

Both the criteria that have shaped the recommendations, as well as the process of developing the recommendations, should be documented in detail.

Expert Input Approaches in Formulating Recommendations

There is no standard approach to obtaining and incorporating input from experts. Where groups of experts are consulted, the method to consensus finding is a key aspect of the input. The AHRQ National Guideline Clearinghouse differentiates expert consensus via Delphi method, consensus via nominal group technique, consensus via consensus development conference, expert consensus not further specified, informal consensus, balance sheets, and not stated.

Research on expert input has addressed the effects of panel composition, process of finding consensus, and format of the panel (e.g., face-to-face meeting). Face-to-face meetings enable fruitful real-time exchanges, but meetings are resource intense (e.g., travel), difficult to schedule given that there are many demands on experts' time, and panel success depends in part on the skills of the panel moderator. Online expert panels can accommodate a larger number of panelists, and it is possible to conduct multiple parallel panels to test for the reproducibility of panel conclusions (Khodyakov et al., 2011). The original Delphi method uses written questionnaires, while variations such as the RAND/UCLA Appropriateness Method (Fitch et al., 2001) typically include a face-to-face meeting. Panelists meet after the first round of ratings and can discuss the ratings in person, focusing on areas of disagreement (Fitch et al., 2001). After the discussion, panelists re-rate individually. Of note, in this approach, no attempt is made to force the panel to consensus; the two-round process is designed to determine whether discrepant ratings are due to real and unresolvable disagreement among experts or due to fatigue or misunderstandings ("artefactual disagreement"). The nominal group technique (Delbecq and Van de Ven, 1971), originally developed for problem identification and solution generation, brings panelists together and asks each person to describe the most important idea on his or her list and continues around the table until everyone's ideas have been listed. After discussion, panelists are asked to individually rank order or rate their judgment. Summary estimates of the effects of alternatives can be represented as a balance sheet depicting benefits and harms for each strategy and can help illustrate the trade offs (Berg, Atkins, and Tierney, 1997).

Professional judgment is an important component of the development of occupational safety and health recommendations, regardless of the systematic review method and other tools used to evaluate the available evidence. Documenting where professional judgment is used in the decisionmaking process and acknowledging the assumptions and uncertainties that underlie recommendations are important to providing a transparent basis of formal recommendations. The format of the expert input should be documented, in particular how consensus was achieved and how disagreements between content experts and stakeholders were handled. Disagreements, for example, may be used as an indication that no recommendation should be formulated, and strong consensus may influence the strength of recommendations.

Strength of Recommendations

The GRADE workgroup has published a framework for linking the quality of the evidence and the strength of recommendations (Guyatt et al., 2008). Part of this approach is that the strength of recommendations is graded. The strength of recommendations indicates the extent to which we can be confident that adherence to the recommendation will do more good than harm (Atkins et al., 2004). The standard system differentiates four cases: *strong recommendation for using an intervention*, *weak recommendation for using an intervention*, *weak recommendation against using an intervention*, and *strong recommendation against using an intervention*.

Ideally, strong recommendations are based on good quality of evidence. Very often, however, in particular for safety questions, there is a lack of strong evidence. Data may stem from observational studies with confounders that hinder confident effectiveness or safety statements. The effects of preventive services are particularly difficult to establish when they aim to prevent rare events. To determine whether the frequency of an already-rare event has changed requires very large sample sizes, limiting data availability and the use of traditional statistical significance testing (Lipscomb and Dement, 2009, and Hempel et al., 2015a). While safety recommendations should be evidence based, they have to take inherent research limitations of the area of interest into account. Safety recommendations need to consider the consequences of false positives and false negatives. Furthermore, the standards of proof and evidence thresholds for safety concerns have to be different than, for example, establishing the effectiveness of an intervention.

The quality of evidence is only one determinant of the strength of recommendations. Other considerations are the balance between desirable and undesirable effects, values and preferences, and costs or resource allocation needed (Guyatt et al., 2008). Factors that could weaken the strength of a recommendation in the WHO Rapid Advice Guidelines (Schünemann et al., 2007a) were *lower quality evidence*, *uncertainty about the balance of benefits versus harms and burdens*, *uncertainty or differences in values*, *marginal net benefits or downsides*, and *uncertainty about whether the net benefits are worth the costs*. The Navigation Guide approach describes grading the strength of recommendations as the last step in their systematic review methodology (Woodruff and Sutton, 2014). This step considers *values and preferences* of stakeholders, *extent of exposures*, *availability of safer alternatives*, and *costs and benefits* in addition to the strength of evidence and information about exposure; however, the workgroup has not yet published on the operationalization of this step.

Reporting of Recommendations

The AHRQ National Guideline Clearinghouse requires registered guidelines to be based on a systematic review of evidence and to document each of the following features in the guideline or its supporting documents:

- an explicit statement that the clinical practice guideline was based on a systematic review

- a description of the search strategy that includes a listing of database(s) searched, a summary of search terms used, the specific time period covered by the literature search including the beginning date (month/year) and end date (month/year), and the date(s) when the literature search was done
- a description of study selection that includes the number of studies identified, the number of studies included, and a summary of inclusion and exclusion criteria
- a synthesis of evidence from the selected studies, e.g., a detailed description or evidence tables
- a summary of the evidence synthesis that relates the evidence to the recommendations, e.g., a descriptive summary or summary tables.

Reporting is particularly important for authoritative recommendations and regulatory industry standards. The exact phrasing of the recommendations is likely to be closely scrutinized. Software may help create templates for recommendation statements to ensure consistency through controlled language (Shiffman et al., 2012).

Increasingly, stakeholders request that conclusions and recommendations are clearly linked to the reviewed evidence. A hallmark of evidence-based recommendations is to be explicit when opinion is used so that the readers understand the basis for the recommendations (Woolf, 2006). The aspects of the recommendations that are based on research reviewed in the systematic review and those based on other input and that do not follow directly from the reviewed evidence should be transparent (see "Analytic Framework" section in Chapter Three).

Different professional bodies have different standards and preferences for presenting conclusions and recommendations and linking the evidence and recommendations. A study surveying occupational exposure limits for metals and other mining-related chemicals identified a number of published approaches and a need to increase transparency in exposure-limit documentation (Haber and Maier, 2002). The authors suggested that documentation should adhere to good risk characterization principles and should identify the methodology used and scientific judgments made, the data used as the basis for the limit calculation, and the uncertainties and overall confidence in the limit derivation. WHO carcinogenic risk assessments require workgroups to report the reasoning used to reach their evaluation, to report conclusions on the strength of evidence together with citations to indicate which studies were pivotal to the conclusions, and an explanation of the reasoning in weighing data and making evaluations (WHO, 2006). Best practices resulting from the weight-of-evidence framework review (Rhomberg et al., 2013) suggested that the reasoning behind the conclusions and the basis for reaching them be clearly documented along with support for all plausible alternative hypotheses. In addition, conclusions should be presented using a narrative discussion accompanied by diagrams, tables, and figures illustrating the logic and results of the evaluation, including alternative explanations of the data.

The GRADE working group suggests documenting the strength of the recommendation together with the content of the recommendation. The group acknowledges that various forms of

presentations exist across organizations issuing recommendations, including the use of numbers and letters or symbols to denote the strength of a recommendation (Guyatt et al., 2008).

