



Toxicology and Epidemiology: Improving the Science with a Framework for Combining Toxicological and Epidemiological Evidence to Establish Causal Inference

Hans-Olov Adami,^{*†} Sir Colin L. Berry,[‡] Charles B. Breckenridge,[§] Lewis L. Smith,[¶] James A. Swenberg,^{||} Dimitrios Trichopoulos,^{*} Noel S. Weiss,^{||}# and Timothy P. Pastoor[§]¹

^{*}Department of Epidemiology, Harvard School of Public Health, Boston, Massachusetts 02215; [†]Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden; [‡]Queen Mary University, London, UK; [§]Toxicology & Health Sciences Department, Syngenta Crop Protection, Greensboro, North Carolina 27419; [¶]MRC Toxicology Unit The College of Medicine, University of Leicester, Leicester, UK; ^{||}Syngenta Crop Protection Ltd, Basle, Switzerland; ^{||}Department of Environmental Sciences and Engineering, University of North Carolina, Chapel Hill, North Carolina 27599; ^{||}Department of Epidemiology, University of Washington, Seattle, Washington 98195; and [#]The Fred Hutchinson Cancer Research Center, Seattle, Washington 98109

¹To whom correspondence should be addressed at PO Box 18300, Greensboro, NC 27419. Fax: (336) 632-2226. E-mail: tim.pastoor@syngenta.com.

Received March 16, 2011; accepted May 4, 2011

Historically, toxicology has played a significant role in verifying conclusions drawn on the basis of epidemiological findings. Agents that were suggested to have a role in human diseases have been tested in animals to firmly establish a causative link. Bacterial pathogens are perhaps the oldest examples, and tobacco smoke and lung cancer and asbestos and mesothelioma provide two more recent examples. With the advent of toxicity testing guidelines and protocols, toxicology took on a role that was intended to anticipate or predict potential adverse effects in humans, and epidemiology, in many cases, served a role in verifying or negating these toxicological predictions. The coupled role of epidemiology and toxicology in discerning human health effects by environmental agents is obvious, but there is currently no systematic and transparent way to bring the data and analysis of the two disciplines together in a way that provides a unified view on an adverse causal relationship between an agent and a disease. In working to advance the interaction between the fields of toxicology and epidemiology, we propose here a five-step “Epid-Tox” process that would focus on: (1) collection of all relevant studies, (2) assessment of their quality, (3) evaluation of the weight of evidence, (4) assignment of a scalable conclusion, and (5) placement on a causal relationship grid. The causal relationship grid provides a clear view of how epidemiological and toxicological data intersect, permits straightforward conclusions with regard to a causal relationship between agent and effect, and can show how additional data can influence conclusions of causality.

Key Words: epidemiology; causation; framework.

THE INTERSECTION OF EPIDEMIOLOGY AND TOXICOLOGY

In 1775, Percivall Pott concluded, on the basis of clinical observations, that scrotal cancer in chimney sweeps was caused

by chimney soot (Potter, 1962). It was almost 140 years before experimental confirmation of this was produced by Yamagiwa and Ichikawa (1918). By repeated painting of rabbit ears with a coal tar extract, they produced epithelial skin tumors, powerfully corroborating what Pott had seen in humans. In this case, an inference of causation in humans was arrived at through a combination of the two scientific disciplines.

Subsequently, animal studies were used to verify other epidemiological findings, serving to establish Koch’s third postulate: the agent should cause the disease when introduced into a healthy organism (Koch, 1884, 1893). Although Koch’s original intent was proving disease causation by microbiological pathogens, this third postulate has also been applied to corroborating chemical-related epidemiological findings in humans, by testing in animals.

Although there are a number of examples of how epidemiology and toxicology intersected over time, perhaps the most notable case is tobacco smoke and lung cancer. By 1964, there was ample epidemiological evidence for a causal connection between lung cancer and smoking tobacco products; at that point, the U.S. Surgeon General (Bayne-Jones *et al.*, 1964) accepted the relationship as causal. Yet at that time, toxicologists could not reproduce similar tumors in animal models. This lack of concordance emphasized the difficulty of using Koch’s postulates that were established for infectious disease in chemical-related pathogenesis. Toxicological corroboration of epidemiological evidence later became the element of “biological plausibility” in Hill’s guidelines for establishing causality (Hill, 1965). Hill and others (Bayne-Jones *et al.*, 1964) effectively modified Koch’s third postulate from an orientation of proof to one of plausibility.

Whereas the concordance was high between cancer-causing agents initially discovered in humans and positive results in animal studies (Tomatis *et al.*, 1989; Wilbourn *et al.*, 1984), the same could not be said for the reverse relationship: carcinogenic effects in animals frequently lacked concordance with overall patterns in human cancer incidence (Pastoor and Stevens, 2005). This lack of concordance between toxicology and epidemiology might arise because the high doses used in animal studies to produce tumors are not typically seen in human populations.

Nonetheless, toxicology took on a predictive rather than a confirmatory role by providing alerts for potential effects in humans, whether carcinogenic, neurotoxic, hepatotoxic, or any other adverse outcome. These alerts became the basis for regulating chemical exposure to humans. The underlying assumption was that restricting exposure well below levels at which adverse effects were seen in animals would prevent harmful outcomes in humans.

