Occupational Safety and Health Administration

Guidance on Data Evaluation for Weight of Evidence Determination:

Application to the 2012 Hazard Communication Standard

1.0 INTRODUCTION

The purpose of OSHA's Hazard Communication Standard (HCS; 29 CFR 1910, 1915, and 1926) is to ensure that the hazards of all chemicals produced or imported are evaluated and that information concerning their potential hazards is transmitted to employers and workers. The HCS has three main components to ensure that workers obtain the relevant information and understand the hazards that they are exposed to at their workplace.

First, chemical manufacturers and importers must evaluate the hazards of the chemicals they produce or import (Paragraph (d) of §1910.1200).

Second, for every chemical found to be hazardous, the chemical manufacturer or importer must develop safety data sheets (SDS) and container labels to be transmitted to downstream users of the chemicals. Employers are required to maintain an SDS in the workplace for each hazardous chemical that they use (Paragraphs (f), (g) of §1910.1200).

Third, all employers must develop a written hazard communication program and provide information and training to workers about the hazardous chemicals in their workplace (Paragraphs (e), (h) of §1910.1200).

In 2012, OSHA modified its HCS to conform to the United Nations' Globally Harmonized System of Classification and Labelling of Chemicals (GHS). The GHS provides a set of criteria which includes the use of weight of evidence (WoE) when evaluating the health hazards of a chemical. OSHA has incorporated the GHS criteria in the HCS. Therefore, the WoE evaluation is part of the process for determining the potential health hazards of a chemical and what information must be disclosed on the label and SDS under OSHA's revised HCS.

This document provides guidance on a systematic application of the WoE approach consistent with the GHS criteria and explains the various types of information that need to be considered in order to establish classifications under the HCS. This guidance is a general statement of OSHA's approach to WoE under the HCS and provides general examples on how to apply GHS criteria to classify chemical hazards.

Note that the HCS is designed specifically for hazard communication, and is aligned with, and derives wording from, the GHS. The evaluation of hazards under the HCS is not designed to determine specific exposure limits, or require exposure controls, for the chemical under consideration. This would entail additional analyses for risk assessment. Hazard communication is intended to provide a full range of hazard information for all chemicals in order to inform workers of the hazards in their workplaces, and to enable employers using the chemical to implement appropriate controls, and prevent the occurrence of adverse effects. Therefore, the reader may notice differences between the discussion of WoE in this guidance document and similar guidance, such as systematic reviews from other authoritative bodies that address a single chemical or group of related chemicals, identify risks, and provide exposure limits and required control measures. Under the HCS, OSHA has established a lower threshold for dissemination of hazard information than would be used for a rule that implements specific, risk-based controls for an individual substance regulated by the Agency.

The HCS is intended to be conservative in nature to ensure that employers are informed about the potential hazards of the products they use and that workers are alerted to and protected against these potential hazards. In other words, where there is uncertainty, OSHA expects that documents produced to meet the requirements of the HCS will err on the side of providing warnings and categorization into more hazardous categories.

1.1 Background information: Hazard identification under the original Hazard Communication Standard

The original HCS provided a threshold for chemicals to be considered hazardous based on a single study, provided that this study was conducted according to established scientific principles and produced a statistically significant result consistent with the definitions of hazard in the standard. OSHA took this "one-

study" approach to reduce the number of situations where conflicting determinations might be made for the same chemical by different suppliers. OSHA provided some guidance on hazard determination in Appendix A and B of the original rule.

Most of the hazard definitions under the original rule simply led to a conclusion that the chemical involved presents that hazard or it does not. For example, for a chemical to be covered under the original HCS as a carcinogen, this determination could be done by relying on classifications produced by an authoritative scientific body such as the International Agency for Research on Cancer (IARC) or the US Department of Health and Human Services National Toxicology Program (NTP) or under the rule based on one study indicating that there is evidence of a potential carcinogenic effect. The original standard did not generally address the degree to which evidence supports carcinogenicity. However, while a one-study determination led to providing information about that hazardous effect on an MSDS, it may not have led to a hazard warning on a label. Previously the HCS only required that such warnings to be "appropriate", and there were situations in which the manufacturer determined that the data did not support warning about the hazard on the label because of other negative studies or information. Thus, even though positive results from one study may have been used to determine initially whether a chemical was to be considered a hazard under HCS, there was a consideration of the weight of evidence when deciding what to include on a label.

OSHA expected the hazard evaluation process to go beyond simply identifying one positive study and to include a complete evaluation of all of the information available when determining what would be appropriate to include on the label and the MSDS. Chemical manufacturers and importers were also allowed to review the weight of evidence in preparing MSDSs, and were permitted to discuss negative evidence and other constraints when reporting the information on the MSDS.

1.2 Background information: Alignment of the Hazard Communication Standard and GHS

In March 2012, OSHA aligned the HCS with the GHS to improve the quality and consistency of the information provided to employers and workers through labels, safety data sheets and training. The GHS achieves this in two ways.

First, the GHS provides a framework and set of guiding scientific criteria to direct the hazard classification process. These criteria help ensure that different evaluators evaluate the same data for the same chemical in a consistent manner. By aligning with the GHS, the HCS now provides specific instructions under Appendix A for the type of evidence used to determine the health endpoints, known as hazard classes. Second, the HCS refines the classification process by establishing categories of hazard within most hazard classes. These categories,

depending on the health hazard class, can either indicate the relative degree of hazard or the strength of (or level of confidence in) scientific evidence supporting the conclusion of the health endpoint. The HCS now allocates the hazard information to be provided for each hazard category, which is tailored to the level of hazard. The classification criteria established in the HCS thus provide the basis for development of the specific hazard information that enhances the protection of workers.

The method to determine a chemical's health hazards varies depending on the health endpoint. Classification for some endpoints can result directly from quantitative test data derived from validated test methods, such as acute toxicity based on values for LD50. When test data are not available or there are conflicting results, some endpoints use a tiered approach for hazard classification, as with skin corrosion/irritation and serious eye damage/eye irritation. Finally, some endpoints, due to the complexity of the mechanism of the endpoint, rely on the WoE approach and thus the hazard categories reflect the degree to which there is evidence that the chemical presents the hazard.

Although OSHA does not expect a major shift in the determination of health effects of chemicals under the new GHS-aligned HCS, it anticipates that hazard classification will be more structured, organized and transparent.

This guidance has been developed to assist manufacturers, importers and employers on how to evaluate scientific studies under a WoE approach for hazard communication purposes. Specifically, the present guidance will outline the types of information to consider in the process of classification of a substance for health hazards, how to evaluate the strength of evidence in classification, the scope and use of WoE, and detailed considerations in the use of WoE. This guidance focuses on the more complex endpoints, which generally need a higher degree of expert judgment to interpret the studies, and includes examples of classification of different substances for each of these endpoints (carcinogens, germ cell mutagens, and reproductive hazards).

2.0 General considerations in the determination of classification under HCS

2.1 The reasons for using weight of evidence

As discussed above, for the more complex hazard endpoints OSHA's HCS classification process not only determines if the chemical is hazardous or not, it also indicates the confidence with which exposure to a given substance or mixture can be causally linked to a resulting health effect. For example, classification of a substance as a Category 1 carcinogen indicates greater confidence in the evidence for carcinogenicity in humans compared to classification as a Category 2

carcinogen. Therefore, OSHA uses WoE as a tool to make judgments on classification in cases in which multiple lines of evidence can be combined to facilitate an informed decision on classification.

For some endpoints, the HCS requires that classification is to be "determined on the basis of the total weight of evidence using expert judgment. This means that all available information bearing on the classification of hazard shall be considered together, including the results of valid *in vitro* tests, relevant animal data, and human experience such as epidemiological and clinical studies and well-documented case reports and observations" (HCS, A.0.3.1).

This can include the integration of a diverse body of epidemiologic studies, experimental animal results, *in vitro* data, and computational (or *in silico*) data in determining the appropriate hazard category. The scientific information is rarely conclusive and information may be incomplete. For example, only a few strains/species are tested, epidemiologic studies are too small to provide a conclusive result, or endpoints are discordant among studies (such as genotoxicity tests that include positive and negative outcomes).