If recommendations are graded, the grading as well as the definition of the grades need to be made explicit. The GRADE strength of recommendations reflect the extent to which we can be confident that the desirable effects of a health care intervention outweigh the undesirable effects (Guyatt et al., 2008). The implications of a strong recommendation are, for patients, that most people would want the recommended course of action; for clinicians, that most people should receive the recommended course of action; and for policymakers, that the recommendations can be adopted as a policy in most cases. The implications of weak recommendations are, for patients, that most people would want the recommended course but many would not; for clinicians, that they should recognize that different choices will be appropriate for different patients; and, for policymakers, that substantial debate and involvement of many stakeholders are required. In some cases, different recommendations may need to be formulated for different populations, e.g., low-, moderate-, or high-risk exposure groups (Schünemann et al. 2007a). The exact wording of the recommendations and the implications for the end user have to be critically reviewed. This is particularly important for occupational safety and health applications that do not represent choices or mere alerts to increase awareness about a specific health hazard, but instead aim to formulate authoritative requirements and standards. An (exaggerated) example is a weak recommendation that may be ignored because of its grading, which defeats the purpose of issuing recommendations. The usefulness and effectiveness of grading occupational safety and health recommendations has not been evaluated.

Translational Products

It can be useful (and sometimes necessary) to provide additional translation products for pertinent stakeholders, such as overviews for policymakers. The AHRQ Effective Health Care Program provides an online platform to identify consumer summaries, clinician summaries, and policymaker summaries for evidence reports in health care research. The Canadian Institute for Work and Health has produced a number of short online summaries accompanying full systematic review reports to facilitate sharing research evidence for specific occupational safety and health questions. In addition to written material, other media—such as toolkits with slide decks and links to online resources for employers and videos—could make recommendations more accessible to stakeholders. Materials designed for implementation, such as checklists and interactive tools, can facilitate uptake in practice.

8. Discussion and Outlook

This report organizes systematic reviews of the literature into basic steps and a chapter each is dedicated to the following:

- Define the question.
- Create a protocol.
- Conduct a literature search and screen for inclusion.
- Document and assess included studies.
- Evaluate and interpret the body of evidence.
- Draw conclusions and develop recommendations based on the systematic review.

Each chapter outlines the processes involved in the individual systematic review step. The chapter outlines considerations specific to occupational safety and health and provides references that are particularly important for this very complex field. A number of resources are presented and links to online resources are provided. Resources were selected to be of use for producers and consumers of systematic reviews in occupational safety and health.

In the course of this project, we did not identify a published systematic review handbook that appeared to addresses all unique occupational safety and health requirements ready for adoption by agencies producing systematic reviews in this field. Systematic reviews in occupational safety and health address a very large field of interdisciplinary research, and a wide range of lines of evidence can contribute to individual systematic reviews and recommendations. While several guideline documents exist that can generally guide systematic reviews in occupational safety and health research, we identified several unique requirements, such as the integration of different lines of evidence that would benefit from further development tailored toward occupational safety and health. However, a large number of pertinent resources can support systematic reviews in occupational safety and health as documented in the individual chapters.

This report documents existing approaches to systematic reviews relevant to occupational safety and health and lists resources to support these reviews but purposefully provides few specific recommendations for the conduct of these reviews. A number of decisions have to be made, such as determining the scale of review efforts for each individual systematic review. In addition, different producers of systematic reviews need to determine standards or guidance for their reviews. Buy-in from the community of stakeholders is key when developing systematic review guidance. Stakeholder engagement will ensure that the guidance meets the need of the field and will ultimately determine whether the guidance will be followed.

Systematic review as a research methodology is a rapidly evolving field. Scientific journals dedicated to systematic reviews and international working groups such as the GRADE group and PRISMA continue to aim to improve methods. Detailed handbooks by organizations producing systematic reviews such as the EPC Methods Guide and the Cochrane handbook are designed as

living documents. Sections are reviewed periodically to ensure the need for updating and the guidance is typically updated by chapter on an as-needed basis. Updates are available from their respective websites. Similarly, a new systematic review guideline that has just been published by NTP is also designed as a living document, and it is anticipated that the guidance will be updated as methodological practices are refined and evaluated and new pertinent strategies are identified to improve the reliability, ease, and efficiency of conducting systematic reviews (U.S. Department of Health and Human Services, 2015).

This report targets systematic reviews and recommendations based on the evidence synthesis. Both products are summaries of the evidence *at the time* of the document development. Researchers have pointed out that systematic reviews can become out of date very fast given the rapidly evolving research fields they are often summarizing. A survival analysis for a sample of 100 systematic reviews reported that, within two years, 23 percent of reviews were out of date, and seven percent were determined to be in need of updating at the time of the journal publication (Shojania et al., 2007). Deciding when a systematic review is out of date is complicated by the fact that while new research is constantly published, the new studies may either confirm the systematic review conclusions or challenge them.

Determining whether a systematic review is out of date presents a key challenge for producers of systematic reviews. The AHRQ EPC program uses a standardized system to detect signals for the need for updating clinical effectiveness systematic reviews in health care. The procedure includes literature searches using the original search strategy but limited to five high-impact general journals and five topic-specific journals. Publications are screened and data are abstracted to determine whether the new research supports the systematic review conclusion. Signals for the need for updating could be a pivotal trial with results opposite to that of the original systematic review or the finding of a superior new treatment. In addition, clinical content experts are contacted to provide input on whether the conclusions of the original report are still valid and to identify any new citations that may invalidate or alter the conclusions. Finally, safety and adverse event alerts relevant to each systematic review are collected from three reporting systems. Surveillance of systematic reviews is performed every six months. Tests showed that, while more systematic reviews were classified as medium or high priority for updating, as both the search time lapse and the number of new relevant publications increased, no threshold existed for either time or number of publications that could accurately predict updating needs (Ahmadzai et al., 2013).

Recommendations based on the systematic review may need to be adjusted as the evidence base matures or changes. Furthermore, recommendations are context specific, and other factors than the evidence base may equally change. For example, the associated costs with a specific preventive measure may have changed over time, reducing the economic burden and making specific recommendations easier to implement. Consequently, a mechanism to periodically review issued recommendations may need to be established to ensure that recommendations continue to represent the best advice at the time.

This report provides resources tailored to systematic reviews in occupational safety and health and provides an overview of general steps in systematic reviews, highlights key methodological considerations, and identifies practical resources to support occupational safety and health evidence synthesis.

References

Agency for Healthcare Research and Quality, *Methods Guide for Effectiveness and Comparative Effectiveness Reviews*, Rockville, Md., AHRQ publication No. 10(14)-EHC063-EF, January 2014.

Ahmadzai, Nadera, Sydne J. Newberry, Margaret A. Maglione, Alexander Tsertsvadze, Mohammed T. Ansari, Susanne Hempel, Aneesa Motala, Sophia Tsouros, Jennifer J. Schneider Chafen, Roberta Shanman, David Moher, and Paul G. Shekelle, "A Surveillance System to Assess the Need for Updating Systematic Reviews," *Systematic Review*, Vol. 2, 2013, p. 104.

AHRQ—*See* Agency for Healthcare Research and Quality.