Thus, the relationship between epidemiology and toxicology has shifted over time. Both disciplines seek to contribute data relating to the causes of human disease and occasionally lean on each other to support propositions of causality. Toxicologists and epidemiologists alike spend considerable time and effort characterizing the relationship between the putative causal agent and a response (Fig. 1). Many of the same fundamental considerations are part of the evidence-based analysis that takes place by scientists in the two disciplines. However, the two fields could arguably be said to work in parallel rather than in concert. Can toxicological experimentation augment a weak positive epidemiological finding? Conversely, when and how does low biological plausibility influence a positive epidemiological finding? Separately, the

fields can derive conclusions based on paradigms illustrated in Figure 1. Together, conclusions of causality can be more firmly based, further investigations can be clearly identified, and improvements in human health protection can be achieved. In addition to highlighting the history of relevant developments in the fields, we suggest a way that the two disciplines can come together to better understand the impact, potential or real, of agents on human health.

CAUSAL INFERENCE

Process for Causal Inference

The disciplines of toxicology and epidemiology ask the question: *can* a substance cause a particular effect in humans. The data obtained in toxicological and epidemiological studies do not always lead to a straightforward interpretation, and often different observers will differ in their conclusions. Even for associations that are widely regarded as causal today—such as ingestion of water contaminated with the bacterium *Vibrio cholerae* and the incidence of cholera or cigarette smoking and the incidence of lung cancer—for some years after relevant data became available, there was considerable disagreement as to the presence of a cause-effect relation in each instance. Indeed, a principle underlying the philosophy of science is that causality cannot be “proven”; it can only be inferred with different degrees of certainty.

Epidemiological investigation of a null hypothesis that postulates that a variable has no effect on a health outcome can never be established to be true (Popper, 1959); there can only be a failure to show that the null hypothesis is false within the

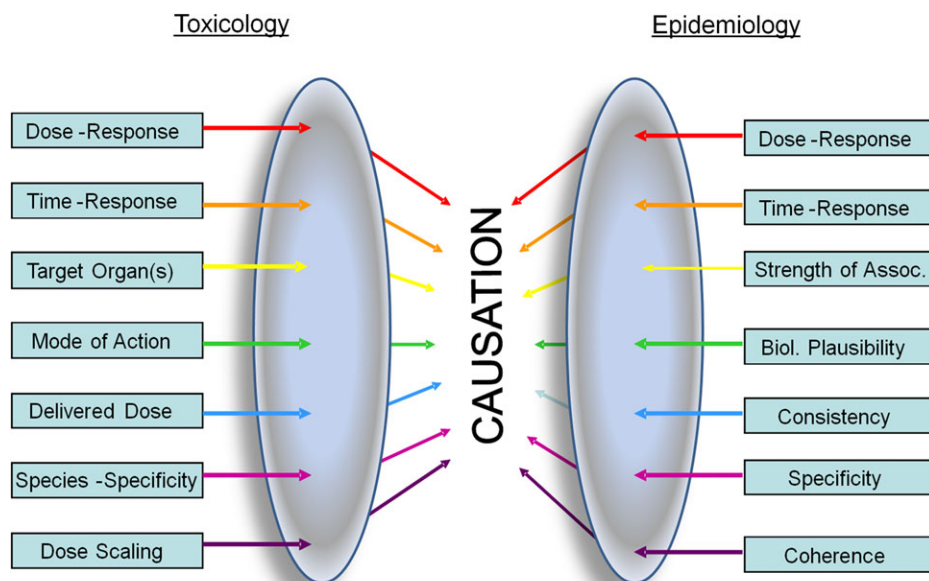


FIG. 1. Contribution of toxicology and epidemiology data to causal inference. Many of the same principles contribute to evidence-based decisions in the two fields. Together, causation can be more accurately inferred.

limits of specific study designs. Theories that integrate observations from multiple studies, or rely on other biological considerations, are useful when they make testable predictions. Hypotheses that are not testable do not fall within the realm of science. Likewise, expert opinion should be supported by evidence for rational science-based decision making (Guzelian *et al.*, 2005).

Because Hill (1965) and others (Bayne-Jones *et al.*, 1964) articulated their perspectives on causal inference, scientists have further described methods to systematically review and characterize the evidence that might be used to support an inference of causality (Cole, 1997; ECETOC, 2009; Kundi, 2007; Phillips and Goodman, 2004; Rothman 1976; Rothman and Greenland, 2005; Susser, 1986; Weed, 2005). We suggest an expert judgment process for integrating the totality of the epidemiological findings in a weight of evidence framework. This integration takes note of the literature cited above but extends this thinking by offering a method to systematically consider biological plausibility and epidemiological evidence in a process to unite epidemiology and toxicology in a framework to infer causality.

