SOT | Society of Toxicology
www.toxsci.oxfordjournals.org

TOXICOLOGICAL SCIENCES, 151(2), 2016, 206–213

doi: 10.1093/toxsci/kfw056
Forum Article

FORUM ARTICLE

Does GLP enhance the quality of toxicological evidence for regulatory decisions?

Christopher J. Borgert,[*,1] Richard A. Becker,[†] Betsy D. Carlton,[‡] Mark Hanson,[§] Patricia L. Kwiatkowski,[¶] Mary Sue Marty,[‖] Lynn S. McCarty,[|||] Terry F. Quill,[||||] Keith Solomon,[#] Glen Van Der Kraak,[**] Raphael J. Witorsch,[††] Kun Don Yi[‡‡]

[*]Dept. Physiol. Sciences, Univ. FL College of Veterinary Medicine, Applied Pharmacology and Toxicology, Inc, and C.E.H.T, Gainesville, Florida 32605; [†]American Chemistry Council, Washington, District of Columbia 20002; [‡]Bluestar Silicones USA Corp 10520 Whitestone Rd., Raleigh, North Carolina 27615, [§]Department of Environment and Geography, University of Manitoba, Winnipeg, Manitoba, Canada, [¶]Bayer CropScience, Research Triangle Park, North Carolina 27709, [‖]Toxicology and Environmental Research and Consulting, The Dow Chemical Company, Midland, Michigan 48674, [|||]L.S. McCarty Scientific Research & Consulting, Newmarket, Ontario L3X 3E2, Canada, [||||]Quill Law Group, LLC, Washington, District of Columbia 20006; [#]University of Guelph, Centre for Toxicology, School of Environmental Science, Guelph, Ontario N1G 2W1, Canada; [**]Department of Integrative Biology, University of Guelph, Guelph, Ontario N1G 2W1, Canada, [††]Department of Physiology & Biophysics, School of Medicine, Virginia Commonwealth University, Richmond, Virginia 23298-0551; and [‡‡]Syngenta Crop Protection Inc, Greensboro, North Carolina 27419-8300

[1]To whom correspondence should be addressed. Fax: (352) 335-8242. E-mail: Cjborgert@apt-Pharmatox.Com.

ABSTRACT

There is debate over whether the requirements of GLP are appropriate standards for evaluating the quality of toxicological data used to formulate regulations. A group promoting the importance of non-monotonic dose responses for endocrine disruptors contend that scoring systems giving primacy to GLP are biased against non-GLP studies from the literature and are merely record-keeping exercises to prevent fraudulent reporting of data from non-published guideline toxicology studies. They argue that guideline studies often employ insensitive species and outdated methods, and ignore the perspectives of subject-matter experts in endocrine disruption, who should be the sole arbiters of data quality. We believe regulatory agencies should use both non-GLP and GLP studies, that GLP requirements assure fundamental tenets of study integrity not typically addressed by journal peer-review, and that use of standardized test guidelines and GLP promotes consistency, reliability, comparability, and harmonization of various types of studies used by regulatory agencies worldwide. This debate suffers two impediments to progress: a conflation of different phases of study interpretation and levels of data validity, and a misleading characterization of many essential components of GLP and regulatory toxicology. Herein we provide clarifications critical for removing those impediments.

Key words: safety evaluation; risk assessment; endocrine toxicology; data quality; evaluation, good laboratory practice

BACKGROUND

There has been ongoing debate in toxicology over whether the requirements of GLP are appropriate standards for evaluating the quality of data used to formulate regulations (Becker *et al.*, 2009; Myers *et al.*, 2009; Tyl, 2009). One side asserts that GLP studies may not be reliable and that agencies should not use GLP and scoring systems that give it primacy as criteria for selecting data used in regulatory decision-making (eg, Myers *et al.*, 2009). The other side argues that both GLP and non-GLP studies provide important, albeit sometimes incomplete information, and that all data should be evaluated and weighed objectively for relevance and reliability using a transparent weight of evidence approach (eg, McCarty *et al.*, 2012; Rhomberg *et al.*, 2013). The argument against GLP has again been raised in a recent commentary (Zoeller and Vandenberg, 2015) on assessing dose response relationships for endocrine disrupting chemicals (EDCs) wherein the authors proffer 'evidence-based logical arguments that (1) the use of the Klimisch score (Klimisch *et al.*, 1997) should be abandoned for assessing study quality; (2) evaluating study quality requires experts in the specific field; and (3) an understanding of mechanisms should not be required to accept observable, statistically valid phenomena.' These arguments were advanced to correct what the authors consider to be deficiencies in an approach proposed by Lagarde *et al.* (2015) wherein Klimisch scoring is used and the biological plausibility of a non-monotonic dose response relationship is assessed. Although there is value in discussing how GLP can be improved to better ensure reliability, transparency, and objectivity, we have two serious concerns about Zoeller and Vandenberg's proposals: (1) some of their arguments mischaracterize GLP and the process for evaluating study quality and acceptability for regulatory decision-making, and (2) their criticisms conflate issues that rightfully pertain to different levels of data validity and study interpretation.