Alli, Benjamin O., *Fundamental Principles of Occupational Health And Safety*, 2nd ed., Geneva, Switzerland: International Labour Organization, 2008.

Ankley, Gerald T., Richard S. Bennett, Russell J. Erickson, Dale J. Hoff, Michael W. Hornung, Rodney D. Johnson, David R. Mount, John W. Nichols, Christine L. Russom, Patricia K. Schmieder, Jose A. Serrano, Joseph E. Tietge, and Daniel L. Villeneuve, "Adverse Outcome Pathways: A Conceptual Framework to Support Ecotoxicology Research and Risk Assessment," *Environmental Toxicology and Chemistry*, Vol. 29, No. 3, March 2010, pp. 730–741.

Atkins, David, Dana Best, Peter A. Briss, Martin Eccles, Yngve Falck-Ytter, Signe Flottorp, Gordon H. Guyatt, Robin. T. Harbour, Margaret C. Haugh, David Henry, Suzanne Hill, Roman Jaeschke, Gillian Leng, Alesandro Liberati, Nicola Magrini, James Mason, Philippa Middleton, Jacek Mrukowicz, Dianne O'Connell, Andrew D. Oxman, Bob Phillips, Holger J. Schünemann, Tessa Tan-Torres Edejer, Helena Varonen, Gunn E. Vist, John W. Williams, Jr., and Stephanie Zaza (members of the GRADE Working Group), "Grading Quality of Evidence and Strength of Recommendations," *British Medical Journal*, Vol. 328, No. 7454, June 2004, p. 1490.

Baker, Adrian, Katharine Young, Jonathan Potter, and Ira Madan, "A Review of Grading Systems for Evidence-Based Guidelines Produced by Medical Specialties," *Clinical Medicine*, Vol. 10, No. 4, August 2010, pp. 358–363.

Balshem, Howard, Mark Helfand, Holger J. Schünemann, Andrew D. Oxman, Regina Kunz, Jan Brozek, Gunn E. Vist, Yngve Falck-Ytter, Joerg Meerpohl, Susan Norris, and Gordon H. Guyatt, "GRADE Guidelines: 3. Rating the Quality of Evidence," *Journal of Clinical Epidemiology*, Vol. 64, No. 4, April 2011, pp. 401–406.

Berg, Alfred O., David Atkins, and William Tierney, "Clinical Practice Guidelines in Practice and Education," *Journal of General Internal Medicine*, Vol. 12, No. S25, April 1997, pp. S25–S32.

Berkman, Nancy D., Kathleen N. Lohr, Mohammed Ansari, Marian McDonagh, Ethan Balk, Evelyn Whitlock, James Reston, Eric Bass, Mary Butler, Gerald Gartlehner, Lisa Hartling, Robert Kane, Melissa McPheeters, Laura Morgan, Sally C. Morton, Meera Viswanathan, Priyanka Sista, and Stephanie Chang, *Methods Guide for Comparative Effectiveness Reviews: Grading the Strength of a Body of Evidence When Assessing Health Care Interventions for the Effective Health Care Program of the Agency for Healthcare Research and Quality: An Update*, Rockville, Md.: Agency for Healthcare Research and Quality, AHRQ Publication No. 13(14)-EHC130-EF, November 2013.

Briss, Peter A., Stephanie Zaza, Marguerite Pappaioanou, Jonathan Fielding, Linda Wright-De Aguero, Benedict I. Truman, David P. Hopkins, Patricia Dolan Mullen, Robert S. Thompson, Steven H. Woolf, Vilma G. Carande-Kulis, Laurie Anderson, Alan R. Hinman, David V. McQueen, Steven M. Teutsch, and Jeffrey R. Harris, "Developing an Evidence-Based Guide to Community Preventive Services—Methods," *American Journal of Preventive Medicine*, Vol. 18, No. 1S, 2000, pp. 35–43.

Burchett, Helen, Muriah Umoquit, and Mark Dobrow, "How Do We Know When Research from One Setting Can Be Useful in Another? A Review of External Validity, Applicability, and Transferability Frameworks," *Journal of Health Services Research and Policy*, Vol. 16, No. 4, October 2011, pp. 238–244.

Bushman, Brad J., and Gary L. Wells, "Narrative Impressions of Literature: The Availability Bias and the Corrective Properties of Meta-Analytic Approaches," *Personality and Social Psychology Bulletin*, Vol. 27, No. 9, September 2001, pp. 1123–1130.

Centre for Reviews and Dissemination, *Systematic Reviews: CRD's Guidance for Undertaking Reviews in Health Care*, Heslington, York, United Kingdom: University of York, 2009.

Concannon, Thomas W., Paul Meissner, Jo Anne Grunbaum, Newell McElwee, Jeanne-Marie Guise, John Santa, Patrick H. Conway, Denise Daudelin, Elaine H. Morrato, and Laurel K. Leslie, "A New Taxonomy for Stakeholder Engagement in Patient-Centered Outcomes Research," *Journal of General Internal Medicine*, Vol. 27, No. 8, April 2012, pp. 985–991.

Concato, John, "Observational Versus Experimental Studies: What's the Evidence for a Hierarchy?" *NeuroRx*, Vol. 1, No. 3, July 2004, pp. 341–347.

Cormier, Joseph W., "Advancing FDA's Regulatory Science Through Weight of Evidence Evaluations," *Journal of Contemporary Health Law and Policy*, Vol. 28, No. 1, August 2011, pp. 1–22.

Dalal, Siddharta R., Paul G. Shekelle, Susanne Hempel, Sydne J. Newberry, Aneesa Motala, and Kanaka D. Shetty, "A Pilot Study Using Machine Learning and Domain Knowledge to Facilitate Comparative Effectiveness Review Updating," *Medical Decision Making*, Vol. 33, No. 3, April 2013, pp. 343–355.

Danz, Margie Sherwood, Susanne Hempel, Yee-Wei Lim, Roberta Shanman, Aneesa Motala, Susan Stockdale, Paul Shekelle, and Lisa Rubenstein, "Incorporating Evidence Review into Quality Improvement: Meeting the Needs of Innovators," *BMJ Quality and Safety*, Vol. 22, No. 11, November 2013, pp. 931–939.

Delbecq, André L., and Andrew H. Van de Ven, "A Group Process Model for Problem Identification and Program Planning," *Journal of Applied Behavioral Science*, Vol. 7, No. 4, July 1971, pp. 466–492.

Dickersin, Kay, Yuan-I. Min, and Curtis L. Meinert, "Factors Influencing Publication of Research Results: Follow-Up of Applications Submitted to Two Institutional Review Boards," *Journal of the American Medical Association*, Vol. 267, No. 3, January 1992, pp. 374–378.

Dickersin, Kay, Roberta Scherer, and Carol Lefebvre, "Identifying Relevant Studies for Systematic Reviews," *British Medical Journal*, Vol. 309, No. 6964, November 12, 1994, pp. 1286–1291.

Dyrvig, Anne-Kristine, Kristian Kidholm, Oke Gerke, and Hindrik Vondeling, "Checklists for External Validity: A Systematic Review," *Journal of Evaluation in Clinical Practice*, Vol. 20, No. 6, December 2014, pp. 857–864.

Eden, Jill, Laura Levit, Alfred Berg, and Sally Morton, eds., *Finding What Works in Health Care: Standards for Systematic Reviews*, Washington, D.C.: Institute of Medicine of the National Academies, 2011.