Applications of Causal Inference in Epidemiology

Epidemiological studies document the occurrence of illness or injury in human populations. Depending on the design, epidemiological studies can provide evidence bearing on a causal relationship. For example, quantification of the efficacy of pharmaceutical agents in humans is often based on randomized controlled studies, where “exposed” and “nonexposed” persons are similar with regard to other characteristics that bear on the outcome in question. However, the focus of this paper is on causal inference for environmental agents (primarily synthetic chemicals). Because randomized trials with environmental agents are rarely feasible or perhaps ethical, this study design will not be discussed here.

Studies aimed at evaluating environmental chemicals and other environmental factors are generally nonrandomized observational studies with an ecologic, case-control, or cohort design. Although these studies are fundamental in gauging possible human health effects, their design may limit the extent to which inferences about causality can be drawn. Because observational studies do not randomly allocate subjects to exposure, interpretation of the results of these studies must take into account any differences, or the possibility of differences, between exposed and nonexposed subjects. A brief description of these observational studies and their strengths and weaknesses follows.

Ecologic Studies Ecologic studies contrast the incidence of disease across populations (or population subgroups) that differ in terms of presence or degree of an exposure to an environmental factor. The incidence of disease among different population subgroups may be evaluated on the basis of, for

example, geographical differences or changes in disease incidence over time within a population.

Ecologic studies have the potential to contribute to our understanding of exposure-disease relationships if . . .

the environmental exposure level can be ascertained with reliability,

there are large differences in exposure,

the incidence of the disease is ascertained in a comparable manner, and

there is little or no difference in the presence of other causes of the disease.

For example, aflatoxin (a toxic product of *Aspergillus flavus*) was found early on in experimental evaluations in animals to be an extremely potent carcinogen. At that time, the only relevant data in humans took the form of correlations of liver cancer mortality rates across population groups with marked differences in estimated aflatoxin intake. The positive correlation observed in these studies was open to alternative interpretations—the populations with the highest rates differed in ways other than exposure to aflatoxin such as the prevalence of Hepatitis B infection—so it was largely the strength of the laboratory evidence that served as a basis for a tentative causal inference. Later, stronger epidemiologic data became available supporting a causal effect. In particular, there were ecologic studies with less potential for confounding and nested case-control studies in which prediagnosis urinary markers of aflatoxin intake could be assessed (Qian *et al.*, 1994; Wang *et al.*, 1996).

In practice, causal inferences that can be drawn from the results of ecologic studies may be limited because:

Within a given population, exposure characterization may not have been carried out (or carried out well and in a similar way) over time or among population subgroups. Furthermore, within a population, the actual variation in exposure levels may be small, making it difficult for an epidemiological study to reliably document the differences in occurrence of disease. The weaker the association, the more difficult it is to distinguish it from an association that arises by chance or confounding.

The completeness of ascertainment of the disease condition can vary by place and time. This is particularly a problem for a condition in which diagnostic criteria are difficult to apply consistently (e.g., autism, Parkinson’s disease, non-Hodgkin lymphoma) but can also be present when differential disease screening occurs as a function of location and time (e.g., prostate specific antigen testing for prostate cancer).

In order to maximize the contrast in exposure prevalence or levels across geographic units, many ecologic studies compare disparate geographic populations for disease occurrence. For example, most ecologic studies that have examined the association between dietary fat and breast cancer incidence have compared national populations from around the world where data on both diet and cancer incidence were available.

This approach allowed for the inclusion of populations in which a variety of dietary intakes was present but made it difficult to interpret whether the association seen (i.e., higher fat intake associated with higher breast cancer incidence) was because of the dietary differences in fat intake or to differences in one or more of the other characteristics of these populations. The absence of a relationship between dietary fat and breast cancer incidence was suggested by the results of cohort studies where little or no association between dietary fat intake and breast cancer risk was observed among individuals within certain populations (Hunter *et al.*, 1996). These data argue that the strong positive association observed in ecologic studies was in fact a reflection of the confounding influence of one or more characteristics that were associated with both diet and breast cancer risk (Colditz *et al.*, 2006).

Case-Control Studies Case-control studies ascertain the proportion of persons who previously experienced one or more exposures among persons with a disease (cases) and a sample of persons representative of the person-time from which the cases were generated (controls).

Exposure ascertainment is a potential source of bias in case-control studies, notably in studies investigating environmental exposures because they may be incompletely or inaccurately reported or recorded and misclassification may vary by case-control status. Thus, the results can indicate either spuriously high or spuriously low estimates of the magnitude of any association. Direct measurements of blood or tissue levels of chemical exposures (or metabolites of these chemicals) obtained after diagnosis in the cases may not reflect earlier levels of exposure because the illness and its treatment may have led to an alteration in these levels. Even if exposure levels of cases were unaffected by disease status among cases, the levels measured at the time of the study may not be indicative of those present earlier in life when critical pathogenic events occurred.

Unless a case-control study can overcome the difficulties of valid retrospective ascertainment of exposure status, it cannot be confidently relied upon to provide a valid estimate of the association between an exposure and a disease. Cross-sectional studies, in which current exposure levels are compared between persons with and without a given condition at the time of the assessment of the exposure (irrespective of when that condition first developed), are particularly problematic in this regard.

