Environment International xxx (2016) xxx–xxx

Contents lists available at ScienceDirect

Environment International

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

Study sensitivity: Evaluating the ability to detect effects in systematic reviews of chemical exposures

Glinda S. Cooper [a,*,1], Ruth M. Lunn [b], Marlene Ågerstrand [c], Barbara S. Glenn [a], Andrew D. Kraft [a], April M. Luke [a], Jennifer M. Ratcliffe [d]

[a] National Center for Environmental Assessment, Office of Research and Development, U.S. Environmental Protection Agency, Washington, DC, USA
[b] 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
[c] Department of Environmental Science and Analytical Chemistry (ACES), Stockholm University, 106 91 Stockholm, Sweden
[d] Integrated Laboratory Systems (ILS), Research Triangle Park, NC, USA

ARTICLE INFO

Article history:
Received 17 July 2015
Received in revised form 15 March 2016
Accepted 16 March 2016
Available online xxxx

Keywords:
Systematic review
Validity
Bias
Environmental health
Chemical hazard assessment
Study sensitivity

ABSTRACT

A critical step in systematic reviews of potential health hazards is the structured evaluation of the strengths and weaknesses of the included studies; risk of bias is a term often used to represent this process, specifically with respect to the evaluation of systematic errors that can lead to inaccurate (biased) results (i.e. focusing on internal validity). Systematic review methods developed in the clinical medicine arena have been adapted for use in evaluating environmental health hazards; this expansion raises questions about the scope of risk of bias tools and the extent to which they capture the elements that can affect the interpretation of results from environmental and occupational epidemiology studies and in vivo animal toxicology studies, (the studies typically available for assessment of risk of chemicals). One such element, described here as "sensitivity", is a measure of the ability of a study to detect a true effect or hazard. This concept is similar to the concept of the sensitivity of an assay; an insensitive study may fail to show a difference that truly exists, leading to a false conclusion of no effect. Factors relating to study sensitivity should be evaluated in a systematic manner with the same rigor as the evaluation of other elements within a risk of bias framework. We discuss the importance of this component for the interpretation of individual studies, examine approaches proposed or in use to address it, and describe how it relates to other evaluation components. The evaluation domains contained within a risk of bias tool can include, or can be modified to include, some features relating to study sensitivity; the explicit inclusion of these sensitivity criteria with the same rigor and at the same stage of study evaluation as other bias-related criteria can improve the evaluation process. In some cases, these and other features may be better addressed through a separate sensitivity domain. The combined evaluation of risk of bias and sensitivity can be used to identify the most informative studies, to evaluate the confidence of the findings from individual studies and to identify those study elements that may help to explain heterogeneity across the body of literature.

1. Introduction

Systematic reviews are designed to evaluate bodies of existing evidence regarding specific hypotheses using rigorous, transparent and unbiased methods or approaches (IOM, 2011). While initially developed to evaluate clinical trial data, often using a meta-analysis to summarize results, their application has been extended to the evaluation of observational and animal studies of environmental hazards (Woodruff and Sutton, 2014; Rooney et al., 2014). The purpose of a hazard assessment is to identify and characterize chemical and other environmental hazards to provide the scientific basis, when warranted, for measures to protect public health. There are several challenges faced in adapting the tools used in clinical medicine to this field, however, including the need for an expanded focus on exposure measures, the greater heterogeneity of the type of studies addressing these questions (e.g., observational epidemiology, animal toxicology, in vitro studies) (Whaley et al., in press), and the greater heterogeneity within each type of study (e.g., among observational epidemiology studies, as noted in Sterne et al., 2016).

A critical step in systematic reviews of potential health hazards is to conduct a systematic evaluation of the strengths and weaknesses of the included studies. The evaluation of internal validity assesses the extent to which a study can provide accurate (unbiased) evidence of a causal relationship between a given treatment or exposure and a given effect

* Corresponding author at: U.S. Environmental Protection Agency, 1200 Pennsylvania Ave NW (8601P), Washington, DC 20460, USA.
E-mail address: cooper.glinda@epa.gov (G.S. Cooper).
[1] Courier address: One Potomac Yard (South Building) S-11342, 2777 S. Crystal Dr, Arlington, VA 22202, USA.

http://dx.doi.org/10.1016/j.envint.2016.03.017
0160-4120/

(e.g., does exposure to substance × cause effect A), if such a causal relationship exists. It is often discussed in terms of "risk of bias" or the degree to which specific types of systematic error may have been introduced into the design or execution of a study; these errors can result in a distortion of the results, such that the study does not provide an accurate answer to the research question (Higgins and Green, 2011).

Study sensitivity, as defined here in relation to experimental and observational environmental studies, is a measure of the ability of a study to detect a true effect. It addresses the question "Is this study able to detect a true effect or hazard, if present, that is due to the exposure?" It is analogous to the concept of sensitivity of an assay or a diagnostic test; an insensitive study will fail to detect a difference that truly exists, leading to a false conclusion of no effect; only a negative result from a highly sensitive study can be interpreted, with confidence, as evidence of no effect. Sensitivity can involve features of the study design and measures, study population, and analysis, and can be viewed as a component of internal validity. For the purpose of chemical hazard assessment, study sensitivity is a critical element to include in the evaluation of a study.

This concept of study sensitivity has not been a focus of risk of bias tools and is infrequently discussed with respect to its importance in evaluating "negative" results (of individual studies, or of a collection of studies). This paper aims to contribute to the discussion and development of recommendations for evaluating studies by providing illustrative examples of study sensitivity and by highlighting the importance of this concept when evaluating epidemiological and toxicological studies. The thesis of this paper is that although the conceptualization of internal validity encompasses study sensitivity, this concept is not necessarily adequately addressed by methods used to operationalize the evaluation of internal validity. The specific consideration of study sensitivity should be included in an evaluation of the evidence to avoid a false conclusion that an exposure has no effect, when the lack of evidence may be due to the insensitivity of the studies to detect an effect.