European Food Safety Authority, "Principles and Process for Dealing with Data and Evidence in Scientific Assessments," *EFSA Journal*, Vol. 13, No. 5, June 2015.

Featherstone, Robin M., Donna M. Dryden, Michelle Foisy, Jeanne-Marie Guise, Matthew D. Mitchell, Robin A. Paynter, Karen A. Robinson, Craig A. Umscheid, and Lisa Hartling, "Advancing Knowledge of Rapid Reviews: An Analysis of Results, Conclusions and Recommendations from Published Review Articles Examining Rapid Reviews," *Systematic Reviews*, Vol. 4, No. 50, 2015.

Fitch, Kathryn, Steven J. Bernstein, Maria Dolores Aguilar, Bernard Burnand, Juan Ramon LaCalle, Pablo Lazaro, Mirjam van het Loo, Joseph McDonnell, Janneke Vader, and James P. Kahan, *The RAND/UCLA Appropriateness Method User's Manual*, Santa Monica, Calif.: RAND Corporation, MR-1269-DG-XII/RE, 2001. As of May 2, 2016: http://www.rand.org/pubs/monograph_reports/MR1269.html

Furlan, Andrea D., Antti Malmivaara, Roger Chou, Chris G. Maher, Rick A. Deyo, Mark Schoene, Gert Bronfort, and Maurits W. van Tulder, "2015 Updated Method Guideline for Systematic Reviews in the Cochrane Back and Neck Group," *Spine*, Vol. 40, No. 21, 2015, pp. 1660–1673.

Ganann, Rebecca, Donna Ciliska, and Helen Thomas, "Expediting Systematic Reviews: Methods and Implications of Rapid Reviews," *Implementation Science*, Vol. 5, 2010.

Giustini, D., and M. N. K. Boulos, "Google Scholar Is Not Enough to Be Used Alone for Systematic Reviews," *Online Journal of Public Health Informatics*, Vol. 5, No. 2, February 2013, pp. 214.

Green, Lawrence W., and Russell E. Glasgow, "Evaluating the Relevance, Generalization, and Applicability of Research: Issues in External Validation and Translation Methodology," *Evaluation and the Health Professions*, Vol. 29, No. 1, March 2006, pp. 126–153.

Gugiu, P. Cristian, "Hierarchy of Evidence and Appraisal of Limitations (HEAL) Grading System," *Evaluation and Program Planning*, Vol. 48, February 2015, pp. 149–159.

Guyatt, Gordon, Andrew D. Oxman, Elie A. Akl, Regina Kunz, Gunn Vist, Jan Brozek, Susan Norris, Yngve Falck-Ytter, Paul Glasziou, Hans deBeer, Roman Jaeschke, David Rind, Joerg Meerpohl, Philipp Dahm, and Holger J. Schünemann, "GRADE Guidelines: 1. Introduction—GRADE Evidence Profiles and Summary of Findings Tables," *Journal of Clinical Epidemiology*, Vol. 64, No. 4, April 2011, pp. 383–394.

Guyatt, Gordon H., Andrew D. Oxman, Regina Kunz, Yngve Falck-Ytter, Gunn E. Vist, Alessandro Liberati, and Holger J. Schünemann, "Going from Evidence to Recommendations," *British Medical Journal*, Vol. 336, No. 7652, May 2008, pp. 1049–1051.

Haber, Lynne T., and Andrew Maier, "Scientific Criteria Used for the Development of Occupational Exposure Limits for Metals and Other Mining-Related Chemicals," *Regulatory Toxicology and Pharmacology*, Vol. 36, No. 3, December 2002, pp. 262–279.

Haddaway, Neal R., Alexandra Mary Collins, Deborah Coughlin, and Stuart Kirk, "The Role of Google Scholar in Evidence Reviews and Its Applicability to Grey Literature Searching," *PLoS ONE*, Vol. 10, No. 9, September 2015, e0138237.

Harder, Thomas, Muna Abu Sin, Xavier Bosch-Capblanch, Bruno Coignard, Helena de Carvalho Gomes, Phillippe Duclos, Tim Eckmanns, Randy Elder, Simon Ellis, Frode Forland, Paul Garner, Roberta James, Andreas Jansen, Gérard Krause, Daneil Lévy-Bruhl, Antony Morgan, Joerg J. Meerpohl, Susan Norris, Eva Rehfuess, Alexa Sánchez-Vivar, Holger Schünemann, Anja Takla, Ole Wichmann, Walter Zingg, and Teun Zuiderent-Jerak, "Towards a Framework for Evaluating and Grading Evidence in Public Health," *Health Policy*, Vol. 119, No. 6, June 2015, pp. 732–736.

Hartling, Lisa, Jeanne-Marie Guise, Susanne Hempel, Robin Featherstone, Matthew D. Mitchell, Makalapua L. Motu'apuaka, Karen A. Robinson, Karen Schoelles, Annette Totten, Evelyn Whitlock, Timothy Wilt, Johanna Anderson, Elise Berliner, Aysegul Gozu, Elisabeth Kato, Robin Paynter, and Craig A. Umscheid, *EPC Methods: AHRQ End-User Perspectives of Rapid Reviews*, Rockville, Md.: Agency for Healthcare Research and Quality, April 2016.

Hartling, Lisa, Jeanne-Marie Guise, Elisabeth Kato, Johanna Anderson, Naomi Aronson, Suzanne Belinson, Elise Berliner, Donna Dryden, Robin Featherstone, Michelle Foisy, Matthew Mitchell, Makalapua Motu'apuaka, Hussein Noorani, Robin Paynter, Karen A. Robinson, Karen Schoelles, Craig A. Umscheid, and Evelyn Whitlock, *EPC Methods: An Exploration of Methods and Context for the Production of Rapid Reviews*, Rockville, Md.: Agency for Healthcare Research and Quality, February 2015a.

Hartling, Lisa, Jeanne-Marie Guise, Elisabeth Kato, Johanna Anderson, Suzanne Belinson, Elise Berliner, Donna M. Dryden, Robin Featherstone, Matthew D. Mitchell, Makalapua Motu'apuaka, Hussein Noorani, Robin Paynter, Karen A. Robinson, Karen Schoelles, Craig A. Umscheid, and Evelyn Whitlock, "A Taxonomy of Rapid Reviews Links Report Types and Methods to Specific Decision-Making Contexts," *Journal of Clinical Epidemiology*, Vol. 68, No. 12, December 2015b, pp. 1451–1462.e1453.

Hartling, Lisa, Michele P. Hamm, Andrea Milne, Ben Vandermeer, P. Lina Santaguida, Mohammed Ansari, Alexander Tsertsvadze, Susanne Hempel, Paul Shekelle, and Donna M. Dryden, "Testing the Risk of Bias Tool Showed Low Reliability Between Individual Reviewers and Across Consensus Assessments of Reviewer Pairs," *Journal of Clinical Epidemiology*, Vol. 66, No. 9, September 2013, pp. 973–981.