In unusual circumstances—when the proportion of ill individuals with a history of a given exposure far exceeds what might be expected—an association can be inferred without the need for a formal control group. For example, because all cases of a form of pneumonia in an area of Spain during a relatively short period of time reported ingestion of adulterated rapeseed oil (Tabuenca, 1981), it was reasonable to infer a causal connection (and to take preventive action) prior to the enrollment of controls into this study.

Cohort Studies It often happens that the same chemicals to which one or more communities are exposed also are encountered in persons who work in the manufacture or distribution of these chemicals. Because these exposures tend to be higher than those received in a community at large, any impact on disease risk from exposure to the agent is likely to be greater in magnitude in the exposed workforce and therefore easier to ascertain in an epidemiologic study.

Because it is often possible to identify workforce members and monitor their status through vital records and disease registers, epidemiologic studies based on the workers' experience are feasible. The results of occupational cohort studies can nevertheless be difficult to interpret because of the presence of multiple chemical exposures on the job and, particularly in retrospective cohort studies, difficulties in accounting for prior work history and other disease-causing exposures not ascertained in available records such as smoking history. Nonetheless, occupational cohort studies have contributed a great deal to our understanding of health effects of chemical exposures and, when available, can assume a prominent place in the evaluation of the safety of exposure to chemicals.

Applications of Causal Inference in Toxicology

In toxicology, the test agent is given to the animal or *in vitro* cellular system under clearly defined exposure conditions (e.g., oral, dermal, inhalation; gavage, diet; short term, long term, etc.). The physiological status of each test group is compared with the untreated group. From this body of data, a toxicologist must then decide which responses are exposure related and determine whether the responses observed are relevant to humans. In the absence of evidence to the contrary, the toxicologist assumes that findings in animals are likely to be relevant to human health.

However, as our understanding of biological systems has evolved, we realize that effects in animals may not be relevant to humans. The need for this important distinction depends on either qualitative differences in biology or quantitative differences between animals and humans in the kinetics of the chemical or the dynamics of the response.

A systematic way of drawing conclusions of human relevance (causality) was first proposed by the International Programme on Chemical Safety (Sonich-Mullin *et al.*, 2001) and later expanded substantially with the development of frameworks for evaluating the human relevance of mode of action (MoA) in experimental animals for carcinogens (Boobis *et al.*, 2006; Cohen *et al.*, 2003; Klaunig *et al.*, 2003; Meek *et al.*, 2003) and for noncancer effects (Boobis *et al.*, 2008; Seed *et al.*, 2005). Julien *et al.* (2009) carried this concept one step further and proposed the key events dose-response framework (KEDRF). KEDRF is a step-wise decision-logic process that provides a foundation for more rigorous and quantitative descriptions of dose-response.

The essential form of MoA analysis asks three questions to establish the likelihood of a chemical's potential effect on humans (Fig. 2):

- 1) Is there sufficient evidence in animal studies to establish a MoA?
- 2) If so, is that mode of action operative in humans? and
- 3) If so—considering pharmacokinetic and dynamic characteristics—would the MoA be operative in humans?

If the answer is YES to all three questions, then the effect seen in animals could plausibly occur in humans. Likewise, if the MoA is considered to be not relevant to humans, then the biological plausibility of the effect being observed in humans through the proposed MoA is highly unlikely.

THE EPID-TOX FRAMEWORK

In one of the initial essays that wrestled with the bases for inferences of the causes of disease, Hill (1965) concluded that it is not possible to “lay down some hard-and-fast rules of evidence that *must* be obeyed before we accept cause and effect.” In practice, tentative inference regarding the presence or absence of a causal relation between exposure and disease is made through a subjective process in which one considers which of the indicated features are present and, in particular, the degree to which they are present. Occasionally, the process is straightforward—all the evidence supports a causal hypothesis—and nearly everyone who addresses the issue arrives at the same conclusion, for example, that cigarette smoke is a cause of lung cancer. The evidence is considered “conclusive”; causation is viewed as “definitely present.”

A similar conclusion occasionally can be reached when the epidemiologic data are overwhelming, even without supporting evidence from other medical disciplines. For example, the

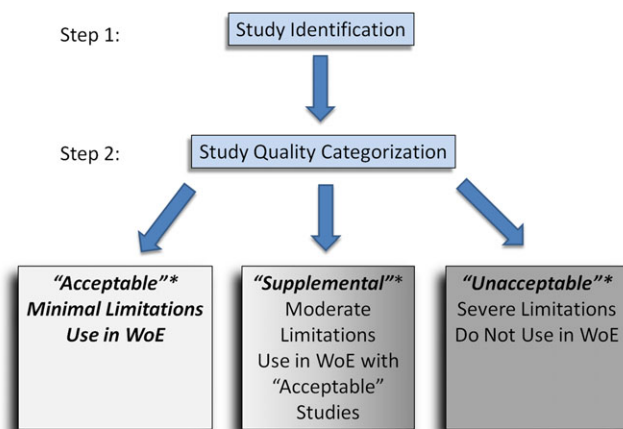


FIG. 2. Steps 1 and 2 of the Epid-Tox framework: study identification and quality categorization.

extremely strong association seen in epidemiologic studies between aspirin use and Reye's syndrome, combined with the absence of any similar association with the use of other analgesics (Halpin *et al.*, 1982; Hurwitz and Schonberger, 1987; Forsyth *et al.*, 1989), served as a solid basis for discouraging aspirin use in children, even without any precise knowledge at that time of how aspirin might have caused a child with flu or chicken pox to develop this illness.

