Regulatory Toxicology and Pharmacology 76 (2016) 187–198

Contents lists available at ScienceDirect

Regulatory Toxicology and Pharmacology

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

Commentary

Systematic comparison of study quality criteria

Heather N. Lynch [a], Julie E. Goodman [a], Jade A. Tabony [b], Lorenz R. Rhomberg [a, *]

[a] Gradient, 20 University Road, Cambridge, MA 02138, USA
[b] Gradient, 600 Stewart Street, Suite 1900, Seattle, WA 98101, USA

ARTICLE INFO

Article history:
Received 18 December 2015
Accepted 25 December 2015
Available online 29 December 2015

Keywords:
Study quality
Systematic review
Evidence integration

ABSTRACT

Approaches for the systematic review and evaluation of chemical toxicity are currently being reconsidered, with a specific focus on the evaluation of individual studies and their integration into the overall body of evidence. This renewed interest has arisen, in part, as a result of several prominent reviews of these approaches by special committees of the National Research Council (NRC), among others. We conducted a critical evaluation of several available frameworks for evaluating study quality. We assessed the criteria separately for human, animal, and *in vitro* studies as well as for systematic reviews. We then evaluated commonalities across disciplines. We also considered the potential implications of applying criteria frameworks and how they bear on fundamental risk assessment questions. We found that the available frameworks within each discipline differed in terms of their intended purpose and level of guidance for decision making. All the frameworks across disciplines shared common themes, however, including the adequate reporting of specific details of study conditions and design/protocol, selection and randomization of study groups (where applicable), outcome assessment methods and applicability (*e.g.*, validity and reliability), avoidance of selective reporting, and the consideration of potential confounders or bias. We identified the most informative study quality considerations, which will enable researchers to implement more objective and standardized methods for evaluating studies and, ultimately, improve risk assessment methods.

1. Introduction

1.1. Study quality and risk of bias

Approaches to the systematic review and evaluation of the toxicity of chemicals are currently being reconsidered. This has arisen, in part, as a result of several prominent reviews by, among others, special committees of the National Research Council (NRC). The United States Environmental Protection Agency (US EPA) has undertaken a process to reform its methods for assessments under its Integrated Risk Information System (IRIS) and seeks input and advice from stakeholders. A key element of these process reforms – one that can be applied routinely and similarly across the studies under consideration as well as across assessments – is the identified need for objective and standardized methods for evaluating the "quality" of individual studies. The aim is to provide a consistent and objective system for bringing study strengths and shortcomings to bear on their evaluation as evidence of potential toxic effects.

Exactly what is meant by a study's "quality" has not been defined precisely, or, at least, it has not been defined in exactly the same way in different discussions of the issue. But the general sense is that a review should include an assessment of the study design's soundness for its immediate intended purposes; the adequacy of precautions taken to avoid potential impacts on the results from unintended or uncontrolled causes (that is, influences aside from the one being tested); the care taken to execute the protocol correctly; and the adequacy and completeness of the analysis as well as documentation of the conduct and the results. All of these factors ultimately bear on whether the results of the evaluated study should be regarded as reliable and unambiguous for the interpretation of causal associations between the substance and the outcome of interest.

The term "risk of bias" has been applied to some, but not all, of the evaluation criteria for data quality. Because it is a new term, it appears that different discussants are assigning it varying meanings, and it is important to understand the intended implications in those discussants' particular uses of the term. The use of the word "bias" implies that the concern is for sources of possible systematic

* Corresponding author.
E-mail address: lrhomberg@gradientcorp.com (L.R. Rhomberg).

http://dx.doi.org/10.1016/j.yrtph.2015.12.017

or directional error, as opposed to mere imprecision that results from statistical fluctuation (for example, because of small numbers of animals investigated or because of difficulty maintaining air concentration precisely at its target level). The use of the term "risk" implies that the issue is the *potential* for directional bias, as opposed to the determination that the biasing factor has indeed affected the results. That is, what is assessed is the sufficiency of steps to eliminate or control the influence of the biasing factor, rather than an assessment of the degree to which particular experimental results might have, in fact, been skewed.

The NRC (2014) review of the IRIS process makes the distinction as follows:

> The committee notes that assessing the quality of the study is **not equivalent** to assessing the risk of bias in the study. An assessment of study quality evaluates the extent to which the researchers conducted their research to the highest possible standards and how a study is reported. Risk of bias is related to the internal validity of a study and reflects study-design characteristics that can introduce a systematic error (or deviation from the true effect) that might affect the magnitude and even the direction of the apparent effect. [emphasis added]

In their guidance on systematic review (which includes details on conducting study quality assessments), the National Toxicology Program's (NTP) Office of Health Assessment and Translation (OHAT) echoed the distinction between imprecision and systematic error but went somewhat farther, by suggesting that the systematic error potential noted in risk of bias is not just related to, but is, in fact, equivalent to the evaluation of a study's internal validity (NTP, 2015a). Note, however, that OHAT's preliminary guidance, issued in 2013, was less clear about the distinction between study quality and "risk of bias," seemingly using the terms interchangeably (NTP, 2013a). Only in the first applications of their risk of bias approach (*e.g.*, the assessment of bisphenol A; NTP, 2013b) did OHAT begin to specifically refer to "risk of bias" as a term that is synonymous with internal validity.