Hempel, Susanne, Melinda Maggard-Gibbons, David K. Nguyen, Aaron J. Dawes, Isomi Miake-Lye, Jessica M. Beroes, Marika J. Booth, Jeremy N. V. Miles, Roberta Shanman, and Paul G. Shekelle, "Wrong-Site Surgery, Retained Surgical Items, and Surgical Fires: A Systematic Review of Surgical Never Events," *Journal of the American Medical Association Surgery*, Vol. 150, No. 8, June 2015a, pp. 796–805.

Hempel, Susanne, Jeremy Miles, Marika J. Suttorp, Zhen Wang, Breanne Johnsen, Sally Morton, Tanja Perry, Diane Valentine, and Paul G. Shekelle, *AHRQ Methods for Effective Health Care. Detection of Associations Between Trial Quality and Effect Sizes*, Rockville, Md.: Agency for Healthcare Research and Quality, AHRQ Publication No. 12-EHC010-EF, January 2012a.

Hempel, Susanne, Lisa V. Rubenstein, Roberta M. Shanman, Robbie Foy, Su Golder, Marjorie Danz, and Paul G. Shekelle, "Identifying Quality Improvement Intervention Publications—A Comparison of Electronic Search Strategies," *Implementation Science*, Vol. 6, August 2011a, p. 85.

Hempel, Susanne, Paul G. Shekelle, Jodi L. Liu, Margie Sherwood Danz, Robbie Foy, Yee-Wei Lim, Aneesa Motala, and Lisa V. Rubenstein, "Development of the Quality Improvement Minimum Quality Criteria Set (QI-MQCS): A Tool for Critical Appraisal of Quality Improvement Intervention Publications," *British Medical Journal Quality and Safety*, Vol. 24, No. 12, August 2015b, pp. 796–804.

Hempel, Susanne, Kanaka D. Shetty, Paul G. Shekelle, Lisa V. Rubenstein, Marjorie S. Danz, Breanne Johnsen, and Siddhartha R. Dalal, *Machine Learning Methods in Systematic Reviews: Identifying Quality Improvement Intervention Evaluations*, AHRQ Publication No. 12-EHC125-EF, Rockville, Md.: Agency for Healthcare Research and Quality, September 2012b.

Hempel, Susanne, Marika J. Suttorp, Jeremy N. V. Miles, Zhen Wang, Margaret Maglione, Sally Morton, Breanne Johnsen, Diane Valentine, and Paul G. Shekelle, *Empirical Evidence of Associations Between Trial Quality and Effect Size*, Rockville, Md.: Agency for Healthcare Research and Quality, AHRQ Publication No. 11-EHC045-EF, June 2011b.

Higgins, Julian P. T., and Sally Green, eds., *Cochrane Handbook for Systematic Reviews of Interventions*, the Cochrane Collaboration, 2008.

———, *Cochrane Handbook for Systematic Reviews of Interventions*, version 5.1.0, the Cochrane Collaboration, March 2011. As of May 2, 2016: http://handbook.cochrane.org/

Hill, Austin Bradford, "The Environment and Disease: Association or Causation?" *Proceedings of the Royal Society of Medicine*, Vol. 58, January 1965, pp. 295–300.

Hillier, Susan, Karen Grimmer-Somers, Tracy Merlin, Philippa Middleton, Janet Salisbury, Rebecca Tooher, and Adele Weston, "FORM: An Australian Method for Formulating and Grading Recommendations in Evidence-Based Clinical Guidelines," *BioMed Central Medical Research Methodology*, Vol. 11, 2011, p. 23.

Horsley, Tanya, Orvie Dingwall, and Margaret Sampson, "Checking Reference Lists to Find Additional Studies for Systematic Reviews," *Cochrane Database of Systematic Reviews*, Vol. 8, August 2011, p. Mr000026.

Howick, Jeremy, Paul Glasziou, and Jeffrey K. Aronson, "The Evolution of Evidence Hierarchies: What Can Bradford Hill's 'Guidelines for Causation' Contribute?" *Journal of the Royal Society of Medicine*, Vol. 102, No. 5, May 2009, pp. 186–194.

Hristozov, Danail R., Stefania Gottardo, Marco Cinelli, Panagiotis Isigonis, Alex Zabeo, Andrea Critto, Martie Van Tongeren, Lang Tran, and Antonio Marcomini, "Application of a Quantitative Weight of Evidence Approach for Ranking and Prioritising Occupational Exposure Scenarios for Titanium Dioxide and Carbon Nanomaterials," *Nanotoxicology*, Vol. 8, No. 2, 2014, pp. 117–131.

Huguet, Anna, Jill A. Hayden, Jennifer Stinson, Patrick J. McGrath, Christine T. Chambers, Michelle E. Tougas, and Lori Wozney, "Judging the Quality of Evidence in Reviews of Prognostic Factor Research: Adapting the GRADE Framework," *Systematic Reviews*, Vol. 2, September 2013, p. 71.

Hujoel, Philippe, "Grading the Evidence: The Core of EBD," *Journal of Evidence-Based Dental Practice*, Vol. 9, No. 3, September 2009, pp. 122–124.

Hull, Ruth N., and Stella Swanson, "Sequential Analysis of Lines of Evidence—An Advanced Weight-of-Evidence Approach for Ecological Risk Assessment," *Integrated Environmental Assessment and Management*, Vol. 2, No. 4, October 2006, pp. 302–311.

Iorio, Alfonso, Frederick A. Spencer, Maicon Falavigna, Carolina Alba, Eddie Lang, Bernard Burnand, Tom McGinn, Jill Hayden, Katrina Williams, Beverly Shea, Robert Wolff, Ton Kujpers, Pablo Perel, Per Olav Vandvik, Paul Glasziou, Holger Schünemann, and Gordon Guyatt, "Use of GRADE for Assessment of Evidence About Prognosis: Rating Confidence in Estimates of Event Rates in Broad Categories of Patients," *British Medical Journal*, 2015, p. 350.

Jiang, Yu-Xia, You-Sheng Liu, Guang-Guo Ying, Hong-Wei Wang, Yan-Qiu Liang, and Xiao-Wen Chen, "A New Tool for Assessing Sediment Quality Based on the Weight of Evidence Approach and Grey TOPSIS," *Science of the Total Environment*, Vol. 537, 2015, pp. 369–376.

Johnson, Paula I., Patrice Sutton, Dylan S. Atchley, Erica Koustas, Juleen Lam, Saunak Sen, Karen A. Robinson, Daniel A. Axelrad, and Tracey J. Woodruff, "The Navigation Guide—Evidence-Based Medicine Meets Environmental Health: Systematic Review of Human Evidence for PFOA Effects on Fetal Growth," *Environmental Health Perspectives*, Vol. 122, No. 10, October 2014, pp. 1028–1039.

Keown, Kiera, Dwayne Van Eerd, and Emma Irvin, "Stakeholder Engagement Opportunities in Systematic Reviews: Knowledge Transfer for Policy and Practice," *Journal of Continuing Education in the Health Professions*, Vol. 28, No. 2, Spring 2008, pp. 67–72.

Khangura, Sara, Kristin Konnyu, Rob Cushman, Jeremy Grimshaw, and David Moher, "Evidence Summaries: The Evolution of a Rapid Review Approach," *Systematic Reviews*, Vol. 1, No. 10, 2012, p. 10.