In other instances, little or no evidence suggests causation, such as in the published literature relating exposure to magnetic fields and the occurrence of cancer. In this case, it is likely that most groups of experts would conclude that there is no evidence for an etiologic connection between exposure to magnetic fields and cancer in adults, that is, there is “no evidence supporting causality.” Because it is not possible to rule out a weak effect of exposure on disease incidence, it is not surprising that some debate continues regarding the safety of exposure to magnetic fields.

These instances of varied information from toxicology and epidemiology argue for a systematic approach that brings comprehensive, disciplined thinking into a complete and rational evaluation of the evidence. Such a systematic treatment lays on the table the complete story and gives practitioners a way to point to specific gaps in knowledge or lapses in logic based on the totality of information.

Overall, the Epid-Tox Framework follows a series of steps that assesses an explicit effect such as a specific cancer, neurological disease, or any tissue or system-specific adverse effect. The following steps would be to:

- 1) collect all relevant studies (toxicology and epidemiology),
- 2) assess the quality of each study and assign it to a quality category,
- 3) evaluate the epidemiological and toxicological weight of the evidence,
- 4) assign a scalable conclusion to the biological plausibility (toxicological) and epidemiological evidence, and
- 5) determine placement in a causal relationship grid.

Collect All Relevant Studies

This may be too obvious, but a serious source of bias is the selective collection of studies. A comprehensive search for all studies relevant to the end point in question should be conducted and documented as part of the process. This step is meant to be as inclusive as possible, bearing in mind that the process begins with a specific question: does agent X cause effect Y? All studies that offer data should be included at this point. One problem that continues to plague both toxicology and epidemiology is the nonpublication of “negative” studies wherein no effects were seen and investigators and journals are reluctant to publish such information. Nonetheless, no-effect studies are an important part of the total available data set and their absence biases the overall judgment in favor of studies showing effects.

Assess Quality and Categorize

Both kinds of studies, in epidemiology and toxicology, may present the observer with a wide range of investigations carried out in variable ways, with differing entry or exclusion criteria, variable ascertainment of effects, a range of exposures or exposure estimations, and observational endpoints. No study should be excluded at this stage of consideration. Having collected all available studies, each study should be included or excluded by using a transparent rationale. Both disciplines have generally accepted criteria for assessing study quality.

Toxicology The U.S. Environmental Protection Agency (USEPA) developed quality criteria that are typically applied in the evaluation of studies submitted for regulatory purposes. Various terms have been used to describe a study's suitability, relevance, conduct, and how well the study satisfies the intent of a particular guideline (USEPA, 1993), including "core guideline," "core minimum," "core supplementary," or "invalid." Core guideline indicates an acceptable study, whereas core minimum indicates that "while some things are missing, the study still fulfills the guideline requirements." Core supplementary has been used to identify studies with "... a significant deficiency or that additional information is required." Terms have changed over the years to "Acceptable" and "Unacceptable," with additional statements as to whether a study is upgradable to Acceptable status (USEPA, 2001). For the purpose of the Epid-Tox framework, the extremes of Acceptable and Unacceptable are useful categories. There are clearly going to be well done studies with verifiable conclusions and on the other hand studies with inapplicable methods, inappropriate data, or unsubstantiated conclusions. As well, an intermediate category is needed to categorize studies that have deficiencies that render them less than fully acceptable, but have sufficient quality that they cannot be regarded as unacceptable. Thus, the suggested categories include Acceptable, Supplemental, and Unacceptable (Fig. 3).

Epidemiology Similar to all scientific investigations, no epidemiological study is perfect; all have limitations to some extent. Nonetheless, experienced epidemiologists can evaluate the strengths and weaknesses of individual studies and categorize them as to whether they can be used to inform a judgment regarding causality. ECETOC (2009) has an excellent rendition of quality criteria that are based on elements of study design, exposure information, and health effects data. However, no objective, numerical yardstick exists to grade the quality of epidemiology studies.

Certainly, there is a subjective element in the categorization process. However, it is better to take a study's quality into account, acknowledging the imperfection of the process, than to give each study an equal weight. How this is done might vary from investigator to investigator, but in any case, the process of quality categorization needs to be transparently documented in the evaluation. What one investigator may find

to be an acceptable study might be rejected by another. The value of this step in the Epid-Tox Framework is to fully reveal and document not just the investigator's quality categorization but the reason for drawing a particular conclusion.

Documentation of these evaluation and categorization decisions can be provided in narrative form for individual studies. Study attributes to be considered include—but may not necessarily be limited to—the number of subjects, the range of exposure levels among these persons, study enrollment methodology, disease and exposure ascertainment methods, range of exposure, potential information bias, identification and measurement of potential confounders, and statistical methodology used to assess associations and to control for confounders.