As a result, conclusions about causality from chemical exposure may vary from very strong to less convincing depending on the particular chemical or substance. In addition, there are qualitative questions (the plausibility of a chemical eliciting a particular effect in exposed humans, i.e., species specificity) and quantitative questions (the accuracy and precision of the projected dose-dependent magnitude of that effect or probability of that magnitude of effect) that must be considered when determining the specific hazard classification category.

A WoE approach provides a systematic way to evaluate a group of health effects studies that vary in quality and provide conflicting information. There are several reasons why the use of WoE improves hazard classification. First, WoE makes use of all available information; this aspect is important especially when there is conflicting information between studies. Second, it may be possible to pool the results of several less conclusive studies to assist in reaching a conclusion on the relevant endpoint. Finally, WoE allows classifiers to combine information from different types of studies (e.g., other species, routes of exposure tested).

2.2 Preparing an Evaluation

For some classes of chemical hazard, classification results directly from data that satisfy the criteria (e.g., acute toxicity). For many other classes of chemical hazard such as carcinogenicity, reproductive toxicity and germ cell mutagenicity, classification is based on the total WoE using expert judgment.

Weight of evidence, as used in the HCS, means that "all available information bearing on the determination of toxicity is considered together," including *in vitro* tests, relevant data from experimental animals, and data from humans. The

identification of study evidence should employ standardized search strategies and reporting formats with clearly stated inclusion/exclusion criteria. Both positive and negative results are assembled together and the quality and consistency of the data are evaluated in order to make a judgment on classification. However, as discussed in Section 2.4, a single positive test that is performed according to good scientific principles, and with statistically and biologically significant positive results, may justify classification. (See also A.0.3.5 of the HCS.)

The first step in an evaluation is to obtain all available information that might pertain to the material and the possible classification in question. In some cases, appropriate reviews have been done previously by independent or government authoritative bodies, and the conclusions of such reviews can be used for purposes of hazard communication under the HCS, as discussed in Section 2.4. When reviewing the conclusions made by these authoritative bodies the classifier should ensure that the conclusions specifically address hazards (or exposures) which are relevant to the workplace.

If such reviews are not available, the relevance of each "packet of information" to decisions on classification is determined on a scale from highly relevant to marginally relevant. A "packet of information" could be an epidemiology study, a toxicity study in lab animals, information on physical/chemical properties, data on mode of action of the chemical, data on ADME (absorption, distribution, metabolism, and excretion), data on structurally similar substances, or other information. The search strategy and protocol should be documented. This includes databases covered,, search terms, range of dates searched, and inclusion/exclusion criteria.

With each packet of information, the WoE approach has two phases that can occur simultaneously. Using experimental studies as an example, the first phase involves a review of available experimental studies to determine how much weight might be given to any particular study based on its design, conduct, interpretation, and relevance. The goal is to establish which studies are the most credible and which, if any, should be included in the evaluation even though there might be weaknesses in some facet(s) of the study, such as the design or conduct of the study. With both relevance and study quality in mind, a weight to each packet of information is assigned in terms of how strongly each supports classification or supports no classification.

The conclusions of this review can then be used in the second phase of WoE, in which the findings of these studies and other relevant packets of information are evaluated collectively to determine the degree of confidence that a given toxicity outcome is causally associated with the substance in question. Attention is given to whether the assembled information appears to present a unified picture of the effects of the material in the body or whether there are discrepancies within the collected assembly of data. The question then is whether the WoE from this collected body of information is sufficient to support classification of the material.

Krauth et al., (*Environ. Hlth. Perspec.* 121:985-992, 2013) presents systematic assessment criteria for bias, methodology, and reporting of studies in laboratory animals. Another tool that some evaluators have used in the evaluation of a particular set of experimental data involves criteria developed by Klimisch et al. (*Regul. Toxicol. Pharmacol*. 25:1-5; 1997). Scores for reliability of the data can be assigned using these criteria and then used to help judge the relative weight to give to various studies on the same substance or its analogs in the overall weight of evidence. However, this method must be used judiciously. The assessment is subjective and only oriented to an evaluation of the quality of the study—it does not assess bias or validity of the study. In addition, its criteria have an inherent preference for standardized studies, which is not employed under the HCS WoE approach. Other methods for judging the acceptability of studies may be used together with or separate from those in these publications.

The common steps in making classification decisions based on WoE for the various endpoints are:

(1) preparation of an outline on the procedure for the review, and criteria for making decisions on classification;

(2) identification of relevant information (epidemiological data, experimental data [*in vivo* and *in vitro*], information on structural analogs, and any other related information);

(3) review of each set of information to determine its quality, potential for bias, validity and usefulness for classification;

(4) decision on whether any existing review by an authoritative body is sufficient to use as a basis for a decision on classification under the HCS;

(5) barring reliance on an existing authoritative review, evaluation and integration of scientific evidence by a systematic, transparent approach using the criteria described in the present guidance;

(6) preparation of short descriptions of available sets of information with stated rationale for weight assigned to each set;

(7) selection at an appropriate classification based on an understanding and acknowledgment of the relative merits of and appropriate weight given to each set of information;

(8) recognition of any existing gaps in information that might be needed to finalize a decision on classification; and

(9) development of a plan to address those gaps. Note that WoE often is a qualitative assessment; quantitative assessments are not advocated here unless they are appropriate to the data.

OSHA recommends that all data and other evidence relevant to understanding the classification decision be maintained for future reference and use, and that these be made available to those who want to better understand the scientific basis for the classification. The HCS requires that the WoE decisions be based firmly in the scientific evidence. Where the classifier has documented the basis for its classification, it will be in the best position to defend it should questions arise in the future. During a compliance inspection, the agency may request that a manufacturer or importer provide the evidence and rationale supporting a particular classification in deciding whether the chemical has been properly classified. The agency may request the classifier to provide this evidence in the event of a compliance inspection if the classifier's conclusion regarding hazard classification is questioned or challenged by the Agency.

In some instances, WoE evaluations might have already been performed by others and, if complete and consistent with the criteria under the HCS, could be used in a hazard evaluation under HCS. In particular, evaluations for carcinogenicity performed by the IARC or the National Toxicology Program (NTP) are considered sufficient for the HCS. Given their expertise and the robust processes used by these bodies, OSHA will generally defer to conclusions on carcinogenicity from IARC or NTP unless clear reasons are presented to do otherwise. See Section 3.2.3.3 for more discussion of this issue.

The required experience and expertise of the individuals who perform the hazard assessments may vary with the situation. For example, a decision on carcinogenicity for a chemical that has been reviewed by IARC could be performed by an individual who is well versed in the HCS but does not have extensive knowledge in toxicology. However, decisions that require reviews of primary epidemiological data might require input from a qualified epidemiologist, just as a qualified toxicologist might need to be involved in reviews of experimental data. In short, the abilities of the reviewer must be adequate to meet the challenges and complexities of a particular review; reviewers who do not have adequate expertise must seek help from others who are sufficiently qualified.

2.3 Approaches related to weight of evidence used by other authoritative bodies

In addition to OSHA, several other authoritative bodies use WoE in evaluations of potential health hazards of substances and mixtures while others use similar approaches. Authoritative bodies that use WoE include the following:

(1) United Nations, Globally Harmonized System of Classification and Labelling of Chemicals (GHS), upon which the HCS is based

(2) US EPA, Guidelines for Carcinogen Risk Assessment (EPA/630/P-03/001F March 2005)

(3) US FDA, Guidance for Industry and Review Staff - Recommended Approaches to Integration of Genetic Toxicology Study Results

(4) US FDA, Guidance for Industry - S1B Testing for Carcinogenicity of Pharmaceuticals (1997)

(5) US EPA, Endocrine Disrupter Screening Program

(6) Agency for Toxic Substances and Disease Registry (ATSDR)

(7) European Chemicals Agency (ECHA)

In addition, approaches similar to WoE are used by other bodies. Some of these approaches were recently discussed in the National Research Council's Review of EPA's IRIS process (NRC, 2014). For example, NTP's Office of Health Assessment and Translation (OHAT) uses systematic review for literature-based health assessments, a more structured process. "This process involves a 7-step framework for systematic review and evidence integration for reaching hazard identification conclusions: problem formulation and protocol development, search for and select studies for inclusion, extract data from studies, assess the quality or risk of bias of individual studies, rate the confidence in the body of evidence, translate the confidence ratings into levels of evidence, and integrate the information from different evidence streams (human, animal, and 'other relevant data' including mechanistic or *in vitro* studies) to develop hazard identification conclusions" (Rooney et al. 2014. *Environ Health Perspect*; DOI:10.1289/ehp.1307972). The OHAT Approach for Systematic Review and Evidence Integration is compatible with the GHS guiding scientific principles to hazard evaluation and can serve as a model for the WoE process discussed in the present guidance.