Khodyakov, Dmitry, Susanne Hempel, Lisa Rubenstein, Paul Shekelle, Robbie Foy, Susanne Salem-Schatz, Sean O'Neill, Margie Danz, and Siddharta Dalal, "Conducting Online Expert Panels: A Feasibility and Experimental Replicability Study," *BMC Medical Research Methodology*, Vol. 11, No. 174, 2011.

Koustas, Erica, Juleen Lam, Patrice Sutton, Paula I. Johnson, Dylan S. Atchley, Saunak Sen, Karen A. Robinson, Daniel A. Axelrad, and Tracey J. Woodruff, "The Navigation Guide—Evidence-Based Medicine Meets Environmental Health: Systematic Review of Nonhuman Evidence for PFOA Effects on Fetal Growth," *Environmental Health Perspectives*, Vol. 122, No. 10, October 2014, pp. 1015–1027.

Krimsky, Sheldon, "The Weight of Scientific Evidence in Policy and Law," *American Journal of Public Health*, Vol. 95, No. S1, 2005, pp. S129–136.

Lam, Juleen, Erica Koustas, Patrice Sutton, Paula I. Johnson, Dylan S. Atchley, Saunak Sen, Karen Robinson, Daniel A. Axelrad, and Tracey J. Woodruff, "The Navigation Guide—Evidence-Based Medicine Meets Environmental Health: Integration of Animal and Human Evidence for PFOA Effects on Fetal Growth," *Environmental Health Perspectives*, Vol. 122, No. 10, October 2014, pp. 1040–1051.

Lau, Joseph, Stephanie Chang, Nancy Berkman, Sara J. Ratichek, Howard Balshem, Michelle Brasure, and David Moher, *EPC Response to IOM Standards for Systematic Reviews*, Rockville, Md.: Agency for Healthcare Research and Quality, AHRQ Publication No. 13-EHC006-EF, April 2013.

Linkov, Igor, Olivia Massey, Jeff Keisler, Ivan Rusyn, and Thomas Hartung, "From 'Weight of Evidence' to Quantitative Data Integration Using Multicriteria Decision Analysis and Bayesian Methods," *Altex-Alternativen Zu Tierexperimenten*, Vol. 32, No. 1, 2015, pp. 3–8.

Linkov, Igor, Paul Welle, Drew Loney, Alex Tkachuk, Laure Canis, J. B. Kim, and Todd Bridges, "Use of Multicriteria Decision Analysis to Support Weight of Evidence Evaluation," *Risk Analysis*, Vol. 31, No. 8, March 2011, pp. 1211–1225.

Lipscomb, Hester J., and John M. Dement, "A Counterview on Data Quality and the Systematic Review Process for Occupational Injury Interventions: Are We Missing the Forest for the Trees?" *American Journal of Preventive Medicine*, Vol. 36, No. 4, April 2009, pp. 377–378.

Lutter, Randall, Linda Abbott, Rick Becker, Chris Borgert, Ann Bradley, Gail Charnley, Susan Dudley, Alan Felsot, Nancy Golden, George Gray, Daland Juberg, Mary Mitchell, Nancy Rachman, Lorenz Rhomberg, Keith Solomon, Stephen Sundlof, and Kate Willett, "Improving Weight of Evidence Approaches to Chemical Evaluations," *Risk Analysis*, Vol. 35, No. 2, December 2015, pp. 186–192.

Maier, A., M. Vincent, E. Hack, P. Nance, and W. Ball, "Derivation of an Occupational Exposure Limit for Inorganic Borates Using a Weight of Evidence Approach," *Regulatory Toxicology and Pharmacology*, Vol. 68, No. 3, 2014, pp. 424–437.

Matthews, Edwin J., Naomi L. Kruhlak, R. Daniel Benz, and Joseph F. Contrera, "A Comprehensive Model for Reproductive and Developmental Toxicity Hazard Identification: I. Development of a Weight of Evidence QSAR Database," *Regulatory Toxicology and Pharmacology*, Vol. 47, No. 2, March 2007, pp. 115–135.

Merlin, Tracy, Adele Weston, and Rebecca Tooher, "Extending an Evidence Hierarchy to Include Topics Other than Treatment: Revising the Australian 'Levels of Evidence,'" *BMC Medical Research Methodology*, Vol. 9, No. 34, 2009.

Moher, David, Ba' Pham, Terry P. Klassen, Kenneth F. Schulz, Jesse A. Berlin, Alejandro R. Jadad, and Alessandro Liberati, "What Contributions Do Languages Other than English Make on the Results of Meta-Analyses?" *Journal of Clinical Epidemiology*, Vol. 53, No. 9, 2000, pp. 964–972.

Munn, Zachary, Sandeep Moola, Karolina Lisy, Dagmara Riitano, and Catalin Tufanaru, "Methodological Guidance for Systematic Reviews of Observational Epidemiological Studies Reporting Prevalence and Cumulative Incidence Data," *International Journal of Evidence-Based Health*, Vol. 13, No. 3, September 2015, pp. 147–153.

Occupational Safety and Health Administration, *Guidance on Data Evaluation for Weight of Evidence Determination: Application to the 2012 Hazard Communication Standard*, Washington, D.C.: United States Department of Labor, 2012.

Office of Health Assessment and Translation, *Handbook for Conducting a Literature-Based Health Assessment Using OHAT Approach for Systematic Review and Evidence Integration*, Washington, D.C.: National Toxicology Program, U.S. Department of Health and Human Services, January 9, 2015.

OHAT—*See* Office of Health Assessment and Translation.

Oxman, Andrew D., Holger J. Schünemann, and Atle Fretheim, "Improving the Use of Research Evidence in Guideline Development: 8. Synthesis and Presentation of Evidence," *Health Research Policy and Systems*, Vol. 4, December 2006, p. 20.

Pemberton, Mark, Lisa A. Bailey, and Lorenz R. Rhomberg, "Hypothesis-Based Weight-of-Evidence Evaluation of Methyl Methacrylate Olfactory Effects in Humans and Derivation of an Occupational Exposure Level," *Regulatory Toxicology and Pharmacology*, Vol. 66, No. 2, July 2013, pp. 217–233.

Pieper, Dawid, Tim Mathes, and Michaela Eikermann, "Can AMSTAR Also Be Applied to Systematic Reviews of Non-Randomized Studies?" *BioMed Central Research Notes*, Vol. 7, 2014, p. 609.

Polisena, Julie, Chantelle Garritty, Craig A. Unscheid, Chris Kamel, Kevin Samra, Jeannette Smith, and Ann Vosilla, "Rapid Review Summit: An Overview and Initiation of a Research Agenda," *Systematic Reviews*, Vol. 4, No. 137, September 2015.

Prueitt, Robyn L., Lorenz R. Rhomberg, and Julie E. Goodman, "Hypothesis-Based Weight-of-Evidence Evaluation of the Human Carcinogenicity of Toluene Diisocyanate," *Critical Reviews in Toxicology*, Vol. 43, No. 5, 2013, pp. 391–435.

Rhomberg, Lorenz. R., Lisa A. Bailey, and Julie E. Goodman, "Hypothesis-Based Weight of Evidence: A Tool for Evaluating And Communicating Uncertainties and Inconsistencies in the Large Body of Evidence in Proposing a Carcinogenic Mode of Action—Naphthalene as an Example," *Critical Reviews in Toxicology*, Vol. 40, No. 8, 2010, pp. 671–696.