As more reports of epidemiology studies include a complete description of the design and analytic methods used, the information needed to perform a quality assessment will be more readily available. A report checklist was developed by von Elm *et al.* (2007) for observational studies, known as the "Strobe statement," and serves as a method to evaluate the quality of reporting of a study. In a similar initiative, the London Principles for Epidemiology itemized the attributes that characterize well-conducted observational epidemiological studies (Graham, 1995; London Principles, 1996).

However, it might be measured, the quality of epidemiological studies will likely be distributed from useful to useless. For practical purposes, studies can be put into discrete categories of quality similar to those used for toxicology studies. Studies that are well designed, relatively free of bias, and have adequate control of known confounders are classified as Acceptable. Supplemental studies would have more serious imperfections and be of lesser quality but still be useable. Unacceptable studies would fail to meet several or all of the quality criteria and would not be used in subsequent steps in the evaluation (Fig. 3).

Evaluate the Weight of Evidence

Toxicology With all Acceptable and Supplemental studies at hand, the question is asked, "Is the effect of interest present." If there is evidence for a specific effect in some or all animal studies, the next stage in the evaluation is to use MoA analysis to determine human relevance (Fig. 2). If the answer to all three MoA Framework questions is "yes," then the effect is considered plausible to occur in humans. If the specific effect of interest is absent from the animal studies or the effect is present but judged by MoA analysis to be not relevant to human health, then the effect is concluded to have low biological plausibility to occur in humans.

Epidemiology Based on the evaluation of the complete set of studies categorized as Acceptable or Supplemental, a judgment is made as to whether or not there is an association between an agent and a given disease in humans as well as the strength of that association. This conclusion must be made

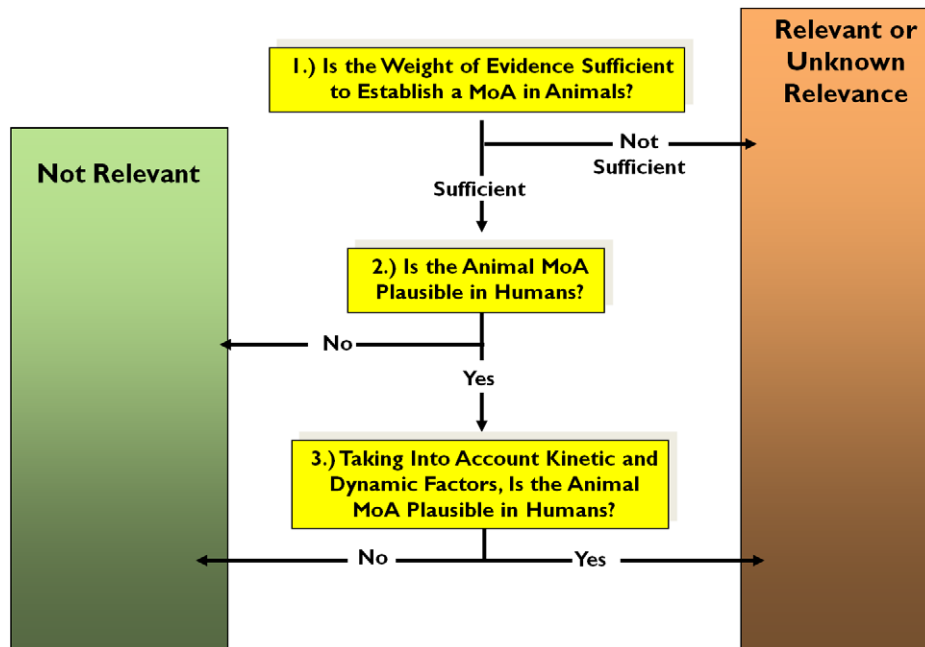


FIG. 3. The human relevance mode of action framework.

from the totality of evidence and may require balancing conflicting studies to produce one encompassing statement about the epidemiological evidence.

Various approaches can be used to produce one encompassing statement, including the systematic use of the Hill criteria (Hill, 1965). But the essence of evaluating the weight of evidence relies upon several central concepts. These include, but are not limited to, an effect within and among the studies that is found with strength, consistency, specificity, and coherence (Cole, 1997; Lagiou *et al.*, 2005). Whereas there are currently no hard-and-fast systematic, numerical characterizations that capture this expert judgment process, most practitioners would acknowledge that faced with an array of epidemiological studies, these concepts would guide their judgment in deriving a reliable encompassing statement of causal inference.

Assign a Scalable Conclusion

The ultimate value of the Epid-Tox Framework is to determine the degree of strength or likelihood of the effect of interest. Therefore, for both the epidemiological and toxicological findings, there needs to be a semiquantitative conclusion that states the degree to which the studies indicate a positive, a negative, or no relationship.

At the beginning of any epidemiological or toxicological evaluation, there has to be a starting point from which evidence pushes a conclusion toward the existence or lack thereof of causality. Starting at one end of a scale is not appropriate. Such a starting point implies that as studies are accumulated,

a positive association will be identified when the reverse, a lack of association, may also become increasingly plausible as scientific evidence accumulates.