To correct those mischaracterizations and to place the issues into proper context, it is necessary to appreciate that interpretation of a scientific study can be considered as occurring in at least three phases. Phase I involves basic study integrity. It requires knowledge of how precisely the procedures were carried out, the integrity of the test system, and the identity and purity of the test substances and positive and negative controls. Phase I would include an evaluation of Borgert *et al.*'s (2011) 'primary validity' of the data, which requires verification of the authenticity and precision of the measurements themselves, proper control of the conditions under which they are measured, and assurance of the replicability of the technique. Phase II focuses on the quality of the study design, the observed results and their variability, and differences relative to controls that are attributable to the test article. Evaluation of Phases I and II requires secondary validity of the data (Borgert *et al.*, 2011), which relates to thorough verification that all required measurements were made and adequately described and reported, including the methodologies used for analysis. Phase III considers the implications of the results, particularly with respect to their ability to inform causality (referred to as 'tertiary validity' per Borgert *et al.*, 2011), and their applicability to the species of concern, which is often best informed by understanding the toxicological mode of action (reviewed in Borgert *et al.*, 2015).

GLP intends to address primarily Phase I and some aspects of Phase II, whereas OECD and EPA test guidelines address primarily Phase II. Thus, guideline studies conducted under GLP focus on Phase I and II concerns, but not issues addressed by Phase III. In contrast, the peer-review process focuses primarily on Phase III, and, depending upon the detail required by the particular journal, also Phase II. Consequently, non-GLP studies published in the open literature leave Phase I concerns largely unaddressed, but are strong in considering Phase III, and can provide assurance for Phase II concerns when full data are made available. McCarty *et al.*, (2012) have described the various gaps within journal peer-review and GLP and note a movement toward converging requirements within each sphere. In short, the strengths of GLP in addressing Phase I and II does not render it superior to peer-review for addressing Phase III, and vice-versa.

SHOULD KLIMISCH SCORES BE ABANDONED FOR ASSESSING STUDY QUALITY?

Klimisch *et al.* (1997) presented a systematic approach for evaluating the quality of data in terms of its reliability, relevance, and adequacy for use in hazard and risk assessment. They defined distinct categories and proposed code numbers corresponding to specific descriptors of study reliability that have since become known as 'Klimisch scores.' The approach, which has been adopted (OECD, 2004) and used by all 34 OECD member countries, favors studies with a high level of documentation regarding methodology, test procedures, and analytical measurements, and hence, tends to confer high scores on guideline toxicology studies conducted according to GLP or GLP-like standards.

Zoeller and Vandenberg (2015) argue that, in the case of evaluating data quality in studies of endocrine disruption, a few criteria proposed by some endocrinologists and environmental health scientists and published in the peer-reviewed literature should be used rather than Klimisch scores. Their proposed criteria are based on opinions of certain experts concerning the use of appropriate negative and positive controls, the use of sensitive animal species and strains, and the use of appropriate endpoints. Their criteria do not address Phase I study interpretation, nor any component of primary validity of the data, which the current Directors of NIEHS and ATSDR testified should be the cornerstone of regulatory science in the U.S. (The Environment and Human Health, 2010). Their criteria touch on some aspects of Phase II and III study interpretation; however, both Klimisch and GLP address the same issues, such as use of appropriate positive and negative controls to provide information about the reliability of the particular test model. As well, the majority of EPA's standardized test guidelines for endocrine screening assays require demonstration of proficiency and/or inclusion of positive and/or negative controls (US EPA, 2009). Zoeller and Vandenberg's preference for expert opinion over Klimisch relies on a selection of literature lacking recent publications that address evaluation of data quality, including aspects touched upon by Zoeller and Vandenberg. The broader literature thoroughly addresses problems common to evaluations of data quality, including deficiencies in the peer-reviewed literature, and provides methodologies for evaluating data quality, study reliability, and replicability that are applicable within and beyond the narrow topic of endocrine disruption (eg, reviewed in McCarty *et al.*, 2012); as well, it outlines refinements and new methods that extend the applicability of Klimisch categories to peer-reviewed publications (ECETOC, 2006; Moermond *et al.*, 2016; Schneider *et al.*, 2009; Segal *et al.*, 2015).

Zoeller and Vandenberg contend that Klimisch scores place too much confidence in conformity to GLP, which, they assert, only applies to record-keeping and reporting requirements in guideline studies. This is a misunderstanding of both Klimisch and GLP. Although the Klimisch scoring system does give the highest ranking to GLP-compliant studies, it explicitly includes both non-guideline studies and studies that are not GLP-compliant, especially in the regulatory context. As well, a review of regulatory methods for evaluating data quality makes it evident that non-guideline studies from the open literature are used routinely (Becker et.al., 2009; EFSA, 2011; US EPA, 2011, 2012). In a recent publication, a group of authors stated: 'The point made by Zoeller *et al.* that nonguideline studies are not considered in risk assessments is not valid. The US EPAs IRIS assessments, REACH dossiers, EFSA evaluations, and most others risk assessments strive to use all available data including mechanistically oriented nonguideline studies. An assessment of reliability of studies, consistency of the database, and a weight-of- evidence approach in the evaluation of inconsistent databases (EFSA, 2015) is established in hazard and risk assessment world-wide and was specifically embraced by the WHO/UNEP report on "endocrine disruption" in 2002' (Autrup et. al., 2015).