Rhomberg, Lorenz R., Lisa A. Bailey, Julie E. Goodman, Ali K. Hamade, and David Mayfield, "Is Exposure to Formaldehyde in Air Causally Associated with Leukemia?—A Hypothesis-Based Weight-of-Evidence Analysis," *Critical Reviews in Toxicology*, Vol. 41, No. 7, April 2011, pp. 555–621.

Rhomberg, Lorenz R., Julie E. Goodman, Lisa A. Bailey, Robyn L. Prueitt, Nancy B. Beck, Christopher Bevan, Michael Honeycutt, Norbert E. Kaminski, Greg Paoli, Lynn H. Pottenger, Roberta W. Scherer, Kimberly C. Wise, and Richard A. Becker, "A Survey of Frameworks for Best Practices in Weight-of-Evidence Analyses," *Critical Reviews in Toxicology*, Vol. 43, No. 9 , 2013, pp. 753–784.

Robinson, Karen A., Roger Chou, Nancy D. Berkman, Sydne J. Newberry, Rongwei Fu, Lisa Hartling, Donna Dryden, Mary Butler, Michelle Foisy, Johanna Anderson, Makalapua Motu'apuaka, Rose Relevo, Jeanne-Marie Guise, and Stephanie Chang, *Methods Guide for Effectiveness and Comparative Effectiveness Reviews: Integrating Bodies of Evidence: Existing Systematic Reviews and Primary Studies*, Rockville, Md.: Agency for Healthcare Research and Quality, AHRQ Publication No. 15-EHC007-EF, February 2015.

Rooney, Andrew A., Abee L. Boyles, Mary S. Wolfe, John R. Bucher, and Kristina A. Thayer, "Systematic Review and Evidence Integration for Literature-Based Environmental Health Science Assessments," *Environmental Health Perspectives*, Vol. 122, No. 7, July 2014, pp. 711–718.

Rorije, Emiel, Tom Aldenberg, Harrie Buist, Dinant Kroese, and Gerrit Schüürmann, "The OSIRIS Weight of Evidence Approach: ITS for Skin Sensitisation," *Regulatory Toxicology and Pharmacology*, Vol. 67, No. 2, 2013, pp. 146–156.

Scherer, Roberta W., Patricia Langenberg, and Erik von Elm, "Full Publication of Results Initially Presented in Abstracts," *Cochrane Database of Systematic Reviews*, Vol. 2, 2007, p. Mr000005.

Schünemann, Holger J., Suzanne R. Hill, Meetali Kakad, Richard Bellamy, Timothy M. Uyeki, Frederick G. Hayden, Yazdan Yazdanpanah, John Beigel, Tawee Chotpitayasunondh, Chris Del Mar, Jeremy Farrar, Tran Tinh Hien, Bülent Özbay, Norio Sugaya, Keiji Fukuda, Nikki Shindo, Lauren Stockman, Gunn E. Vist, Alice Croisier, Azim Nagjdaliyev, Cathy Roth, Gail Thomson, Howard Zucker, and Andrew D. Oxman, "WHO Rapid Advice Guidelines for Pharmacological Management of Sporadic Human Infection with Avian Influenza A (H5N1) Virus," *The Lancet Infectious Diseases*, Vol. 7, No. 1, January 2007a, pp. 21–31.

Schünemann, Holger J., Suzanne R. Hill, Meetali Kakad, Gunn E. Vist, Richard Bellamy, Lauren Stockman, Torbjørn Fosen Wisløff, Chris Del Mar, Frederick Hayden, Timothy M. Uyeki, Jeremy Farrar, Yazdan Yazdanpanah, Howard Zucker, John Beigel, Tawee Chotpitayasunondh, Tran Tinh Hien, Bülent Özbay, Norio Sugaya, and Andrew D. Oxman, "Transparent Development of the WHO Rapid Advice Guidelines," *PLoS Medicine*, Vol. 4, No. 5, 2007b, p. e119.

Schünemann, Holger J., Peter Tugwell, Barnaby C. Reeves, Elie A. Akl, Nancy Santesso, Frederick A. Spencer, Beverley Shea, George Wells, and Mark Helfand, "Non-Randomized Studies as a Source of Complementary, Sequential, or Replacement Evidence for Randomized Controlled Trials in Systematic Reviews on the Effects of Interventions," *Research Synthesis Methods*, Vol. 4, No. 1, March 2013, pp. 49–62.

Selph, Shelley S., Alexander D. Ginsburg, and Roger Chou, "Impact of Contacting Study Authors to Obtain Additional Data for Systematic Reviews: Diagnostic Accuracy Studies for Hepatic Fibrosis," *Systematic Reviews*, Vol. 3, 2014, p. 107.

Shamseer, Larissa, David Moher, Mike Clarke, Davina Ghersi, Alessandro Liberati, Mark Petticrew, Paul Shekelle, Lesley A. Stewart, and PRISMA-P Group, "Preferred Reporting Items for Systematic Review and Meta-Analysis Protocols (PRISMA-P) 2015: Elaboration and Explanation," *British Medical Journal*, Vol. 349, January 2015, p. g7647.

Shea, Beverley J., Jeremy M. Grimshaw, George A. Wells, Maarten Boers, Neil Andersson, Candyce Hamel, Ashley C. Porter, Peter Tugwell, David Moher, and Lex M. Bouter, "Development of AMSTAR: A Measurement Tool to Assess the Methodological Quality of Systematic Reviews," *BMC Medical Research Methodology*, Vol. 7, 2007, p. 10.

Shekelle, Paul G., Peter J. Pronovost, and Robert M. Wachter, "Assessing the Evidence for Context-Sensitive Effectiveness and Safety of Patient Safety Practices: Developing Criteria," Rockville, Md.: Agency for Healthcare Research and Quality, AHRQ Publication No. 11-0006-EF, December 2010.

Shekelle, Paul G., Robert M. Wachter, Peter J. Pronovost, Scott Lucas, Meredith Noble, James Reston, Karen Schoelles, Nancy Sullivan, Fang Sun, Kelley Tipton, Jonathan R. Treadwell, Amy Tsou, Sallie J. Weaver, Bradford D. Winters, Elizabeth Pfoh, Renee Wilson, Kathryn Martinez, Sydney M. Dy, Zack Berger, Breanne Johnsen, Jody Wozar Larkin, Aneesa Motala, Roberta Shanman, Kathryn M. McDonald, Sumant R. Ranji, Stephanie Rennke, Eric Schmidt, Kaveh Shojania, Sydne J. Newberry, and Mary E. Vaiana, *Making Health Care Safer II: An Updated Critical Analysis of the Evidence for Patient Safety Practices*, Evidence Report/Technology Assessment Number 211, Rockville, Md.: Agency for Healthcare Research and Quality, AHRQ Publication No. 13-E001-EF, March 2013.

Shiffman, Richard N., George Michel, Richard M. Rosenfeld, and Caryn Davidson, "Building Better Guidelines with BRIDGE-Wiz: Development and Evaluation of a Software Assistant to Promote Clarity, Transparency, and Implementability," *Journal of the American Medical Informatics Association*, Vol. 19, No. 1, January–February 2012, pp. 94–101.