2. Examples of study sensitivity

In this section, we explore in more detail some examples of study sensitivity for epidemiology and animal toxicology studies. Although some examples are unique to one of these types of studies, others are common to both.

Factors in epidemiology studies related to study sensitivity include, but are not limited to, evidence of substantial exposure (e.g., level, duration, frequency, or probability) during an etiologically relevant time window of exposure, and an adequate range of exposure levels or duration to evaluate exposure-response relationships. For example, the studies that found a positive association with trichloroethylene and kidney cancer were studies of highly exposed workers (NTP, 2015a). The validity of the outcome ascertainment methods, or the ability of the method to accurately differentiate between "diseased" and "non-diseased", is also important. For cancers with high survival rates, use of cancer mortality, for example from death certificate data, may be an insensitive outcome measure. Similar issues apply to other kinds of outcomes. Adequate length of follow-up in cohort studies for a specific endpoint should also be considered, and will differ depending on the specific exposure, outcome, and potential mechanisms in play. For example, the optimal latency for mesothelioma in relation to asbestos exposure may differ from that of lung cancer, and the optimal latency for breast cancer may differ for an estrogenic chemical than for a chemical acting through a different mode of action. The number of exposed cases is especially important in the evaluation of rare cancers in cohort studies (such as nasopharyngeal cancer, which is linked to formaldehyde exposure) or rare exposures in population-based case-control studies. Although a systematic review may consider study size in the evaluation of precision of the study, this may not be possible when there are no observed cases in a cohort study, or no exposed cases or controls in a case-control study (Fu et al., 2011), or when it is not possible to develop a combined effect estimate across multiple studies. These factors should be evaluated in a systematic manner with the same rigor as the evaluation for the potential for biases and confounding.

Dilution of risk estimates comparing exposed and referent groups, with a reduction in sensitivity, can arise when there is a great deal of variation in the probability, frequency and level of exposure in the group defined as exposed. This issue was specifically noted in a 2006 NRC report on key scientific issues for assessing the health effects of trichloroethylene. The committee noted the potential for dilution of the risk estimate when effect estimates are calculated for "ever exposed" in studies with large numbers of individuals with low levels of exposure (e.g., based on average or cumulative exposure measures) (NRC, 2006). Other examples of this type of low sensitivity can be found in studies of lung cancer and asbestos (Marsh et al., 2001), cause-specific mortality in a polyvinylchloride manufacturing cohort and a lead smelting cohort (Parodi et al., 2007), and pregnancy outcomes in relation to formaldehyde exposure among nursing personnel in surgical departments or sterilization units in general hospitals (Hemminki et al., 1982; Hemminki et al., 1985).

Sensitivity is also important in the consideration of animal studies. As with the epidemiology studies, important aspects include the exposure duration and levels, assessing a relevant time window of exposure, and the appropriate timing of endpoint assessment. For example, the most informative cancer bioassays are generally those that expose and observe animals for as long as possible without introducing end-of-life health complications (e.g., a 2-year bioassay), with shorter assays drawing into question the reliability of null findings. This may not always be the most sensitive study protocol, however. While studies of arsenic carcinogenicity in adult animals did not reveal substantial effects, more recent studies of exposure to arsenic or its metabolites suggested that gestational and early postnatal exposure may be a time of particular sensitivity in terms of carcinogenesis (IARC, 2012).

The reliability, specificity, and validity of the endpoint ascertainment in animal studies also requires systematic evaluation, some features of which are routinely covered during outcome assessment in risk-of-bias approaches (e.g., evaluating blinding of outcome assessors; ensuring consistent application of protocols across groups). Other features, such as whether the endpoint was measured at animal ages during which the endpoint being tested was sensitive to change (e.g., based on known biological maturation of the organ or function in question) or whether the specific endpoint evaluation protocol employed might be more or less sensitive for detecting changes in the endpoint being evaluated, may not be adequately addressed. Validation of the non-standard assays that are often the only data available for environmental health assessments through the use of positive and negative controls may be necessary to ensure that the assay can appropriately detect the effect under study. It is also important to consider the specificity of the assay protocols for measuring the outcome of interest. For example, while routine histopathology of all organs at necropsy may be capable of detecting overt damage to the tissue of interest, a more specific evaluation of the target tissue using stereological methods or histopathology from interim sacrifice data for age-related pathological lesions may be able to detect effects that would otherwise be missed.

For endocrine disrupting chemicals, sensitivity to endocrine disruption is highest during tissue development (UNEP-WHO, 2013). Thus, the sensitivity of a study would be reduced by a study design that does not include exposure during the developmental period, does not include the length of follow-up needed to assess latent developmental effects, or for some endpoints, such as pubertal development, examines effects at a time that is too late to detect effects on early maturation. Within the context of neurodevelopment, functional maturation of the nervous system continues through puberty, with an age-dependency to the development and variability of different behavioral functions (Semple et al., 2013; Rice and Barone, 2000). Testing for effects of exposure prior to the full maturation of the specific behavior in question could make it difficult to detect effects. While endpoint timing

Please cite this article as: Cooper, G.S., et al., Study sensitivity: Evaluating the ability to detect effects in systematic reviews of chemical exposures, Environ Int (2016), http://dx.doi.org/10.1016/j.envint.2016.03.017