Environment International xxx (2016) xxx–xxx

Contents lists available at ScienceDirect

Environment International

journal homepage: www.elsevier.com/locate/envint

How credible are the study results? Evaluating and applying internal validity tools to literature-based assessments of environmental health hazards

Andrew A. Rooney [a], Glinda S. Cooper [b], Gloria D. Jahnke [c], Juleen Lam [d], Rebecca L. Morgan [e], Abee L. Boyles [a], Jennifer M. Ratcliffe [f], Andrew D. Kraft [b], Holger J. Schünemann [e], Pamela Schwingl [f], Teneille D. Walker [b], Kristina A. Thayer [a], Ruth M. Lunn [c,*]

[a] Office of Health Assessment and Translation, Division of the National Toxicology Program, National Institute of Environmental Health Sciences (NIEHS), National Institutes of Health (NIH), Research Triangle Park, NC, USA
[b] National Center for Environmental Assessment, Office of Research and Development, U.S. Environmental Protection Agency, Washington, DC, USA
[c] Office of the Report on Carcinogens, Division of the National Toxicology Program, National Institute of Environmental Health Sciences (NIEHS), National Institutes of Health (NIH), Research Triangle Park, NC, USA
[d] University of California San Francisco, Program on Reproductive Health and the Environment, San Francisco, CA, USA
[e] McMaster University, Department of Clinical Epidemiology and Biostatistics, Hamilton, Ontario, Canada
[f] Integrated Laboratory Systems (ILS), Morrisville, NC, USA

ARTICLE INFO

Article history:
Received 20 July 2015
Received in revised form 2 December 2015
Accepted 10 January 2016
Available online xxxx

Keywords:
Risk of bias
Internal validity
Systematic review
Environmental health
Hazard assessment

ABSTRACT

Environmental health hazard assessments are routinely relied upon for public health decision-making. The evidence base used in these assessments is typically developed from a collection of diverse sources of information of varying quality. It is critical that literature-based evaluations consider the credibility of individual studies used to reach conclusions through consistent, transparent and accepted methods. Systematic review procedures address study credibility by assessing internal validity or "risk of bias" — the assessment of whether the design and conduct of a study compromised the credibility of the link between exposure/intervention and outcome. This paper describes the commonalities and differences in risk-of-bias methods developed or used by five groups that conduct or provide methodological input for performing environmental health hazard assessments: the Grading of Recommendations Assessment, Development, and Evaluation (GRADE) Working Group, the Navigation Guide, the National Toxicology Program's (NTP) Office of Health Assessment and Translation (OHAT) and Office of the Report on Carcinogens (ORoC), and the Integrated Risk Information System of the U.S. Environmental Protection Agency (EPA-IRIS). Each of these groups have been developing and applying rigorous assessment methods for integrating across a heterogeneous collection of human and animal studies to inform conclusions on potential environmental health hazards. There is substantial consistency across the groups in the consideration of risk-of-bias issues or "domains" for assessing observational human studies. There is a similar overlap in terms of domains addressed for animal studies; however, the groups differ in the relative emphasis placed on different aspects of risk of bias. Future directions for the continued harmonization and improvement of these methods are also discussed.
Published by Elsevier Ltd

1. Introduction

The assessment of study quality has long been considered an important part of synthesizing evidence to answer questions in toxicology and environmental health sciences (e.g., IARC, 1990; WHO, 1999). However, the term "study quality" is broad and can vary widely across the fields of systematic review and environmental health (e.g., see terminology discussion in Viswanathan et al., 2012). Recent initiatives in the environmental and occupational health community have emphasized the goal of increasing transparency and objectivity of the evaluation process by adopting systematic review methods (e.g., Birnbaum et al., 2013; EFSA, 2010; Woodruff and Sutton, 2014). As a result of these efforts, there is an increased focus on transparently evaluating one aspect of study quality — the assessment of systematic errors that can result in a biased (over- or under-estimated) effect estimate referred to as risk of bias or internal validity. Risk of bias is a measure of whether the design or conduct of a study alters the effect estimate or compromises the credibility of the reported association (or lack thereof) between exposure/treatment and outcome (Guyatt et al., 2011a; IOM, 2011; Viswanathan et al., 2012). The use of the risk-of-bias terminology has

Abbreviations: AHRQ, Agency for Healthcare Research and Quality; NTP, National Toxicology Program; OHAT, Office of Health Assessment and Translation; ORoC, Office of the Report on Carcinogens; EPA-IRIS, Integrated Risk Information System of the U.S. Environmental Protection Agency; GRADE, Grading of Recommendations Assessment, Development, and Evaluation.

* Corresponding author at: NIEHS, P.O. Box 12233, Mail Drop K2-04, RTP, NC 2770; 530 Davis Drive, Morrisville, NC 27560, USA.
E-mail address: lunn@niehs.nih.gov (R.M. Lunn).

http://dx.doi.org/10.1016/j.envint.2016.01.005
0160-4120/Published by Elsevier Ltd

been supported by systematic review guidance groups such as the Cochrane Collaboration and the Agency for Healthcare Research and Quality (AHRQ) because it reduces ambiguity between the quality of reporting and the quality of the actual conduct of the research (Higgins and Green, 2011; Rooney et al., 2014; Viswanathan et al., 2012).