IARC uses a systematic approach that considers the same factors used in an evaluation of the WoE of a potential carcinogen, but it does not use the term "weight of evidence." Instead, IARC uses "strength of evidence" and "degree of evidence", and IARC's "strength of evidence" has a more general meaning than the "strength of evidence" as defined under the procedures used by GHS and OSHA for evaluation of carcinogenic hazard. Nonetheless, the procedure used by IARC is similar to the WoE of other bodies and can also serve as a model for the WoE process discussed in the present guidance.

If a classifier reaches a final WoE conclusion that differs from that of the NTP or IARC, OSHA would look, in the event of a compliance inspection, for a clear justification for the different classification. (See Section 3.2.3.3.) If OSHA disagrees with the classifier's classification after evaluating the classifier's justification, OSHA may issue a citation.

2.4 Classification Based on a Single Positive Study

The words "positive" and "negative" are often applied to toxicological and epidemiological studies, but care should be taken in the assignment of these words. In a positive study, exposure to a chemical would have a demonstrable significant effect in specific endpoints in exposed individuals. The nature of the effect could vary from an increase in redness of the skin in a dermal irritation study to a statistically significant increase in specific types of tumors in a carcinogenicity study, and many criteria for positive studies are presented in Appendix A of the HCS. No demonstrable significant effects are seen in a negative study. In some studies the designation of results as "positive" or "negative" is not as clear-cut as the words imply and expert judgment is needed (such as with studies that might be false negatives).

The HCS states that, in some cases, a single positive study, performed according to established scientific principles and with statistically or biologically significant positive results may justify classification. However, the decision-making process must also take into account any other available data. Positive human evidence is not required for a positive hazard determination; positive animal studies should not be discounted without a strong scientific rationale for doing so. (See A.0.3.3.) As discussed below, a classifier generally may not rely on a single negative study to justify non-classification for a specific hazard class where there are positive studies for that endpoint.

In addition, if one positive study is being considered as a basis for classification, the design and conduct of that study must first be evaluated carefully. Studies conducted in accordance with internationally accepted test guidelines, such as those from the OECD and US EPA, are generally acceptable, provided that there are no significant deviations from those guidelines. The study must meet the criteria of being considered scientifically robust, conducted according to internationally recognized scientific principles, scientifically sound, and validated according to international procedures. Validation, as described in section 1.3.2.4.2 of the GHS, is "the process by which the reliability and the relevance of a procedure are established for a particular purpose." This includes considerations such as appropriate control groups, minimizing bias, control of confounding variables, and reliability of exposure characterization. Stated another way, the test methods must be standardized so that "the results are reproducible with a given substance, and the standardized test yields 'valid' data for defining the hazard class of concern." (HCS A.0.2.3) The overriding intent of these considerations is to ensure that data derived from the test method are credible, reproducible, and relevant to the determination of classification. In other words, classification can be made based on information from one study, but that approach must be performed while taking into account the validity of that study, the scientific strength of results from the study, and all other available data on the chemical.

Using carcinogenicity as an example, among the information reviewed during a WoE evaluation there might be one study that meets the criteria for classification by itself. A well-performed epidemiology study with conclusive positive results and that appropriately addresses confounding factors can justify classification. Alternately, an experimental study in laboratory animals that resulted in a statistically significant, conclusive, and mechanistically consistent increased incidence in tumors can also justify classification. However, a decision to base classification on one experimental study must also consider any limitations of the study as discussed in the HCS (e.g., Appendix F to § 1910.1200) and in sections 3.2.3.1 and 3.2.3.2 of this guidance. A decision based on either one positive epidemiological study or one positive laboratory study must also address any information that supports or conflicts with the decision on classification.

In some cases, a single positive study might be chosen as the basis for classification, but other studies do not demonstrate a similar effect. In such instances, the discrepancy must be reasonably resolved before the decision on classification is finalized. Stated differently, both positive and negative results shall be considered together in a single WoE determination (HCS, A.0.3.2), however, "positive effects which are consistent with the criteria for classification, whether seen in humans or animals, shall normally justify classification" (HCS, A.0.3.3). General guidance for classification is provided in the HCS non-mandatory Appendix F and are similar to the guiding principles used for IARC cancer classification.

While a positive study normally justifies classification, negative or marginal studies may help in determining the hazard category or the degree of evidence. In cases where one or more negative studies with high quality and validity providing a strong degree of evidence conflict with the positive study, the classifier would need to reconcile the conflicting evidence. The positive study would take precedence in classification unless there were compelling reasons not to do so, such as fundamental flaws in the positive study or data that clearly established that positive results are not relevant to humans. When there is no clear reason to reconcile the difference, OSHA would expect the classifier to rely primarily on the positive study for classification.

The purpose of the HCS is to ensure that all employers receive the information they need to design and implement worker protection programs and properly train their workers on the hazards of chemicals that are used in the workplace. Therefore, it is important to avoid a false negative or under classification of a chemical where an employee may believe that a chemical is safe when it is not. In such cases, the worker might have a false sense of safety.

3.0 Hazard classification based on WoE for different hazards

In the classification criteria for carcinogens, the HCS uses the terms "strength of evidence" and "weight of evidence." "Strength of evidence" describes the enumeration of tumors in human and animal studies and the determination of their level of statistical significance. "Weight of evidence" involves the consideration of other factors, beyond strength of evidence, that influence the likelihood that a chemical may pose a carcinogenic hazard, such as tumor type and background incidence, multisite responses, mode of action, and the comparison of absorption, distribution, metabolism and excretion between test animals and humans.

In this context, the consideration of both strength of evidence and WoE in evaluating carcinogens for HCS is similar to the concepts used by other authoritative bodies such as IARC and NTP (although the NTP and NAS IRIS are moving towards the phase "evidence integration" to describe this concept (Rooney et al. 2014. *Environ Health Perspect*; DOI:10.1289/ehp.1307972), which is consistent with the approach suggested by NAS (http://www.epa.gov/iris/iris-nrc.htm)). The HCS uses a similar approach for other hazards, even though it does not use the words "strength of evidence" for those hazards. The HCS requires the same general approach to evaluate studies in determining hazard classification; this will be discussed in general terms under the detailed discussion of WoE in this section.

For germ cell mutagenicity, classification is expected to be based on WoE. It is likely that more than one study will ordinarily be performed for potential germ cell mutagens in order to obtain sufficient data to support a decision on classification. In some cases classification may be warranted based solely on one study where there is a particularly definitive epidemiology study. This approach is discussed further in subsequent sections of this guidance.

As with germ cell mutagenicity, classification for reproductive toxicity is expected to be based on WoE. The HCS provides considerable guidance on factors to consider, including maternal toxicity and the possibility that observed effects might be nonspecific secondary results that were indirectly caused by effects of the substance on other organs in the body. In some cases a single positive study "performed according to good scientific principles and with statistically or biologically significant positive results may justify classification" (See Section 2.4.) Professional evaluation and judgment are central to reach decisions on possible reproductive toxicants, particularly for instances in which data from available studies might be considered less than optimal or sufficient. These judgments should be clearly outlined in the analysis.