Zoeller and Vandenberg further imply that the high cost and personnel-intensive requirements of GLP are necessary only for industry-funded studies, which they assume are unpublished, to protect against fraudulent data submissions. This is not correct. Because of the high-cost of GLP certification requirements and stringent Quality Assurance/Quality Control standards, GLP-compliant studies have tended to be carried out in large, well-staffed and equipped laboratories. The studies have tended to be funded by industry, as well as by governmental agencies and others who have the means to support GLP studies, eg, the National Toxicology Program. Moreover, industry-funded guideline toxicity studies are quite often published in the peer-reviewed literature. For example, using only 'reproduction' OR 'carcinogenicity,' as search terms, PubMed identifies many guideline and non-guideline toxicology studies sponsored or authored by scientists affiliated with industry (see Supplemental Materials). Klimisch et al. (1997) and regulatory agencies have placed a high value on study reports that include sufficient detail to allow reanalysis of data to independently confirm results and support additional analysis using alternative methods of data evaluation. Until recently (eg, supplemental data published online; Hanson *et al.*, 2011), however, the abbreviated format common to scientific journal articles lacked a mechanism for inclusion of the hundreds of pages of raw data included in GLP study reports and still lack a reliable means of assuring data protection.

Although GLP was implemented as legal regulations in the US in 1979 as a result of misconduct and fraud by contract research organizations that submitted test data to federal agencies (Baldeshwiler, 2003), transparent recommendations from the larger scientific community have developed it into a well-understood standard that is accepted by regulatory agencies around the globe. The intent of GLP requirements is to ensure consistency, reliability, comparability, and harmonization in the conduct of various types of studies around the world. Scientific fraud is unfortunate and damaging in any setting, regardless of the source, but data quality, reliability, reproducibility, and fraud have been recently highlighted as major areas of concern in the scientific community in general (eg, Nature, 2013, 2014; Science, 2014), despite having been successfully addressed decades ago within the regulatory arena by adoption of GLP. To imply that GLP requirements are necessary only for industry-funded studies ignores the considerable inadequacies that are the topic of an entire body of literature focused on correcting and improving scientific and journal peer-review processes. It also ignores the considerable costs that have ensued following the publication of fraudulent academic research. This may directly comprise only 1%–2% of the NIH research budget but further entails significant indirect tangible and intangible costs (Scientific American, 2014). Full documentation and disclosure of the details of conducting a study and generating and recording the data are critical for evaluating the quality of the data, identifying controlled versus uncontrolled variables, and, if present, detecting fraudulent data; therefore, these practices have value for all scientific studies. Consequently, many top scientific journals are implementing GLP-like reporting standards in the peer-review process to enhance its effectiveness in identifying scientific fraud (eg, academic studies from Arnold *et al.*, 1996; McLachlan, 1997; Online Universities, 2012), irreproducible data (eg, Prinz *et al.*, 2011) and reporting of associations that fail to meet minimum statistical requirements (eg, Bosker *et al.*, 2013). These and other issues and perspectives on information quality in regulatory decision-making and in the peer-review process are discussed in a recent review (McCarty *et al.*, 2012) not cited by Zoeller and Vandenberg (2015).

Zoeller and Vandenberg suggest that by assigning high scores to GLP-compliant studies, the Klimisch scoring system mistakenly equates quality in record keeping and reporting with quality in study design and execution. Their point would be valid if GLP-compliance was purely a matter of recording and reporting, but it is not. GLP requires justification of the test system and procedures; training certification and documentation for investigators and technicians involved in each scientific procedure; the measurement of a comprehensive set of study parameters, including: analytical characterization of test materials; analytical verification of concentration, stability, and homogeneity of dosing solutions; validation and calibration of instruments; adherence to standard operating procedures in conducting experimental activities; independent auditing of study procedures and data; adherence to animal welfare requirements; measurement of laboratory and animal room conditions, among many other important assurances of experimental quality. The goal of these requirements is to ensure that critical study parameters—parameters that go to the core of study quality—are, in fact, measured. Furthermore, GLP requires that a study protocol be selected and justified a priori, that any deviations from that protocol occurring during the study are documented with appropriate written explanation or justification, and that these elements are audited by an independent party before finalization of the study report. Notably, the National Toxicology Program requires that studies be conducted in compliance with GLP and, as such, employ standard operating procedures (NTP, 2011): 'The review and revision of Standard Operating Procedures (SOPs) are a continuing process. Along with the protocol, SOPs are considered essential to the successful conduct, documentation, inspection, and auditing of a study.' Although such transparency is seldom provided in non-GLP studies, it is difficult to argue against these requirements due to the credibility they bring to decision-making (Schreider *et al.*, 2010). Regulatory agencies have found such requirements to be effective in enhancing reliability and reproducibility of studies. In particular, GLP specifically meets their requirements for carefully recorded and audited data covering relevant exposure routes and doses. Table 1, provides a comparison of GLP requirements to the criteria proposed by Zoeller and Vandenberg (2015).