Shojania, Kaveh G., Margaret Sampson, Mohammed T. Ansari, Jun Ji, Steve Doucette, and David Moher, "How Quickly Do Systematic Reviews Go Out of Date? A Survival Analysis," *Annals of Internal Medicine*, Vol. 147, No. 4, August 2007, pp. 224–233.

Singh, Sonal, Stephanie M. Chang, David B. Matchar, and Eric B. Bass, "Chapter 7: Grading a Body of Evidence on Diagnostic Tests," *Journal of General Internal Medicine*, Vol. 27, No. S1, June 2012, pp. 47–55.

Suter, Glenn W. II, and Susan M. Cormier, "Why and How to Combine Evidence in Environmental Assessments: Weighing Evidence and Building Cases," *Science of the Total Environment*, Vol. 409, No. 8, 2011, pp. 1406–1417.

Swaen, Gerard, and Ludovic van Amelsvoort, "A Weight of Evidence Approach to Causal Inference," *Journal of Clinical Epidemiology*, Vol. 62, No. 3, 2009, pp. 270–277.

Tang, K-C., B. C. K. Choi, and R. Beaglehole, "Grading of Evidence of the Effectiveness of Health Promotion Interventions," *Journal of Epidemiology and Community Health*, Vol. 62, No. 9, September 2008, pp. 832–834.

Tricco, Andrea C., Jesmin Antony, Wasifa Zarin, Lisa Strifler, Marco Ghassemi, John Ivory, Laure Perrier, Brian Hutton, David Moher, and Sharon E. Straus, "A Scoping Review of Rapid Review Methods," *BioMed Central* Medicine, Vol. 13, 2015, p. 224.

Tsertsvadze, Alexander, Yen-Fu Chen, David Moher, Paul Sutcliffe, and Noel McCarthy, "How to Conduct Systematic Reviews More Expeditiously?" *Systematic Reviews*, Vol. 4, 2015, p. 160.

Uhlig, K., A. Macleod, J. Craig, J. Lau, A. S. Levey, A. Levin, L. Moist, E. Steinberg, R. Walker, C. Wanner, N. Lameire, and G. Eknoyan, "Grading Evidence and Recommendations for Clinical Practice Guidelines in Nephrology: A Position Statement from Kidney Disease: Improving Global Outcomes (KDIGO)," *Kidney International*, Vol. 70, No. 12, 2006, pp. 2058–2065.

U.S. Department of Health and Human Services, National Toxicology Program, *Handbook for Conducting a Literature-Based Health Assessment Using OHAT Approach for Systematic Review and Evidence Integration*, January 9, 2015. As of May 24, 2016: https://ntp.niehs.nih.gov/ntp/ohat/pubs/handbookjan2015_508.pdf

Van Staa, T. P., L. Smeeth, I. Persson, J. Parkinson, and H. G. M. Leufkens, "Evaluating Drug Toxicity Signals: Is a Hierarchical Classification of Evidence Useful or a Hindrance?" *Pharmacoepidemiology and Drug Safety*, Vol. 17, No. 5, 2008, pp. 475–484.

Verbeek, Jos, and Frank van Dijk, *A Practical Guide for the Use of Research Information to Improve the Quality of Occupational Health Practice for Occupational and Public Health Professionals*, Protecting Workers' Health Series No. 7, Geneva, Switzerland: World Health Organization, 2006.

Verbeek, J., J. Salmi, I. Pasternack, M. Jauhiainen, I. Laamanen, F. Schaafsma, C. Hulshof, and F. van Dijk, "A Search Strategy for Occupational Health Intervention Studies," *Occupational and Environmental Medicine*, Vol. 62, No. 10, October 2005, pp. 682–687.

Viswanathan, Meera, Timothy S. Carey, Suzanne Belinson, Elise Berliner, Stephanie Chang, Elaine Graham, Jeanne-Marie Guise, Stanley S. Ip, Margaret A. Maglione, Douglas McCrory, Melissa McPheeters, Sydne J. Newberry, Priyanka Sista, and Michael C. White, *Methods Research Report: Identifying and Managing Nonfinancial Conflicts of Interest for Systematic Reviews*, Rockville, Md.: Agency for Healthcare Research and Quality, AHRQ Publication No. 13-EHC085-EF, May 2013.

Von Elm, Eric, Michael C. Costanza, Bernhard Walder, and Martin R. Tramèr, "More Insight into the Fate of Biomedical Meeting Abstracts: A Systematic Review," *BioMed Central Medical Research Methodology*, Vol. 3, July 2003, p. 12.

Watt, Amber, Alun Cameron, Lana Sturm, Timothy Lathlean, Wendy Babidge, Stephen Blamey, Karen Facey, David Hailey, Iinger Norderhaug, and Guy Maddern, "Rapid Reviews Versus Full Systematic Reviews: An Inventory of Current Methods and Practice in Health Technology Assessment," *International Journal of Technology Assessment in Health Care*, Vol. 24, No. 2, April 2008a, pp. 133–139.

———, "Rapid Versus Full Systematic Reviews: Validity in Clinical Practice?" *AZN Journal of Surgery*, Vol. 78, No. 11, November 2008b, pp. 1037–1040.

Weed, Douglas L., "Weight of Evidence: A Review of Concept and Methods," *Risk Analysis*, Vol. 25, No. 6, December 2005, pp. 1545–1557.

West, Suzanne, Valerie King, Timothy S. Carey, Kathleen N. Lohr, Nikki McKoy, Sonya F. Sutton, and Linda Lux, "Systems to Rate the Strength of Scientific Evidence," *Evidence Reports/Technology Assessments*, No. 47, March 2002, pp. 1–11.

WHO—*See* World Health Organization.

Woodruff, Tracey J., and Patrice Sutton, "An Evidence-Based Medicine Methodology to Bridge the Gap Between Clinical and Environmental Health Sciences," *Health* Affairs, Vol. 30, No. 5, May 2011, pp. 931–937.

———, "The Navigation Guide Systematic Review Methodology: A Rigorous and Transparent Method for Translating Environmental Health Science into Better Health Outcomes," *Environmental Health Perspectives*, Vol. 122, No. 10, October 2014, pp. 1007–1014.

Woolf, Steven H., "Weighing the Evidence to Formulate Dietary Guidelines," *Journal of the American College of Nutrition*, Vol. 25, No. 3S, 2006, pp. 277S–284S.

World Health Organization, International Agency for Research on Cancer, *IARC Monographs on the Evaluation of Carcinogenic Risks to Humans: Preamble*, Lyon, France, 2006. As of May 24, 2016:
http://monographs.iarc.fr/ENG/Preamble/CurrentPreamble.pdf

Young, Taryn, and Sally Hopewell, "Methods for Obtaining Unpublished Data," *Cochrane Database of Systematic Reviews*, No. 11, 2011.

Zaza, Stephanie, Linda K. Wright–De Agüero, Peter A. Briss, Benedict I. Truman, David P. Hopkins, Michael H. Hennessy, Daniel M. Sosin, Laurie Anderson, Vilma Carande-Kulis, Steven M. Teutsch, and Marguerite Pappaioanou, "Data Collection Instrument and Procedure for Systematic Reviews in the Guide to Community Preventive Services," *American Journal of Preventive Medicine*, Vol. 18, No. 1S, 2000, pp. 44–74.