Substances shall be placed in Category 2 for reproductive toxicity when (see HCS Figure A.7.1(a).) the following conditions are present:

(1) There is some evidence from humans or experimental animals, possibly supplemented with other information, of an adverse effect on sexual function and fertility or on development;

(2) This evidence occurs (a) in the absence of other toxic effects or (b) together with other toxic effects and the adverse effect on reproduction is considered not to be a secondary, non-specific consequence of the other toxic effects; and

(3) The evidence of reproductive toxicity is not sufficiently convincing to place the substance in Category 1.

For other health hazard classes the WoE plays less of a central role in the classification process. For example, the criteria are relatively straightforward for acute toxicity (oral, dermal, and inhalation) and are based on the numerical values of LD50, LC50, or acute toxicity estimates (ATE). For skin corrosion/irritation, a decision on classification for skin irritation can be based on numerical scores for erythema/eschar or edema with caveats to adjust for variability among animals, delayed reactions, and the duration of lesions. Dermal corrosion is also based on the incidence of corrosion among multiple animals. Alternately, a tiered approach can be also used to evaluate available information, including data from both humans and laboratory animals, to make a decision based on WoE. A similar process applies for eye damage/irritation.

Classification of substances for respiratory or dermal sensitization uses (1) a WoE approach involving criteria presented in the HCS, (2) reliable and good quality evidence from human cases or epidemiological studies, and/or (3) appropriate studies in laboratory animals. Guidance is given in the HCS on specific criteria for classification based on studies in laboratory animals. Using these criteria, determination of classification could be performed using results of one definitive study. However, a careful review of all data is appropriate. Although the criteria are different from those used for sensitization, the general approach in the HCS to classification of substances of specific target organ toxicity (STOT) with both (1) single or (2) repeated or prolonged exposures is similar and includes the use of human and/or animal data.

The present guidance focuses mainly on carcinogens, germ cell mutagens, and reproductive toxicants since these are more complex endpoints and generally need a higher degree of expert judgment to interpret the studies. Therefore, discussions of these three particular categories receive a greater level of detail here.

Classification of a substance or mixture as a carcinogen is based on the inherent hazard properties of the substance or mixture. (i.e., hazard-based scheme). It does not take into account the amount of substance needed to trigger the adverse outcome or the level of risk to exposed workers. Therefore, classification can be based on studies in laboratory animals at doses greater than those that people would normally encounter. The potential exposures or level of risk to people during the use of the substance or mixture in a given workplace are not considered

for purposes of determining hazard classes. However, the standard allows risk to be addressed in workplace-specific training, where employers can provide information about exposures in that workplace and the control measures implemented to address those exposures.

3.1 Hierarchy of weight given to different data

A general hierarchy exists for the weight that might be ascribed to a particular set of data during a WoE evaluation. The underlying principle is that greater weight is given to those studies that are well designed, are well performed, and can be most readily extrapolated to conditions in people. Typically, data obtained from studies in humans would receive more weight than those in laboratory animals. However, chemically induced toxicity observed in experimental animals is assumed to also occur in humans unless there are clear and convincing data to establish otherwise. Therefore, classification of chemicals is often based on data from experimental animals absent data on people.

Among lab animals, studies in mammals receive more weight than those in non-mammalian animals. Likewise, *in vitro* studies performed in human cells and tissues receive more weight than those in bacteria and other prokaryotes (species without a cell nucleus).

The conditions of the study also influence the weight given to it. *In vivo* (in life) studies performed in a living organism have greater weight than *in vitro* ("in glass") studies that are performed in an artificial environment and not in a living organism. For example, mutagenicity studies in mice have more weight than mutagenicity studies in bacteria which are cultured in the laboratory, incubated with the test material plus metabolizing enzymes, and then tested for their ability to reproduce in specific growth media. (In other words, weight for mammals > non-mammals, weight for eukaryotes > prokaryotes, and weight for *in vivo*>*in vitro.*)

Greater weight is given to those studies that have the statistical power to demonstrate a statistically significant difference in quantitative data between treated animals and the respective control animals. Statistical significance provides a quantitative indication of the likelihood that the difference is not due to random variation. However, in some instances there might be a difference that is not statistically significant but contributes to the WoE due to information that strongly supports the likelihood of such an effect. In other words, even though the observed effect is not statistically significant, it might be biologically significant (e.g., occurrence of a rare neoplasm in animal studies that is highly relevant to humans). In such cases, the evaluation of the WoE can include consideration of this observed effect. Greater weight is also given to study evidence that demonstrates a large (statistically significant) magnitude of effect, significant dose-response relationship, is consistent with findings from other evidence (e.g.,

across animal species or dissimilar populations) and conforms with a well-accepted mode of action (i.e., biological plausibility).

Decisions made while using this typical hierarchy of toxicity studies can be affected by other factors, including those discussed in later sections of this guidance. If, for example, there is a conflict between findings in humans and animals, the quality and reliability of the evidence from both sources must be evaluated in order to resolve the question of classification. Reliable, good-quality human data shall generally have precedence over other data. However, even well-designed and well-conducted epidemiological studies may lack a sufficient number of subjects to detect relatively rare but still significant effects, or to assess potentially confounding factors. Therefore, positive results from well-conducted animal studies are not negated by a lack of similar findings in people unless the epidemiologic study is a particularly well-conducted one, with adequate power to detect small differences in risk. Instead, an assessment of the robustness, quality, and statistical power is needed for both the human and animal data.

3.2 Weight of evidence for carcinogens

Classification of substances as carcinogens is to be based on strength of evidence as well as additional considerations under WoE. That is, both are to be used in combination with each other to determine the categorization of the carcinogen. This basic approach and the criteria to consider are similar to those used by IARC and NTP.

3.2.1 Strength of Evidence

The "strength of evidence" generally refers to the degree in which human and/or animal tumor incidence across the available studies support a causal association with the carcinogenic substance under consideration. An evaluation of the strength of evidence includes factors such as study design, study quality, consistency of effect, risk of bias, strength of association, statistical significance, and dose-response relationships. "Sufficient human evidence demonstrates causality between human exposure and the development of cancer, whereas sufficient evidence in animals shows a causal relationship between the agent and an increased incidence of tumors" [1] (A.6.2.4 of the HCS). Limited evidence occurs when data exists to suggest a positive association or carcinogenic effect but the evidence is considered less than sufficient to demonstrate causality. IARC explains that categorization of a substance based solely on the strength of evidence indicates that the substance is "carcinogenic and not the extent of its carcinogenic activity (potency)" (IARC, 2006). Beyond the determination of the

[1]Where the weight of evidence for the carcinogenicity of a substance does not meet the criteria for classification as a carcinogen, any positive study conducted in accordance with established scientific principles, and which reports statistically significant findings regarding the carcinogenic potential of the substance, must be noted on the safety data sheet (footnote to Category 2 in Table A.6.1)

strength of evidence for carcinogenicity, a number of other factors influence the overall likelihood that an agent may pose a carcinogenic hazard in humans (A.6.2.5). These additional considerations include: route of exposure, mode of action, relevance to humans, and structural similarity to other known carcinogens. Further discussion of the WoE is in the following sections.

3.2.2 Weight of evidence for carcinogens: Epidemiological studies

Data on carcinogenicity in humans typically are given greater weight than data obtained using laboratory animals during a WoE evaluation, but results from a human epidemiology study cannot necessarily be assumed to be valid until the methods, results, and interpretation of the study have been carefully reviewed and evaluated. This caution results from the fact that a number of factors can potentially affect the accuracy of an epidemiology study. Some of the general factors that can influence these studies are discussed briefly here.

Some epidemiologic studies that help establish causal relationships are called case-control. In these studies, a group of subjects that have the disease (or other adverse effect) being studied is identified, and that group's exposure rate to a substance of interest is compared to the exposure rate of group of subjects without the disease. Details of the levels of previous exposures and also exposures to other agents are generally obtained retrospectively. Information can be incomplete or inaccurate due to bias in the recall of exposures by individuals.

In contrast, during cohort studies individuals are selected based on their exposure to one or more particular substances, or to an environment, and they are then followed prospectively over time or in the case of a retrospective study the investigator collects data from past records. The development of disease can be followed, yielding an idea of the relative risk related to exposure to the particular agents or environment.