No scientific study is perfect, and the more complex the study, the greater the likelihood that uncontrolled variables may affect the results. The combination of conducting a standardized OECD or EPA test guideline in accordance with GLP reduces the likelihood of systematic errors. However, as Zoeller and Vandenberg correctly point out, GLP compliant studies do not ensure that test substance contamination of the untreated control group will not occur; nevertheless, GLP does ensure that this contamination will be detected and reported. Obviously, the status of the untreated control group is unknown in studies that do not analytically characterize dose solutions or diets. Although GLP-compliance does not provide 100% assurance that all tissues will be dissected appropriately, it does help to ensure that dissections will be done consistently within the same laboratory due to adherence to detailed, step-by-step, standard operating procedures and documented training of laboratory personnel. Furthermore, archiving requirements of GLP enable laboratories to compare control values with previously conducted studies to ensure that the control values are representative of the specific species/strain/age of animals. In addition, guideline-required histopathology can verify the normal appearance of control tissues. In addition to transparency in reporting and documentation, GLP allows for full evaluation of the strengths and weaknesses of a study so that these points can be considered in interpretation of the data. This is not always the case for non-GLP studies and most studies published in the peer-reviewed literature, where such strengths can only be assumed.

Zoeller and Vandenberg argue that Klimisch scoring misses the problem of using insensitive species in EDC studies and inadequate study designs employing too few dose levels. However, GLP requires justification of the test system within the study protocol and Klimisch *et al.* (1997) recommends evaluating the appropriateness of the test system for its relevance to the hazard identification or risk characterization under consideration. Such justification and evaluation requires an understanding of the sensitivity of the test species to the target effects of the test, and the adequacy of the doses selected to allow responses to be characterized (eg, Van Der Kraak *et al.*, 2014). Justification for the species used for specific toxicity tests and the doses tested are routine in GLP-compliant study reports, as are corresponding aspects of validation, including verification of animal strains, purity of cell lines, and responsiveness of models. Species sensitivity is but one consideration, and whereas it should not be ignored, neither should it be given preeminence over more fundamental aspects of validity. Zoeller and Vandenberg's argument focuses selectively on using the most sensitive species, ignoring the fact that the most sensitive species and/or endpoint may be the least relevant for regulatory purposes. Moreover, issues of validity of measurements are specifically addressed in the broader literature on evaluation of data quality, which Zoeller and Vandenberg do not address. For example, accuracy, precision, replication, and potential confounding by extraneous variables (eg, insensitivity of the species or strain) are addressed comprehensively under the evaluation of primary validity of the data in the method of Borgert *et al.*, (2011; Supplemental Material), and would lead to a reduction in the quality assigned to studies confounded by the use of insensitive species or endpoints.

For regulatory purposes, the desire to use the most recent scientific methods must be balanced against the need to evaluate endpoints whose reliability and relevance has been validated (McCarty *et al.*, 2012). As well, sufficiency of study design, addressed specifically under the evaluation of tertiary validity of the data in the schema described in Borgert *et al.* (2011), is much more thorough than merely an evaluation of dose selection as mentioned by Zoeller and Vandenberg (2015). Tertiary validity includes evaluating whether the endpoints measured are the most probative for the hypotheses or types of toxicity tested and whether counterfactual methods are employed to verify relationships between dependent and independent variables (Borgert *et al.*, 2011). The importance of counterfactual methods for conclusions used in risk assessment is recognized generally (Meek *et al.*, 2014). With respect to development and revision of standardized test guidelines, OECD has developed, and follows, specific guidance for establishing and evaluating the validity of a test method which entails establishing the reliability and relevance of a particular test for a specific purpose (OECD, 2005). This discussion underscores our contention that failing to discern the different phases of study interpretation and different levels of data validity leads to criticizing one aspect—in this case, Klimisch scoring—for failing to address issues covered specifically by other processes, eg, by the development of formalized study guidelines or by the risk assessment process, both of which are integral components of regulatory toxicology.

CAN ONLY SUBJECT–MATTER EXPERTS ASSESS STUDY QUALITY?

Zoeller and Vandenberg charge that GLP and Klimisch is not a scientific strategy but rather a bureaucratic one that allows non-experts to evaluate data quality and exclude data from further consideration. First, as stated directly in Klimisch *et al.* (1997), '. . .it is not the intention of this procedure to automatically exclude all unreliable data from further consideration by experts in risk assessment. . .studies with higher reliability should have greater weight for being used in risk assessment.' Moreover, their criticism is misdirected because the GLP-compliant guideline studies favored by Klimisch are designed specifically by subject-matter experts from a variety of disciplines. Their assertion that only a small group of scientists active in publishing on endocrine disruption should have the authority to make pronouncements about data quality specific to endocrine systems ignores the interdisciplinary nature of complex toxicology tests that seek to evaluate all aspects of a general system, as do bioassays for reproductive and developmental toxicity. The development of guideline reproductive and developmental studies specifically utilizes input from subject matter experts in endocrinology, physiology, toxicology, pharmacology, and risk assessment. Hence, assessing study quality indeed requires subject-matter experts, but not only subject-matter experts from a single area of research. Finally, we would also point out that certain aspects of data quality, such as adequate recording and reporting of data, are so fundamental to the pursuit of scientific enquiry that their absence must lower confidence in any set of data in any field.

SHOULD MECHANISTIC UNDERSTANDING BE CONSIDERED?