Therefore, the scaling of strength for an epidemiological or toxicological evaluation begins at the center of the scale and, depending on the presence or absence of the effect, the scaling moves accordingly in the positive or negative direction. By starting at the center of each scale (the middle of the grid), *evidence of absence* can be distinguished from *absence of evidence* for an association. For example, with few sufficient quality epidemiology studies, one may have to state that there is an *absence of evidence* to conclude one way or another that there is a causal association. On the other hand, *evidence for an absence* of an epidemiological relationship can take either of two forms:

1) There may be a sufficient number of epidemiological studies to conclude that an association does not exist. For example, relative risks are around 1.0 and no statistical differences are seen within the studies. There is, therefore, evidence for an absence of an effect. As a consequence, the scaling shifts toward the left, indicating that there is epidemiological evidence “against” a causal link. With the accumulation of more and more studies not showing a given effect, the confidence for evidence against an association is strengthened.

2) Along with data sets showing no association, there may also be epidemiological data sets that actually indicate a protective effect. In this case, relative risks would be less than 1.0 and statistically significant.

Determine Placement in a Causal Relationship Grid

Figure 4 shows the Epid-Tox graphical template for establishing a causal relationship. Starting at the intersection of the x- and y-axes (middle of graph), the degree of biological plausibility (toxicology) is scaled on the y-axis and the degree (weight) of epidemiological evidence on the x-axis. The intersection of the toxicological and epidemiological scaling leads to an appropriate, evidence-based conclusion regarding causality.

The structure and appearance of the causal relationship graphic is fundamental to ensuing decisions about causality. Several factors led to the development of its form, including the impact of the degree of “positive” or “negative” data and the relative weighting of epidemiological studies versus toxicological studies. At the beginning of any analysis, there may be a dearth of either toxicological or epidemiological studies. In such a case, where the scaling remains at or near the center point, there is “insufficient information” to draw any conclusions. Note that the area of insufficient information is oblong and extended for the biological plausibility axis. The reason is that animal studies require a greater degree of evidence relative to epidemiology. Animal studies are surrogates for actual human data and as such require higher levels of evidence.

In addition, as more studies and information become available, the scaling of either the toxicological plausibility or epidemiological evidence can change in a way that can be easily illustrated with the two-dimensional graphic.

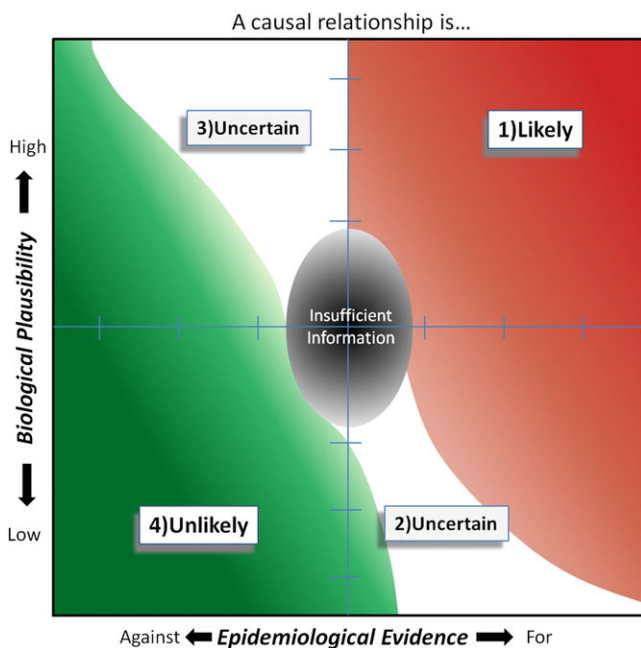


FIG. 4. The causal inference grid: how strong is the evidence for or against a causal relationship in humans?

At this point, the evaluator can clearly see where the epidemiological and toxicological evidence intersects and, based on that location on the graphic makes an overall conclusion that starts with, “A causal relationship is . . .” and completes the conclusion with words that describe the resultant area. Short, descriptive phrases are used here, but the underlying data and weight of evidence should be well understood at this point. The categories are as given below.

Likely A causal relationship is “Likely” between the environmental factor and the disease condition. This implies that consistent, reliable evidence from epidemiological and animal studies permits a causal inference to be made. Two examples of this outcome are asbestos as a cause of mesothelioma and tobacco smoke as a cause of lung cancer.

Uncertain A causal relationship is “Uncertain” between the environmental factor and the disease condition. In this case, there may be epidemiological evidence that can reasonably be interpreted as indicating a causal link. However, there may be little or no biological plausibility based on animal studies. Note in the lower right-hand corner that the transition between Likely and Uncertain favors epidemiological evidence. That is, with a high degree of epidemiological evidence, significant, and compelling data for a lack of biological plausibility must exist to transition from Likely to Uncertain. This again stresses the primacy of epidemiological evidence.

For example, Kaposi’s sarcoma, a normally rare tumor in humans, showed such a remarkable increased incidence following HIV infection (Sarid *et al.*, 2002) that epidemiological criteria for a causal relationship were met (Fig. 5). However, at the time, no laboratory studies had verified the pathogen. Therefore, the association was categorized as likely but of low biological plausibility. Later, extensive laboratory investigations led to the discovery of a specific herpes virus (HHV8 or KSHV) that strengthened the inference of causality because of an increased knowledge regarding a likely pathogenesis of Kaposi’s sarcoma.