Both types of epidemiology studies may be used in meta-analyses, which involve methods that combine the results from multiple studies in an effort to increase the size of population studies and provide more data to examine whether an effect exists and, if so, the size of that effect. In addition to these methods, case reports can also provide useful information in developing warnings about the potential effects of exposure to particular substances.

Regardless of the study design, an assessment of a particular study entails consideration of several factors, including those listed below:

(1) Were there any other factors, such as concurrent disease, that might have affected the outcome of the study?

(2) Was the size of the exposed population sufficient to detect a carcinogenic effect relative to the control population?

(3) Did the study consider other factors (like smoking) that may be associated with tumor incidence?

(4) Was the information on exposure qualitative or quantitative? If qualitative, is the source of the information likely to be accurate and unbiased? If quantitative, were the methods and record-keeping sufficiently valid to provide reliable data on the actual doses or exposures levels to the substance? Such exposures could occur in the workplace, home, or elsewhere depending on the circumstances.

(5) Were individuals exposed either concurrently or at other times to another agent that might have had similar toxicity?

(6) Was information on effects obtained from self-reporting by the individual subjects (potentially affected by bias in recall) or by objective observation by qualified personnel?

(7) Were the methods and criteria for statistical evaluation adequate and appropriate? Was an association demonstrated between exposure to the agent in question and tumors in the exposed population?

One aid that may be useful in examining the causal relationship between exposure and disease in humans are the criteria developed in 1965 by Austin Bradford Hill (*Proceedings of the Royal Society of Medicine*, http://www.edwardtufte.com/tufte/hill). The criteria that he recommended considering were strength of the association, consistency of findings, specificity of the association, temporality (effect occurring after the cause), biological gradient (exposure-response), plausibility, coherence between epidemiological and laboratory findings, experimental evidence, and analogy to effects of similar factors.

Although the endpoints and other specifics might differ between epidemiological studies on cancer and other adverse effects, many of the factors to consider during the review of the studies remain the same or are similar. A body of evidence that satisfies most of the Hill criteria increases the confidence and validity of categorization as a Category 1A carcinogen based on the human evidence. However, the absence of one or more of the Hill criteria does not preclude the existence of a causal relationship between an exposure and an outcome. Furthermore, since the Hill criteria are primarily based on human evidence, absence of several of the criteria should not be used as a reason to categorize an exposure as non-carcinogenic; they can only be used as an aid in evaluating a substance's human carcinogenicity.

Many other sources of information on epidemiological studies are available and other study designs are discussed in them. These sources include *Cancer Epidemiology: Principles and Methods* (http://www.iarc.fr/en/publications/pdfs-online/epi/cancerepi) from the IARC and *Guidelines for Carcinogen Risk Assessment* from the US EPA (http://www2.epa.gov/risk/guidelines-carcinogen-risk-assessment).

One factor to consider in the evaluation of epidemiological studies (and studies in laboratory animals as well) is bias. The *Review of EPA's Integrated Information System (IRIS) Process* by the National Research Council states that the 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. An assessment of risk of bias is a key element in systematic-review standards; potential biases must be assessed to determine how confidently conclusions can be drawn from the data." The report describes approaches to the assessment of the risk of bias and therefore can serve as a useful reference on this topic during the review of data for classification of substances.

3.2.3 Weight of evidence for carcinogens: Experimental studies

There are two parts to WoE used in the determination of classification for carcinogenicity based on experimental studies. The first involves a review of the design, conduct, and interpretation of carcinogenicity studies. The second part of the WoE determination is broader and encompasses all other available information that is relevant to the determination of classification.

3.2.3.1 Weight of evidence for carcinogens: Evaluating experimental carcinogenicity studies

Many factors are involved in reviewing the WoE of a specific carcinogenicity study in laboratory animals. The following examples are not all-inclusive, but they do provide some suggestions of the types of issues to consider. Additional factors to consider can be found in Krauth et al., (*Environ. Hlth. Perspec.* 121:985-992; 2013) and Rooney et al. (*Environ. Hlth. Perspec.* http://dx.doi.org./10.1289/ehp.1307972; 2014).

(1) Were the studies conducted in accordance with established test guidelines, such as those from the OECD or the US EPA? Studies conducted using these guidelines are given significant weight. Other *in vivo* carcinogenicity studies need particularly critical review to determine if their design and conduct are sufficient for the determination of carcinogenicity. In either case, does the study report/publication contain sufficient information for judgment of the validity of the study?

(2) Was the number of animals in each control and treated group sufficient to provide adequate statistical power to evaluate tumor incidence at the end of the study? That is, what is the minimum incidence of tumors in treated animals that could be determined to be statistically different from the incidence of those tumors in the control group? Chronic studies in laboratory animals typically have 50 animals/sex/dose, based in part on these types of considerations.

(3) Was survival or other indicators of overt toxicity affected in the treated animals? If so, consideration should be given both to the statistical power in the analysis of tumor incidence with reduced numbers of animals and to potential effects that toxic effects on organs might have played in the development of tumors.

(4) Were the doses adequately high to test for carcinogenicity, but not so high as to cause toxicity that would result in tumors by a secondary mechanism? Exaggerated doses are necessary to test for carcinogenicity in the relatively small population of animals in each group. However, tumors only at high doses that are accompanied by severe toxicity would contribute less to the WoE.

(5) Was the duration of treatment sufficient for tumors to develop and was the total duration of the observation time before termination of the study sufficient for tumors to be seen? When possible, animals in carcinogenicity studies receive doses over much of their expected lifespan and are examined for tumors at the end of the study or in the case of unplanned death. Similarly, was the frequency of treatment adequate? Doses are often given 5 days/week.

(6) Were appropriate vehicle or treatment controls provided? Such controls are needed to reduce the likelihood of confounding factors that might reduce the weight, given to a study. Similarly, all groups of treated animals should receive similar treatment except for the dose of the tested substance that they are given.

(7) Information on the relevance of the experimental route of exposure in the animals to the expected route(s) of human exposure is important. For example, if ingestion in drinking water is the expected route in humans, then a study by that route would have greater weight than one by a different route, unless volatility of the compound makes drinking water concentrations uncertain or unpalatability of the compound significantly decreases water consumption. In some cases, a route-specific classification may be justified (HCS, paragraph A.6.2.1).

(8) A thorough review of the data on the incidence of tumors is important. Did tumors occur in one species or sex, or did they occur in multiple species and sexes? Greater weight can usually be given to tumors that occur among multiple species and sexes. Did tumors occur at more than one site? Did only preneoplastic lesions or benign tumors occur or were malignant tumors (cancers) identified? Did unusual tumors not normally observed in this species and sex occur, even if incidence is not statistically significant? If so, they may be given additional weight. Was there evidence that tumors occurred earlier (reduced latency) in the treated groups than in the controls or groups treated with lower doses?

(9) A dose-related increase in the incidence of tumors would generally be given more weight than an increase that is not dose-related.

(10) When evaluating the results of a study, was the incidence of tumors in a treated group within the range of incidences seen in historical control groups? If so, the incidence in the treated group could have less weight since it was within the range that could normally occur in untreated animals. Also, how extensive are data on historical controls for the particular species and strain under the conditions used in the study being evaluated? A larger number of studies with historical controls gives greater credibility to the cited range of control data.

(11) The level of statistical significance of difference in the incidence of specific tumors in test animals compared to control animals is a common way to judge the likelihood that observed tumors were not due to random variation. The probability that the observed tumors were treatment-related increases as the level of statistical significance increases.

(12) Details of the experimental method are important. For example, were the methods for processing tissues sufficient to locate and identify tumors? Were appropriate methods for statistical analysis used? Was the composition of the tested substance analyzed and was its stability ascertained?