If Lagarde *et al.* (2015) had proposed that a mechanistic understanding were required in order to accept non-monotonic dose responses, then we would agree with Zoeller and Vandenberg's third criterion. However, Lagarde *et al.* (2015) did not require a mechanistic understanding, but rather proposed using a variety of factors to evaluate the biological plausibility that statistically significant differences in dose-response data reflect a true non-monotonic dose-response relationship. The topic of biological plausibility is more complex than can be covered adequately here; but in general, it asks whether an effect of a chemical is consistent with the state of biological, physiological, biochemical, bio-physical, and pharmacokinetic knowledge. The more that is known about a chemical, its behavior within and upon organisms and the related biology, physiology, and pharmacokinetics, the finer the distinction that can be made regarding the qualitative and quantitative nature of its putative effects. Regulatory evaluations increasingly probe the consistency and biological plausibility that an effect is related to the test substance and is of a certain qualitative (eg, mutagenic vs teratogenic) or quantitative (eg, linear vs non-linear) type (Meek *et al.*, 2014 and references therein). Biological plausibility is not limited to mechanistic understanding, but includes the possible physiological modes of action. A mode of action identifies the key events in producing an adverse effect, but does not entail an elucidation of the full 'mechanism' (Meek *et al.*, 2014; Boobis *et al.*, 2008; Borgert *et al.* 2004; Butterworth *et al.*, 1995; Dellarco and and Wiltse, 1998; Schlosser and Bogdanffy, 1999). Lagarde *et al.* propose evaluating possible molecular mechanisms as well as possible physiological modes of action that can produce non-monotonic dose responses to verify apparent non-monotonic dose-response relationships. This proposal is well reasoned and consistent with internationally accepted methodologies used to discern whether statistical significance aligns with biological relevance.

CONCLUSIONS

In summary, we find that the debate in toxicology over whether the requirements of GLP are appropriate standards for evaluating the quality of data used to formulate regulations suffers two impediments to progress. First, there has been a conflation of different phases of study interpretation and levels of data validity. This can be resolved by appreciating that different components of regulatory toxicology address different aspects of study integrity. Second, there has been a misleading characterization of the issue of as an either/or choice between three criteria, proffered for application to a narrow area of science on the one hand, and, on the other, a mischaracterized view of a well-established, well-documented, extensively used regulatory science method that has seen more than 25 years of open international development and refinement, which is documented in a substantial body of peer-reviewed evaluation and analysis. This is troublesome because the proffering of those criteria ignores literature that obviates the conflict asserted and selectively discusses only the literature that appears to support them. Such an approach is not evidence-based or logical. To improve the quality of scientific debate, we suggest that argumentation advanced in favor of a particular perspective should be judged by the thoroughness with which it addresses and properly cites the relevant literature, both pro and con. Scientists from all sectors, regardless of affiliation or funding source, should support the use of objective criteria for determining data quality and study reliability, as well as procedures to systematically integrate evidence from all relevant studies, both GLP and non-GLP.

SUPPLEMENTARY DATA

Supplementary data are available online at http://toxsci.oxfordjournals.org/.

ACKNOWLEDGMENTS AND DISCLOSURES

The authors thank Dr. James S. Bus of E^{x}ponent® for helpful suggestions regarding GLP protocols. The authors declare no competing interests in the subject matter of this submission. In the course of their employment, the authors may conduct, direct, oversee, or interpret studies that are required by statute to comply with GLP, however, their employment status, compensation, advancement, and other benefits do not depend in any way upon the requirement for compliance with or use of GLP. CJB received compensation from the Endocrine Policy Forum for the initial drafts.

REFERENCES

Arnold, S. F., Klotz, D. M., Collins, B. M., Vonier, P. M., Guillette, L. J., and McLachlan, J. A. (1996). Synergistic activation of estrogen receptor with combinations of environmental chemicals. *Science* **272**, 1489–1492.

Autrup, H., Barile, F. A., Blaauboer, B. J., Degen, G. H., Dekant, W., Dietrich, D., Domingo, J. L., Gori, G. B., Greim, H., Hengstler, J. G., *et al.* (2015). Principles of pharmacology and toxicology also govern effects of chemicals on the endocrine system. *Toxicol. Sci.* **146**(1): 1–5.

Baldeshwiler, A. M. (2003). History of FDA good laboratory practices. *Qual. Assur. J.* **7**, 157–161.

Becker, R. A., Janus, E. R., White, R. D., Kruszewski, F. H., and Brackett, R. E. (2009). Good laboratory practices and safety assessments. *Environ. Health Perspect.* **117**, A482–A483.

Boobis, A. R., Doe, J. E., Heinrich-Hirsch, B., Meek, M. E., Munn, S., Ruchirawat, M., Schlatter, J., Seed, J., and Vickers, C. (2008). IPCS framework for analyzing the relevance of a noncancer mode of action for humans. *Crit. Rev. Toxicol.* **38**, 87–96.

Borgert, C. J., Mihaich, E. M., Ortego, L. S., Bentley, K. S., Holmes, C. M., Levine, S. L., and Becker, R. A. (2011). Hypothesis-driven weight of evidence framework for evaluating data within the US EPA's endocrine disruptor screening program. *Regul. Toxicol. Pharmacol.* **61**, 185–191.

Borgert, C. J., Quill, T.F., McCarty, L.S., Mason, A.M. (2004). Can mode of action predict mixtures toxicity for risk assessment?. *Toxicol. Appl. Pharmacol.* **201**(2): 85–96.

Bosker, T., Mudge, J. F., and Munkittrick, K. R. (2013). Statistical reporting deficiencies in environmental toxicology. *Environ. Toxicol. Chem.* **32**, 1737–1739.

Butterworth, B. E., Conolly, R. B., Morgan, K. T. (1995). A strategy for establishing mode of action of chemical carcinogens as a guide for approaches to risk assessments. *Cancer Letters.* **93**(1): 129–146.