In this paper, we begin with a discussion of the application of systematic review methods to environmental health. We then present an overview of risk-of-bias approaches that have been developed or used to assess environmental health data by five different groups (the Grading of Recommendations Assessment, Development, and Evaluation [GRADE] Working Group; the Navigation Guide; the National Toxicology Program's [NTP] Office of Health Assessment and Translation [OHAT]; the NTP's Office of the Report on Carcinogens [ORoC]; and the Integrated Risk Information System of the U.S. Environmental Protection Agency [EPA-IRIS]). This analysis is based on discussions that occurred during 2014–2015 to address common interests in understanding, developing, or refining methods for assessing the credibility of individual studies as part of reaching conclusions on specific environmental health questions. Commonalities and differences in the approaches taken across the groups are highlighted along with a discussion of opportunities and challenges for harmonization as methods are refined and further developed over time. To ensure clear communication with a variety of scientific disciplines, definitions for terms commonly used in environmental health reviews and publications are provided (Table 1).

1.1. Application of systematic review methods to environmental health

A systematic review is a literature-based evaluation focused on a specific question that uses explicit, pre-specified methods to identify, select, assess, and synthesize scientific evidence (IOM, 2011). These methods increase the transparency, objectivity, and rigor in the review process. The systematic review methods being applied to environmental health questions have been built on the structure of established approaches for evaluating evidence in clinical medicine and public health, such as the Cochrane Collaboration (Higgins and Green, 2011), the Evidence-based Practice Center (EPC) methods guides for the AHRQ (AHRQ, 2013) and the GRADE Working Group (Atkins et al., 2004; Guyatt et al., 2011a). These approaches typically consider human evidence from different study designs (i.e., randomized controlled trials and observational studies) and have been applied widely to clinical medicine and public health.

Table 1
Definitions of common terms.

Term	Definition
Domain (also used: Category or Question)	Issue or topic within risk of bias such as "confounding" or "selective outcome reporting"
Indirectness (also used: Applicability or external validity)	Measure of how well a study addresses the specific question of the systematic review or the extent to which results inform the review question
Reporting quality (also used: study quality)	Measure of how thoroughly details on study design, experimental procedures, results and analyses were reported (Reporting only addresses a portion of the larger concept of Study Quality; however, sometimes the terms are conflated)
Risk of bias (also used: internal validity, study quality)	Measure of the credibility of study findings that reflects the ability of a study's design and conduct to protect against systematic errors that may bias (over- or under estimate) the results or estimate of effect (Risk of Bias only addresses a portion of the larger concept of Study Quality; however, sometimes the terms are conflated)
Sensitivity	The ability of a study to detect a true risk(similar to the concept of a sensitive assay); an insensitive study will fail to show a difference that truly exists, leading to a false conclusion of no effect. Example considerations include having adequate numbers of exposed cases, exposure levels, durations, ranges, windows of exposure, and lengths of follow-up.
Study quality	A complex idea with different meanings for different groups including one or more of the following: reporting quality, applicability and risk of bias. For systematic review methods study quality generally includes risk of bias assessment.
Systematic review	A review of literature focused on a specific question that uses explicit, pre-specified methods to identify, select, assess, and synthesize scientific evidence

There is considerable variability in the study designs and data sources available to evaluate potential health effects from exposure to environmental chemicals, necessitating some modification of methods developed in clinical medicine. Unlike questions in clinical medicine, environmental datasets rarely include controlled human exposure studies because ethical considerations generally rule out exposing human subjects to chemicals suspected to pose a health hazard. When available, controlled human exposure studies are typically limited to short-term exposures and temporary or reversible health endpoints such as the series of investigations on inflammatory and cardiovascular indicators associated with exposure to diesel exhaust (see Ghio et al., 2012); these types of studies may be of limited relevance to questions regarding effects of longer term exposures. Studies of "natural experiments" wherein researchers take advantage of unplanned exposures or external factors that interrupt exposure [e.g., reduced air pollution associated with the Beijing Olympics allowing an examination of the impact of air pollution on birth weight (Rich et al., 2015)], can provide another useful source of human health effects data (Craig et al., 2012). However, availability of such data is very limited. More typically, human data are derived from a variety of observational designs, including cohort studies, case–control studies, and clinic-based or population-based surveys, as well as from ecological studies or case series or reports.

Questions in environmental health often require the assessment of a broad range of relevant data including animal and mechanistic studies as well as human studies. Experimental animal data, primarily from in vivo laboratory studies in rodents, provide a large proportion of the toxicology data used for hazard identification and risk assessment. Studies of wildlife or animals living in heavily contaminated sites using an observational design may provide health effect data for chemicals that are widely distributed in the environment. Mechanistic data can be found in a wide variety of in vitro and in vivo studies, or studies of molecular, biochemical and cellular events in humans, rather than studies of the disease phenotype (i.e., molecular epidemiology studies). These data may explain how a chemical produces particular adverse health effects and can inform the hazard conclusions.

For environmental health questions, the most widely available in vivo data generally come from experimental animal and observational human epidemiology studies. Whatever the evidence base is, critical assessment of individual studies is needed to evaluate each of the evidence streams (human, animal, and mechanistic studies) with clear consideration of the strengths and weaknesses of different study designs.

2. Overview of current methods (frameworks and tools)

The five groups are involved in conducting systematic reviews that may differ in focus (e.g., cancer or non-cancer endpoints; short term or lifetime hazard evaluations; derivation of risk estimates), scope (individual health endpoints or comprehensive toxicological evaluations; simple or complex literature databases considered), underlying guidance (e.g., agency guidelines that must be adhered to), and use of the systematic reviews by regulatory agencies. The approach taken for evaluating risk of bias and incorporating that evaluation into the systematic review should match the intended purpose of the review for the organization involved. For example, the product of an OHAT systematic review will vary depending on the question and the extent of the available evidence, and may take the form of NTP hazard identification conclusions, opinions on whether substances may be of concern given what is

Please cite this article as: Rooney, A.A., et al., How credible are the study results? Evaluating and applying internal validity tools to literature-based assessments of environment..., Environ Int (2016), http://dx.doi.org/10.1016/j.envint.2016.01.005