3.2.3.2 Weight of evidence for carcinogens: Other factors

The second, broader evaluation of WoE involves sources of information other than long-term carcinogenicity studies. This information is related to the carcinogenicity of the substance and aids in a decision on classification of that substance. More specific examples include the following:

(1) Absorption, distribution, metabolism, and excretion (ADME): Information on the similarity between test animals and humans of absorption of a substance into the body via a relevant route of administration, distribution within the body, metabolism, and excretion can be a significant factor in decisions on the relevance of carcinogenicity tests in lab animals to humans. Interspecies similarity of metabolic products in the body is another significant consideration. Data that help justify extrapolation of experimental results to humans will increase the weight given to results on carcinogenicity in the species. If no data are available, it can be assumed that the carcinogenicity results are relevant to humans and classification of the substance can proceed accordingly. However, if there is convincing scientific information that demonstrates that the mechanism or mode of action in animals is not relevant to humans, a lower classification or non-classification may be warranted (HCS, paragraph A.0.3.4).

(2) Similarity of the mechanism of action (mode of action) of the substance in laboratory animals and humans gives significant weight to the extrapolation of data from studies in those animals to people. As an

example, mechanistic studies performed *in vitro* might provide additional weight if, combined with other information, they help to elucidate the mode of action of a substance on a target site and, by doing so, provide information on the potential relevance of observed toxicity in laboratory settings to toxicity that might occur in humans. Similarly, data from a valid study in lab animals can be given additional weight if it is accompanied by a study showing that the test material has a mode of action in humans that is consistent with the effects observed in lab animals and so the data can be reasonably used to decide if the same mode of action is expected to occur in people.

(3) Conversely, studies that demonstrate species-specific modes of action in laboratory animals might cast doubt on the relevance of findings in those animals to humans. When there is clear and compelling scientific evidence demonstrating that the mechanism or mode of action is not relevant to humans (as in the case of some non-genotoxic carcinogens), the chemical should not be classified (HCS, paragraph A.0.3.4).

(4) Structural similarity to a substance for which there is good evidence of carcinogenicity can be a useful component in WoE. Information on the similarity of ADME between the untested substance and the tested analog can be useful in this regard. If the formation of active metabolites is a step in the mode of action of the tested analog, information on the similarity of metabolites formed from an untested substance would be highly useful in decisions on the appropriateness of the use of a structural analog.

(5) A number of biological tests are available that measure endpoints related to carcinogenicity without the costs in animals, time, and money involved in a full rodent carcinogenicity assay. Many of these are designed to measure mutations *in vivo* or *in vitro*. Descriptions of some of the widely used and accepted tests are in the Health Effects Test Guidelines from the US EPA's Office of Chemical Safety and Pollution Prevention. The OECD Guidelines for the Testing of Chemicals provides descriptions of a similar set of tests for health effects. Examples of other related tests include *in vivo* organ-specific studies (e.g., preneoplastic lesions in rodent liver as indicators of potential carcinogenicity, skin-painting in mice to assess dermal carcinogenicity, or accelerated development of tumors in genetically susceptible animals) and studies in transgenic animals that are genetically engineered with reporter genes for detection of mutations. Some transgenic models can be used as screening assays for carcinogenicity, particularly in combination with other relevant data. Even though these studies do not directly determine carcinogenicity, they can provide useful information that can help in the determination of the overall WoE for decisions on classification. However, these tests generally are not sufficient to establish a lack of carcinogenicity.

(6) In some instances, non-guideline tests are used that provide valuable information that might lead to classification of a substance. The HCS

does not specify the tests that are to be performed. In fact, paragraph A.0.2.3 states that "Any test that determines hazardous properties, which is conducted according to recognized scientific principles, can be used for purposes of a hazard determination for health hazards. Test conditions need to be standardized so that the results are reproducible with a given substance, and the standardized test yields 'valid' data for defining the hazard class of concern." Therefore, non-guideline tests can be used if they meet these criteria and the use of the results from the tests for hazard classification is scientifically valid.

(7) Additional information for the WoE evaluation might be obtained from other sources, such as modeling or an evaluation of structure-activity relations with other substances. In such cases, the scientific validity of the use of this information must be carefully reviewed and, if appropriate, the information added to the WoE.

3.2.3.3 Weight of evidence for carcinogens: Non-mandatory guidance

The HCS also provides non-mandatory guidance described in Appendix F on the classification of carcinogens. This additional guidance includes (1) background guidance from GHS that is based on the preamble to "Monographs on the Evaluation of Carcinogenic Risks to Humans" published by IARC, (2) information from IARC on classification of carcinogens, (3) listing criteria from the "Report on Carcinogens" from NTP, and (4) a comparison of classifications under GHS, IARC, and NTP.

The background guidance provided by IARC offers additional discussion of factors to consider in making decisions by WoE, similar to those discussed in this document. IARCs information on classification and NTP's listing criteria explain the basis for their respective decisions on classification of carcinogens. A table on the "Approximate Equivalences among Carcinogen Classification Schemes" is provided to help classifiers use existing classifications under IARC and/or NTP to establish classification of the same substance or mixture under GHS and thus under the HCS.

Expert committees at IARC and NTP review existing epidemiological and/or toxicity studies, including data on the relevance of MOA/mechanism in laboratory animals to potential hazards in humans. Therefore, in some cases, classifications based on review of such information may have already been performed and the resulting conclusions may be used.

Alternately, instances might occur in which new information becomes available but it has not yet been reviewed by IARC, US EPA, or other bodies. If, for example, a substance has been previously classified as a possible carcinogen by IARC based on data from laboratory animals, and new information demonstrates that the mode of action (MOA) shows that the experimental findings are not relevant to humans, then a classification that is lower than that from IARC (or

other well-regarded bodies) might be justified. While this situation is rare, one example is Di(2-ethylhexyl) phthalate (DEHP). IARC classified DEHP in 1987 as a *Group 2B* – human carcinogen based on sufficient evidence of liver tumors in rats and mice. This IARC classification would trigger a GHS hazard classification of 'presumed/suspected' human carcinogen. In 2000, IARC re-classified [i.e., downgraded] DEHP as a *Group 3* - with the following explanation:

DEHP is *not classifiable as to its carcinogenicity to humans (Group 3)* because peroxisome proliferation has not been documented in human hepatocyte cultures exposed to DEHP nor in the liver of exposed non-human primates. Therefore, the mechanism by which DEHP increases the incidence of hepatocellular tumors in rats and mice is not relevant to humans. (IARC, 2000).

However, OSHA considers the IARC and NTP classification process equivalent to the HCS weight of evidence approach and, therefore, their determination as sufficient evidence in establishing the classification of a carcinogen. If the classifier uses the determinations of IARC or NTP then they do not have to conduct their own weight of evidence evaluation with regards to carcinogenicity. However, if the classifiers do perform their own hazard evaluation and their determination differs from that of IARC and/or NTP, OSHA would look, in the event of a compliance inspection, for a clear justification for the difference. OSHA may request the results and underlying data that support the classification and the weight-of-evidence determination. The classifier must ensure that **all** reliable and relevant information on a chemical has been reviewed, and the classifier also should clearly explain why the weight of the negative evidence would lead to a different conclusion/classification than the conclusions of IARC and/or NTP. In particular, OSHA would consider whether the classifier used the same general strategies as discussed above addressing issues of data quality and bias in the studies they considered. However, regardless of the classification, the classifier must note on the SDS that the chemical has been to found to be a potential carcinogen by IARC or is listed in NTP's Report on Carcinogens.

3.2.3.4 Weight of evidence for carcinogens: Examples of the use of WoE

Following are examples of the use of WoE in the determination of classification of carcinogenicity under the HCS. Each example begins with a table that contains a summary of available data in the left-hand column. The other column shows the possible level of evidence, a tentative interpretation of the study until a determination of the weight of evidence of the study is made. The studies that would not be sufficient to justify classification by themselves are identified as "supporting." While they do not individually lead to decisions on classification, they are used collectively in a WoE evaluation, which is presented in each example here in an abbreviated form. Note that ADME means the absorption, distribution, metabolism, and excretion of the substance being evaluated.

Example #1 of Test Results and Classification for Carcinogenicity	
Data	Possible Level of Evidence
Clearly positive chronic study in 2 rodent species	Sufficient evidence in lab animals
Positive HPRT assay for gene mutation in cultured hamster cells (*in vitro*)	Supporting
Positive micronucleus assay in mice *in vivo* (chromosomal damage)	Supporting
Structural analogs: none	
ADME similar in humans and rodents	Supporting

Weight of Evidence: Studies were performed in accordance with established test guidelines and no confounding factors were noted in the studies. The substance was clearly carcinogenic in two species in rodent bioassay, providing sufficient evidence of carcinogenicity for classification. Chromosomal damage in mammals *in vivo* and data supporting extrapolation from rodents to humans provided supporting evidence. Mutagenicity in mammalian cells *in vitro* was consistent with these results.