The Environment and Human Health: HHS' Role. (2010). Hearing before The Subcommittee on Health of the Committee on Energy and Commerce, House of Representatives, One Hundred Eleventh Congress, Second Session, April 22, 2010, Serial No. 111–112, pp. 50–51.

Dellarco, V.L., Wiltse, J.A. (1998). US environmental protection agency s revised guidelines for carcinogen risk assessment: incorporating mode of action data. *Mutat. Res.* **405**: 273–277.

EFSA (European Food Safety Authority). (2011). Submission of scientific peer-reviewed open literature for the approval of pesticide active substances under Regulation (EC) No 1107/2009 (OJ L 309, 24.11.2009, pp. 1–50). EFSA *J.* **9**, 2092. [49 pp.]. doi:10.2903/j.efsa.2011. 2092. Available Online at www.efsa.europa.eu. http://www.efsa.europa.eu/en/efsajournal/pub/2092. Accessed March 30, 2016.

ECETOC (European Centre for Ecotoxicology and Toxicology of Chemicals). (2006). Appendix A: Criteria for Reliability Categories in Synthetic Amorphous Silica. http://members.ecetoc.org/Documents/Document/JACC%20051.pdf. Accessed December 11, 2015.

Hanson, B., Sugden, A., and Alberts, B. (2011). Making data maximally available [Editorial]. *Science* **331**, 649.

Klimisch, H. J., Andreae, M., and Tillmann, U. (1997). A systematic approach for evaluating the quality of experimental toxicological and ecotoxicological data. *Regul. Toxicol. Pharmacol.* **25**, 1–5.

Lagarde, F., Beausoleil, C., Belcher, S. M., Belzunces, L. P., Emond, C., Guerbet, M., and Rousselle, C. (2015). Non-monotonic dose-response relationships and endocrine disruptors: a qualitative method of assessment. *Environ. Health* **14**, 13.

McCarty, L. S., Borgert, C. J., and Mihaich, E. M. (2012). Information quality in regulatory decision-making: peer review versus good laboratory practice. *Environ. Health Perspect.* **120**, 927–934.

McLachlan, J. A. (1997). Retraction: synergistic activation of estrogen receptor with combinations of environmental chemicals. *Science* **277**, 462–463.

Meek, M. E., Boobis, A., Cote, I., Dellarco, V., Fotakis, G., Munn, S., Seed, J., and Vickers, C. (2014). New developments in the evolution and application of the WHO/IPCS framework on mode of action/species concordance analysis. *J. Appl. Toxicol.* **34**, 1–18.

Moermond, C., Kase, R., Korkaric, M., and Agerstrand, M. (2015). CRED: Criteria for reporting and evaluating ecotoxicity data. *Environ. Toxicol. Chem.* doi:10.1002/etc.3259.

Myers, J. P., vom Saal, F. S., Akingbemi, B. T., Arizono, K., Belcher, S., Colborn, T., Chahoud, I., Crain, D. A., Farabollini, F., Guillette, L. J., *et al.* (2009). Why public health agencies cannot depend on good laboratory practices as a criterion for selecting data: The case of bisphenol A. *Environ. Health Perspect.* **117**, 309–315.

Nature. (2014). Journals unite for reproducibility; Consensus on reporting principles aims to improve quality control in biomedical research and encourage public trust in science. Editorial. http://www.nature.com/news/journals-unite-for-reproducibility-1.16259. Accessed June 18, 2015.

Nature. (2013). Announcement: Reducing our irreproducibility. Editorial. http://www.nature.com/news/announcement-reducing-our-irreproducibility-1.12852. Accessed June 18, 2015.

NIH (National Institutes of Health). (2015). Principles and Guidelines for Reporting Preclinical Research http://www.nih.gov/about/reporting-preclinical-research.htm. Accessed June 18, 2015.

NTP. (2011). Specifications for the Conduct of Studies to Evaluate the Toxic and Carcinogenic Potential of Chemical, Biological and Physical Agents in Laboratory Animals for the National Toxicology Program (NTP). http://ntp.niehs.nih.gov/ntp/test_info/finalntp_toxcarspecsjan2011.pdf. Accessed July 16, 2015.

OECD (Organisation for Economic Cooperation and Development). (2005). OECD Series on Testing and Assessment Number 34: Guidance Document on the Validation and International Acceptance of New or Updated Test Methods for Hazard Assessment. http://www.oecd.org/general/searchresults/?q=OECD Series on Testing and Assessment Number 34&cx=012432601748511391518:xzeadub0b0a&cof=FORID:11&ie=UTF-8 Accessed March 30, 2016..

OECD (Organisation for Economic Cooperation and Development). (2004). Manual for Investigation of HPV Chemicals, Chapter 3: Data Evaluation. (http://www.oecd.org/chemicalsafety/risk-assessment/49191960.pdf). Accessed December 20, 2015.

Online Universities. (2012). The 10 Greatest Cases of Fraud in University Research. http://www.onlineuniversities.com/blog/2012/02/the-10-greatest-cases-of-fraud-in-university-research/ Blog, Feb 27, 2012.

Prinz, F., Schlange, T., and Asadullah, K. (2011). Believe it or not: How much can we rely on published data on potential drug targets? *Nat. Rev. Drug Discov.* **10**, 712.