The concepts of risk of bias and study quality are nonetheless related. Procedures to ensure quality, such as the use of standardized methods, quality control procedures, and transparent reporting, are put in place to minimize the possibility of introducing sources of bias. Standard protocols (such as Organisation for Economic Co-operation and Development [OECD] guidelines for specific test types) are developed primarily to mandate steps to avoid known pitfalls and to avoid inadvertent introduction of unknown extraneous factors into an untested aspect of study design. Even unbiased random error is unbiased only in the long run, over multiple iterations of an experiment or study. In any one experiment, however, the random fluctuations can, by chance, be skewed away from the true mean values. If quality and bias measures are to be applied to individual studies, as many of these criteria systems intend, there is not a very clear distinction between study-specific random skewing and more systematic skewing potential. If the *possibility* of systematic skewing is at issue, the question still remains whether the factor actually affected the results of an individual study and, if so, how much and in what direction. Frequently, the direction and magnitude of skewing may be unknown. Moreover, it would seem that if one examines two studies with identical risk-of-bias profiles, except that one has a markedly larger number of animals or number of distinct dose groups, then this measure of quality (and not of bias) would affect the perceived dependability of the measured results as an index of the true causative relationship being investigated – that is, it would affect internal validity.

Another caveat about the use of the term "risk of bias" is that there are some aspects of bias that operate at the level of collections of studies – most notably, publication bias. For a study for which all the potential extraneous factors have been controlled, most of the evaluation systems will concur about its internal validity. But the choice of what studies to publish, what results of those studies to feature and document in the publications, which among several alternative analysis processes to pursue, whether those manuscripts are accepted for publication, and even what studies to undertake in the first place all can affect and bias the array of outcomes available in the literature, even if each individual study result that is reported is objective and reliable. Care must be taken in projecting the evaluations of quality or bias for individual studies into a characterization of the overall reliability of the body of studies collectively.

1.2. Methods

We conducted an evaluation of several of the available study quality criteria systems (Table 1), pulling out and systematically comparing the specific criteria that the systems require evaluators to examine for each rated study. The intent is that this analysis will help further the general discussion about available study quality evaluation systems and the features that should be adopted in any system that might be established as part of the larger risk assessment methods improvement effort. Because the types of studies considered (*i.e.*, animal, *in vitro*, human, and systematic reviews) have different designs – and, thus, different considerations for quality – we have evaluated these types of evidence separately and have summarized them in four separate tables (Tables 3–6).

We assessed the following systems for evaluating the quality of studies or systematic reviews: the Klimisch system (Klimisch et al., 1997); OECD Guidance Document (GD) 34 (OECD, 2005); the Toxicological Data Reliability Assessment Tool (ToxRTool) (European Commission, Undated); the approaches that have been used in recent IRIS systematic review documents (US EPA, 2013), notably the recent risk-of-bias evaluations for inorganic arsenic (US EPA, 2014a)[1]; the framework being developed by NTP's OHAT (NTP, 2013a, 2015a,b), as applied to the risk-of-bias assessment of perfluorinated compounds (NTP, 2013b,c); Animal Research: Reporting of *In Vivo* Experiments (ARRIVE) guidelines for animal research (Kilkenny et al., 2010); the Navigation Guide for systematic reviews (Koustas et al., 2013, 2014; Woodruff and Sutton, 2014; Johnson et al., 2014; Lam et al., 2014), the "assessment of multiple systematic reviews" (AMSTAR) system (Shea et al., 2007); the "strengthening the reporting of observational studies in epidemiology" (STROBE) system (von Elm et al., 2007a,b,c,d,e); and the Systematic Approach for Scoring Human Data, as developed by Money et al. (2013). The types of studies (*i.e.*, human, animal, *in vitro*, systematic reviews) addressed in each system are summarized in Table 1. The specific criteria for each system are summarized in four tables, one each for: human studies (Table 3), animal studies (Table 4), *in vitro* studies (Table 5), and systematic reviews (Table 6). Several of the criteria systems cover more than one of the study types. Although these 10 particular criteria systems do not exhaust the list of those that have been put forth, they provide a good overview of the main systems aimed at broad usage. Overall, the criteria systems differ in their purpose and specific recommendations – some only provide suggestions for information that should be reported by study authors, others suggest specific criteria that should be fulfilled by study authors, and others provide rating scales for use to assess the relative level of quality for a study based on the scores it receives. The entries used in the tables correspond to each of these

[1] Note that, at this time, there is no comprehensive guidance document for the IRIS risk of bias criteria system.