Conclusion: Category 1B, Presumed human carcinogen. An assessment of the relevance to humans would be advisable to help solidify this conclusion.

Note that in this example the possible classification as Category 1B seems relatively straightforward because of the clearly positive study in rodents. However, this classification could be downgraded to a Category 2 or possibly "not classified" if, for example, conclusive additional data were available that showed the substance acted exclusively through a mechanism that was specific to the tested species and had little or no relevance to humans. Alternately, classification as Category 1A might be more appropriate for this substance if additional human data considered to be sufficient evidence becomes available. Similar caveats apply to the other following examples, stressing the need for careful case-by-case review of all information.

Example #2 of Test Results and Classification for Carcinogenicity	
Data	Possible Level of Evidence
Clearly positive chronic study in rats, marginal in mice	Limited evidence in lab animals
Positive HPRT assay for gene mutation in cultured hamster cells (*in vitro*)	Supporting
Negative micronucleus assay in mice *in vivo* (chromosomal damage)	
Structural analog is Category 2, suspected carcinogen based on limited evidence. Kinetics and metabolism very similar to test substance	Supporting
ADME similar in humans and rodents	Supporting

Weight of Evidence: Studies were performed in accordance with established test guidelines and no confounding factors were noted in the studies. Tumorigenicity in the chronic study was limited mainly to one of two species, providing limited evidence of carcinogenicity. A structural analog is already classified as HCS Category 2 and the metabolism and kinetics of the tested substance and analog are similar and support extrapolation to humans. Mutagenicity in mammalian cells *in vitro* provided supporting evidence. The classification could be upgraded with some limited human evidence of carcinogenicity combined with a more convincing mode of action relevant to humans.

Conclusion: Category 2, Suspected human carcinogen based on limited evidence. If the classifier has or obtains additional evidence (e.g., the relevance of mode of action to humans) the classification must be reevaluated and may result in reclassification.

Example #3 of Test Results and Classification for Carcinogenicity	
Data	Possible Level of Evidence
Clearly positive tumor study in transgenic mice.	Limited evidence in lab animals
Positive HPRT assay for gene mutation in cultured hamster cells *in vitro*	Supporting
Structural analog is Category 2, suspected carcinogen based on limited evidence. Kinetics and metabolism are very similar to test substance.	Supporting
Equivocal interspecies comparison of ADME	Supporting

Weight of Evidence: Studies were performed in accordance with established test guidelines and no confounding factors were noted in the studies. Results in transgenic mice might meet the criteria for Category 2 (limited evidence in lab animals), but a thorough evaluation of the specific transgenic model, (i.e., its effectiveness in predicting carcinogenicity, and its acceptance by regulatory bodies) is needed. A structural analog is already classified as HCS Category 2 and the metabolism and kinetics of the tested substance and analog are similar. However, data on interspecies extrapolation from mice to humans were marginal. Mutagenicity in mammalian cells *in vitro* provided supporting data for classification as a suspected carcinogen.

Conclusion: Category 2, Suspected human carcinogen. However, this is dependent of strength of the results and validation of the assay in transgenic mice. Consistently positive results in well validated transgenic animal assays could raise the classification to Category 1b. This would be based on sufficient evidence in animals supported by positive *in vitro* data.

Example #4 of Test Results and Classification for Carcinogenicity	
Data	Possible Level of Evidence
Positive mouse spot test for mutagenicity in mammals *in vivo*	Supporting
Positive test in mammalian bone marrow for chromosomal aberration *in vivo*	Supporting
Positive Ames test for mutations in bacteria *in vitro*	Supporting
Structural analog considered mutagenic, but no decision on carcinogenicity	Supporting
No data on similarity of ADME between humans and lab animals	

Weight of Evidence: Studies were performed in accordance with established test guidelines and no confounding factors were noted in the studies. There are no direct data on carcinogenicity *in vivo*. The substance was mutagenic in mammals *in vivo* and caused chromosomal aberrations in mammals *in vivo*. Positive test for mutagenicity in bacteria provides supporting information. Data on mutagenicity of a structural analog do not add significant weight to a decision. There are no data on ADME of the substance.

Conclusion: Not classified. However, the data should be put in Section 11 of the SDS to inform users of a potential hazard. Also, the chemical manufacturer might consider further investigation to resolve the question of potential carcinogenicity in order to avoid issues such as potential future liability. NTP might also regard this as a strong candidate for testing.

3.3 Weight of evidence and germ cell mutagens

In lieu of positive evidence from human epidemiological studies that would lead to classification, decisions on classification for germ cell mutagens under the HCS can be based on total WoE or a single well-conducted test with "clear and unambiguously positive results." The classification for heritable effects in human germ cells can be based on mutagenic and/or genotoxic effects in germ and/or somatic cells of exposed animals. Effects in *in vitro* tests shall also be considered and the hierarchy of weight given to different types of data in section 3.1 of this guidance applies to mutations in germ cells. Also, although the specifics of the tests are different, the general approaches to the determination of the WoE of each test are the same here as in section 3.2.3.1 of this guidance.

The following are examples of the use of WoE in the determination of classification of germ cell mutagens under the HCS. Each example begins with a table that contains a summary of available data in the left-hand column. The other column shows the possible level of evidence, which is a tentative interpretation of the study until a determination of the weight of evidence of the study is made. Studies that are consistent with classification but are not sufficient to justify classification by themselves are identified as "supporting." While they do not

individually lead to decisions on classification, they are used collectively in a WoE evaluation, which is presented in each example here in an abbreviated form.

Example #1 of Test Results and Classification for Germ Cell Mutation	
Data	Possible Level of Evidence
Clearly positive results in mouse visible specific locus test	Sufficient in Lab Animals
Positive HPRT assay for gene mutation in cultured hamster cells *in vitro*	Supporting
No structural analogs	
ADME similar in humans and rodents	Supporting

Weight of Evidence: Studies were performed in accordance with established test guidelines and no confounding factors were noted in the studies. The positive results in a mouse visible specific locus test were definite and provided evidence of mutations in germ cells of a mammalian species *in vivo*. This result is sufficient for classification as a Category 1B. The positive test for mutations in mammalian cells *in vitro* provides supporting data. The similarity of kinetics and metabolism in mice and humans provides additional weight for considering germ cell mutations to be possible in humans.

Conclusion: Category 1B, substances regarded as if they induce heritable mutations in germ cells of humans as shown by the specific locus test.

Example #2 of Test Results and Classification for Germ Cell Mutation	
Data	Possible Level of Evidence
Positive mouse spot test for mutations in somatic cells *in vivo*	Limited evidence in lab animals
Positive HPRT assay for gene mutation in cultured hamster cells *in vitro*	Supporting
Positive mutation assay in mouse lymphoma cells *in vitro*	Supporting
No data on structural analogs	
No data on ADME	

Weight of Evidence: Studies were performed in accordance with established test guidelines and no confounding factors were noted in the studies. These tests are not designed to assess mutations in germ cells directly. However, the HCS states that somatic cell mutagenicity tests *in vivo* (in mammals) or other *in vivo* somatic cell genotoxicity tests which are supported by positive results from *in vitro* mutagenicity assays can be used to establish classification as Category 2. The positive results in the spot test indicated mutations in somatic cells of fetal mice before birth when their mother is exposed to the test substance. These *in vivo* results were supported by positive *in vitro* assays for mutations in two types of

mammalian somatic cells. Collectively these results clearly fulfill the criteria for classification as Category 2.

Conclusion: Category 2, Substances which cause concern for humans owing to the possibility that they may induce heritable mutations in the germ cells of humans. Category 1B was not chosen because, although the *in vivo* test in somatic cells was positive, there was not supporting "evidence that the substance has potential to cause mutations to germ cells" (as stipulated in the HCS).