Rhomberg, L. R., Goodman, J. E., Bailey, L. A., Prueitt, R. L., Beck, N. B., Bevan, C., Honeycutt, M., Kaminski, N. E., Paoli, G., Pottenger, L. H., *et al.* (2013). A survey of frameworks for best practices in weight-of-evidence analyses. *Crit. Rev. Toxicol.* **43**, 753–784.

Science. (2014). Journals unite for reproducibility. Editorial. http://www.sciencemag.org/content/346/6210/679.full. Accessed June 18, 2015.

Scientific American. (2014). Dollar Costs of Scientific Misconduct Smaller Than Feared. http://www.scientificamerican.com/article/dollar-costs-of-scientific-misconduct-smaller-than-feared/ August 28, 2014. Accessed March 30, 2016.

Segal, D., Makris, S. L., Kraft, A. D., Bale, A. S., Fox, J., Gilbert, M., Bergfelt, D. R., Raffaele, K. C., Blain, R. B., Fedak, K. M., *et al.* (2015). Evaluation of the toxrtool's ability to rate the reliability of toxicological data for human health hazard assessments. *Regul. Toxicol. Pharmacol.* **72**, 94–101.

Schlosser, P. M., Bogdanffy, M. S. (1999). Determining modes of action for biologically based risk assessments. *Regul Toxicol Pharmacol.* **30**: 75–79.

Schneider, K., Schwarz, M., Burkholder, I., Kopp-Schneider, A., Edler, L., Kinsner-Ovaskainen, A., Hartung, T., and Hoffmann, S. (2009). 'ToxRTool', a new tool to assess the reliability of toxicological data. *Toxicol. Lett.* **189**, 138–144.

Schreider, J., Barrow, C., Birchfield, N., Dearfield, K., Devlin, D., Henry, S., Kramer, M., Schappelle, S., Solomon, K., Weed, D. L., and., *et al.* (2010). Enhancing the credibility of decisions based on scientific conclusions: Transparency is imperative. *Toxicol. Sci.* **116**(1): 5–7.

Science. Journals unite for reproducibility. Editorial. http://www.sciencemag.org/content/346/6210/679.full 2014; Accessed 18 June, 2015.

Scientific American. Dollar Costs of Scientific Misconduct Smaller Than Feared. http://www.scientificamerican.com/article/dollar-costs-of-scientific-misconduct-smaller-than-feared/ August 28, 2014.

Tyl, R. W. (2009). Basic exploratory research versus guideline-compliant studies used for hazard evaluation and risk assessment: bisphenol A as a case study. *Environ. Health Perspect.* **117**, 1644–1651.

US EPA (United States Environmental Protection Agency). (2009). Series 890 – Endocrine Disruptor Screening Program Test Guidelines. http://www.epa.gov/test-guidelines-pesticides-and-toxic-substances/series-890-endocrine-disruptor-screening-program. Accessed December 20, 2015.

US EPA (United States Environmental Protection Agency). (2011). Evaluation Guidelines for Ecological Toxicity Data in the Open Literature. https://www.epa.gov/pesticide-science-and-assessing-pesticide-risks/evaluation-guidelines-ecological-toxicity-data-open#guidance Accessed March 30, 2016.

US EPA (United States Environmental Protection Agency). (2012). Guidance for Considering and Using Open Literature Toxicity Studies to Support Human Health Risk Assessment. Office of Pesticide Programs. https://www.epa.gov/pesticide-science-and-assessing-pesticide-risks/guidance-considering-and-using-open-literature Accessed March 30, 2016.

Van Der Kraak, G. J., Hosmer, A. J., Hanson, M. L., Kloas, W., and Solomon, K. R. (2014). Effects of atrazine in fish, amphibians, and reptiles: An analysis based on quantitative weight of evidence. *Crit. Rev. Toxicol.* **44**, 1–66.

Zoeller, R. T., Bergman, Å., Becher, G., Bjerregaard, P., Bornman, R., Brandt, I., Iguchi, T., Jobling, S., Kidd, K. A., Kortenkamp, A., *et al.* (2015). A path forward in the debate over health impacts of endocrine disrupting chemicals. *Environ. Health* **14**, 118.

Zoeller, R. T., and Vandenberg, L. N. (2015). Assessing dose-response relationships for endocrine disrupting chemicals (edcs): a focus on non-monotonicity. *Environ. Health* **14**, 42.

TABLE 1. Comparison of GLP Requirements to Criteria Proposed by Zoeller and Vandenberg (2015)