Example #3 of Test Results and Classification for Germ Cell Mutation	
Data	Possible Level of Evidence
Positive bacterial reverse mutation test (Ames test)	Supporting
Negative micronucleus assay in mice *in vivo* (chromosomal damage)	
Structural analogs: none	
ADME similar in humans and rodents	Supporting

Weight of Evidence: Studies were performed in accordance with established test guidelines and no confounding factors were noted in the studies. The positive mutation assay was not sufficient for classification since it was in bacteria (not mammals or in other organisms with nucleated cells) and *in vitro* rather than *in vivo*. As such, results from this test without other positive results in other tests are not sufficient for a Category 2 classification. The micronucleus assay provided no evidence of genotoxicity.

Conclusion: Not classified. However, the positive mutation assay might warrant follow-up studies.

3.4 Weight of evidence for reproductive toxicants

The HCS provides considerable guidance on what must be considered during the determination of whether a substance is a potential reproductive hazard. Particular attention is given to the influence of maternal toxicity and also to the need for determination of whether observed effects are directly attributable to the substance or those effects are nonspecific secondary results of effects of the substance on other organs in the body.

In lieu of positive evidence from human epidemiological studies that would lead to classification, decisions on classification for reproductive hazards under the HCS can be based on total WoE. In some cases a single positive study "performed according to good scientific principles and with statistically or biologically significant positive results may justify classification" (A.7.2.3) Although the specifics of the tests are different, the general approaches to the determination of the WoE of each test are the same here as in section 3.2.2.1 of this guidance.

Following are examples of the use of WoE in the determination of classification of reproductive hazard under the HCS. Each example begins with a table that contains a summary of available data in the left-hand column. The other column shows the possible level of evidence, a tentative interpretation of the study until a final determination of the WoE of the study is made. The studies that are consistent with classification but are not sufficient to justify classification by themselves are identified as "supporting." While they do not individually lead to decisions on classification, they are used collectively in a WoE evaluation, which is presented in each example here in an abbreviated form.

Example #1 of Test Results and Classification for Reproductive Hazard	
Data	Possible Level of Evidence
Clearly positive results at doses of 100 and 400 mg/kg in OECD 414 (Prenatal Developmental Toxicity Test) with pronounced dose-related effects on fetuses and minimal maternal toxicity	Sufficient evidence in laboratory animals
No effects on reproductive organs (organ weights and histopathology) in 90-day study at 100 and 400 mg/kg. Effects on other organs only at 400 mg/kg.	Supporting
No data on structural analogs	
No data on ADME	

Weight of Evidence: Studies were performed in accordance with established test guidelines and no confounding factors were noted in the studies. Clear adverse effects occurred in fetuses of treated female rats at two doses. Maternal toxicity was minimal. Fetal effects were dose-related and did not appear to be secondary to effects on the dams (mothers). No adverse effects were seen on reproductive organs of the dams or in a 90-day study in rats at the same doses. Fetal effects appear to be specific to the fetuses. No information on ADME or structural analogs is available.

Conclusion: Category 1B, presumed human reproductive toxicant based on evidence in experimental animals. Note that no deficiencies were seen in the study on developmental toxicity. However, the evidence would be less convincing if significant deficiencies were found and, in such a case, classification as Category 2 might be more appropriate than Category 1B.

Example #2 of Test Results and Classification for Reproductive Hazard	
Data	Possible Level of Evidence
Dose-related higher incidence of fetal resorptions and lower fetal weight at doses ≥10 mg/kg in OECD 414 (Prenatal Developmental Toxicity Test). Dose-related effects on multiple organs in mothers at same doses.	Sufficient evidence of fetal effects in laboratory animals, but associated with maternal toxicity
No effects on reproductive organs (organ weights and histopathology) in 90-day study at same doses, but effects on other organs were the same as in the developmental toxicity study.	Supporting
Effects on fetuses and adults with structural analogs are similar to this substance, but classification of analogs is not yet finalized.	
No data on ADME	

Weight of Evidence: Studies were performed in accordance with established test guidelines and no confounding factors were noted in the studies. Maternal toxicity was seen over the same range of doses as higher incidence of fetal resorptions and lower fetal weight, but the effects on mothers were not as significant as increased fetal deaths. Therefore, although it is not known if some of the fetal effects are secondary to maternal toxicity, the fetal effects are serious enough that this possibility does not lessen the need for classification as a reproductive toxicant. No effects were seen on reproductive organs during the developmental toxicity or the subchronic study, but effects on other organs were consistent between both studies. Information on structural analogs is available, but does not aid with classification.

Conclusion: Category 2 suspected human reproductive toxicant, based on definite fetal toxicity in the presence of maternal toxicity.

Example #3 of Test Results and Classification for Reproductive Hazard	
Data	Possible Level of Evidence
A trend for dose-related malformations in fetuses in an OECD 414 (Prenatal Developmental Toxicity Test) with minimal maternal toxicity. Statistical significance was not achieved with the fetal malformations and incidence was at the upper end of data from historical controls.	Non-statistically significant trends in laboratory animals
No effects on reproductive organs (organ weights and histopathology) or other organs in 90-day study at same doses.	Supporting
The same test performed on each of two structural analogs yielded essentially identical suggestive, but inconclusive, results.	Supporting
No data on ADME	

Weight of Evidence: Studies were performed in accordance with established test guidelines and no confounding factors were noted in the studies. Data from identical studies with three structural analogs were inconclusive individually, showing only dose-related trends rather than definite effects. However, collectively these studies indicate substance-related effects on the fetuses are likely. No adverse effects were seen on reproductive organs of the dams or in a 90-day study in rats at the same doses. Fetal effects appear to be specific to the fetuses. No information on ADME.

Conclusion: Category 2, presumed human reproductive toxicant based on evidence in experimental animals, based on the collective data from multiple studies. Classification is warranted, although investigation of the mode of action should be considered.

Example #4 of Test Results and Classification for Reproductive Hazard	
Data	Possible Level of Evidence
Clear dose-related reductions in fertility index, sperm count, and testicular weight in F1 generation of two-generation study in rats. Testicular atrophy seen histologically.	Sufficient evidence in lab animals
Definite dose-related reduction in testicular weight in 90-day study in rats. Testicular atrophy seen histologically.	Supporting
No data on structural analogs	
ADME generally similar in rats and humans	Supporting

Weight of Evidence: Studies were performed in accordance with established test guidelines and no confounding factors were noted in the studies. Dose-related testicular atrophy, lowered testicular weight, lowered sperm counts were associated with dose-related reductions in fertility index. Quantitative values were statistically significantly different from those of controls. Effects on other organs, particularly liver, were observed but were judged not to be associated with effects on reproduction. Similarity of ADME in rats and humans supports potential use of the results with people. No data on structural analogs were available. Route of dosing was relevant to expected human exposure.

Conclusion: Category 1B, presumed human reproductive toxicant based on results in laboratory animals.

Example #5 of Test Results and Classification for Reproductive Hazard	
Data	Possible Level of Evidence
Marginal alterations in estrous cycle at highest dose (1000 mg/kg/day) in F1 generation of two-generation study in rats. Marginally (not statistically significant) lower fertility index at highest dose in F1 generation. Ovarian weights normal.	Equivocal evidence in lab animals
Marginally lower ovarian weights at highest dose (1000 mg/kg/day) in 90-day study in rats. No alterations were seen histologically.	Supporting
No data on structural analogs	
ADME generally similar in rats and humans	Supporting

Weight of Evidence: Studies were performed in accordance with established test guidelines and no confounding factors were noted in the studies. Marginal effects were seen for estrous cycle and fertility index only at the highest dose recommended in the guidelines. Effects were in the F1 generation and not in the parental generation. Marginally lower ovarian weights observed in a shorter 90-day study were not seen in the two-generation study, and no morphological effects were noted in the ovaries. Similarity of ADME in rats and humans supports potential use of the results with people. No data on structural analogs were available. Route of dosing was relevant to expected human exposure.

Conclusion: Not classified, although OSHA recommends that the relevant results of the two-generation study should be stated on the SDS.