Criteria of Zoeller and Vandenberg (2015)	
a.	the use of appropriate negative and positive controls
b.	the use of sensitive animal species and strains
c.	the use of appropriate endpoints
d.	evaluation by subject experts
GLP Requirements[a]	
Subpart B—Organization and Personnel	
a.	education and proper training of personnel
b.	sanitation and health precautions to avoid contamination of test, control, and substances
c.	designated study director
d.	quality assurance unit
e.	appropriate testing of all substances and mixtures
f.	documentation and corrective action taken for deviations from regulations
g.	approval of protocol and changes to protocol
h.	accurate recording of all experimental data
i.	documentation of unforeseen circumstances that may affect study
j.	archival of all raw data, documentation, protocols, specimens, and final reports
k.	inspection of study at intervals by quality assurance unit to ensure integrity of study and maintain proper documentation
l.	Review of final study report by quality assurance unit
Subpart C—Facilities	
a.	facilities of suitable size and construction
b.	sufficient number of animal rooms or test areas
c.	isolation of studies being done with substances known to be biohazardous
d.	include provisions to regulate environmental conditions
e.	have storage areas that protect against infestation or contamination of substances, feed, soil, etc.
f.	separate areas for receipt and storage of substances, mixing of substances, and storage of mixtures
Subpart D—Equipment	
a.	equipment of appropriate design and adequate capacity
b.	methods for the adequate inspection, cleaning, and maintenance of equipment
c.	methods for the adequate testing, calibration, and standardization of equipment
d.	written records of all inspection, maintenance, testing, calibrating, and/or standardization of equipment
Subpart E—Testing Facilities Operation	
a.	written standard operating procedures
b.	operating procedures must include:
i.	test system room preparation
ii.	test system care
iii.	receipt, identification, storage, handling, mixing, method of sampling of all substances
iv.	test system observations
v.	laboratory or other tests
vi.	handling of test systems found moribund or dead during study
vii.	necropsy of test systems of post—mortem examination of test systems
viii.	collection and identification of specimens
ix.	histopathology
x.	data handling, storage, and retrieval
xi.	maintenance and calibration of equipment
xii.	transfer, proper placement, and identification of test systems
c.	immediately available manuals and standard operating procedures in each laboratory
d.	historical file of all standard operating procedures
e.	standard operating procedures for the housing, feeding, and care of animals
f.	isolation and evaluation of all newly received test systems from outside sources
g.	all test systems should be free of any disease or condition at the initiation of study
h.	appropriate identification of warm blooded animals
i.	units clearly marked with identification information

(continued)

j. test systems of different species shall be housed in separate rooms
k. test systems of the same species used in different studies should be housed in separate rooms
l. periodic analysis of feed, soil, and water used for test systems and documentation of such analyses
m. documentation of use of any pest control materials. Cleaning and pest control materials that interfere with study shall not be used

Subpart F—Test, Control, and Reference Substances

a. characteristics which appropriately define any substance shall be determined for each batch and documented before its use
b. solubility of each substance determined when relevant
c. stability of any substance determined before experimental start date, or according to standard operating procedure
d. adequate labeling of each storage container for a substance
e. reserve samples for each batch of any substance shall be retained for studies of more than 4 weeks
f. stability of all substances under storage conditions at test site shall be determined
g. procedures established for handling of all substances
h. distribution of a substance designed to preclude contamination, deterioration, or damage
i. proper identification maintained throughout distribution process
j. documentation of receipt and distribution of each batch
k. For each substance that is mixed with a carrier, tests by appropriate analytical methods shall be conducted:
    i. to determine uniformity, concentration of substance in mixture
    ii. to determine solubility of each substance in the mixture, when relevant to study
    iii. to determine the stability of the substance in the mixture
    iv. expiration date of a mixture clearly shown on the container

Subpart G—Protocol Requirements[b]

a. specification of age, weight, and acclimatization of animals
b. specification of weight variations of animals relative to mean weight for each sex
c. specification of numbers by gender of test animals used at each dose
d. specification of adequate randomization procedures in allocation of animals to test and control groups
e. animal identification by unique number; all animal parts labeled with corresponding ID number
f. specification of temperature of the experimental rooms
g. specification of relative humidity of the experimental rooms
h. specification of use of artificial lighting
i. specification of feed and nutrition for control and test groups; feed analysis for possible influential impurities
j. specification of period of acclimatization/quarantine of animals prior to initiation of study
k. specifications for analyzing the test, reference, and control substances for identity, purity, stability in dosing solutions, feed or media
l. if applicable, specification of incorporation of test substance into feed or another vehicle, methods for analyzing mixture, and achieved dose as a time—weighted average
m. specification of control groups, and the use of both vehicle and untreated control groups if test substance is administered through a vehicle of unknown toxic properties
n. specification of dosage and range of toxic effects; data must produce a dose-response curve
o. specification of intermediate dose levels
p. specification of toxicity of lowest dose level
q. specification of identical dosing method for all animals during entire experimental period
r. specifications for administration procedures for doses administered by gavage
s. specifications for clinical observations of animals and clinical conditions, including time and period of effect after dosing
t. specifications for conducting analyses and collecting data (including procedures to limit variation in observations and specifications of record keeping)
u. specification and measurements of food and water consumptions
v. specification for tracking weight of test animals
w. specification of accepted statistical methods used and significance criteria
x. specification of test reporting criteria in addition to requirements under EPA Good Laboratory Practice Standards
y. specification of systems developed and maintained to assure and document performance of laboratory staff and equipment

Subparts H–I [Reserved]

Subpart J—Records and Reports

a. specification of content of final report (including certification by study director and independent QA officer)
b. specification of procedures for storage and retrieval of study data and records
c. (c) requirements for retention of raw data, study records, samples of test substances, and specimens

[a] http://www.gpo.gov/fdsys/pkg/CFR-2011-title40-Vol15102/pdf/CFR-2011-title40-Vol15102-part792.pdf.

[b] Example using specifications from EPA 870.3100 90-day oral rodent subchronic study test guideline. http://www.epa.gov/ocspp/pubs/frs/publications/Test_Guidelines/series870.htm.