Based on some early suggestive results, a number of epidemiologic studies have been done on the possible relation between exposure to electromagnetic fields (EMF) and the occurrence of brain cancer. The results from occupational studies—which typically involve higher levels of exposure than residential studies—have been summarized (Kheifets *et al.*, 2008). The relative risk associated with EMF exposure was statistically significant (RR = 1.14, 95% CI = 1.07–1.22). The authors of the review, however, concluded that “the lack of a clear pattern of EMF exposure and outcome risk does not support a hypothesis that these exposures were responsible for the excess risk.” Biological plausibility is low in this example because “*in vitro*, *in vivo*, or mechanistic evidence has not provided clues” as to a basis for an association between exposure to EMF and the development of brain cancer (Kheifets *et al.*, 2009). As shown in Figure 5, the initial analysis of epidemiological studies showed some evidence;

however, in combination with low plausibility, the causal relationship would be considered Uncertain. With time, more recent studies—generally with relatively better exposure ascertainment—tended to observe an even smaller association than did earlier studies. The updated evaluation would move the categorization from Uncertain to Unlikely.

Uncertain A causal relationship is “Uncertain” but plausible between the environmental factor and the disease condition. In this instance, the weight of evidence analysis of epidemiological studies shows little or no evidence of any effect although toxicological studies may indicate the plausibility of an effect in humans.

For example, as shown in Figure 6, melamine bladder and kidney toxicity seen in animal studies was considered relevant to human health, albeit only at very high exposures. But no epidemiological evidence supported a causal relationship. An initial evaluation placed melamine in this category (Uncertain but plausible). However, the unfortunate incidents in China after the adulteration of milk with melamine and resultant rise in the number of children with melamine crystals detected in the urinary bladder and death because of kidney damage confirmed that the mode of action understood from animal models is relevant to humans at high levels of exposure (World Health Organization, 2009). This additional epidemiological evidence moved the conclusion of causality from Uncertain to Likely.

Unlikely A causal relationship is “Unlikely” between the environmental factor and the disease condition. Both epidemiological and toxicological evidence is compatible with the absence of effect.

For example, because D-limonene causes kidney toxicity in male rats, the biological plausibility was high and, without

epidemiological evidence, would be considered Uncertain but plausible (Fig. 6). Subsequent investigations showed that D-limonene-induced kidney toxicity is not relevant to humans (Swenberg and Lehman-McKeeman, 1999; Meek *et al.*, 2003), which moves the conclusion of a causal relationship to the Unlikely category.

Another example, not shown on the grid, is phenobarbital. Phenobarbital increased the incidence of liver tumors in long-term rodent bioassays (Whysner *et al.*, 1996) by a mode of action that would be plausible in humans. Biological plausibility would be considered high for phenobarbital. However, epidemiological studies have found no evidence of liver tumors in patients on lifetime anti-epilepsy treatment with phenobarbital (IARC, 2001; Whysner *et al.*, 1996). Without epidemiological evidence, the categorization would be Likely or Uncertain, but with sufficient epidemiological evidence for an absence of an effect in humans, the categorization would be Unlikely.

As mentioned previously, a checkbox approach to characterize the nature of the evidence that would lead an expert team of epidemiologists and toxicologists to reach a weight of evidence decision is not practical. However, the Epid-Tox Framework described here and shown schematically in Figure 7 provides by structure and example a way of systematically working through all evidence and reaching a conclusion that can be tracked, debated, and modified with further data.

FUTURE NEEDS AND DIRECTIONS

The proposed new framework represents a concerted effort to bridge the fields of epidemiology and toxicology in a way that can impact, and hopefully improve, human risk

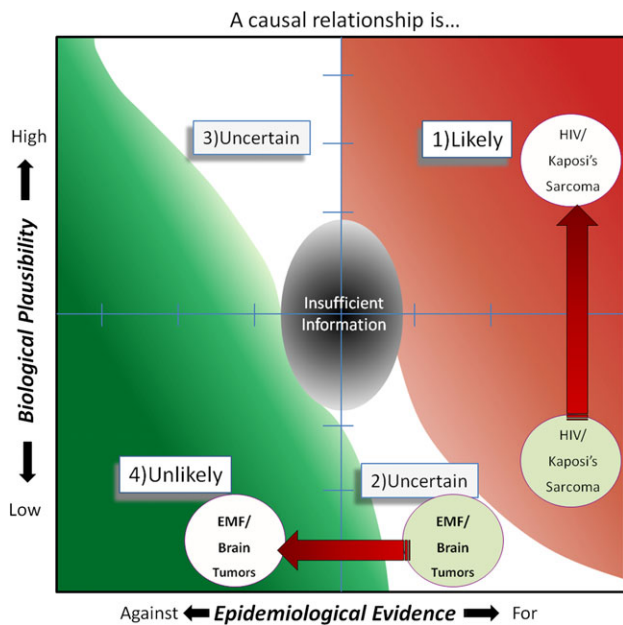


FIG. 5. Applications of the Epid-Tox framework: HIV/Kaposi's sarcoma and EMF and brain tumors.

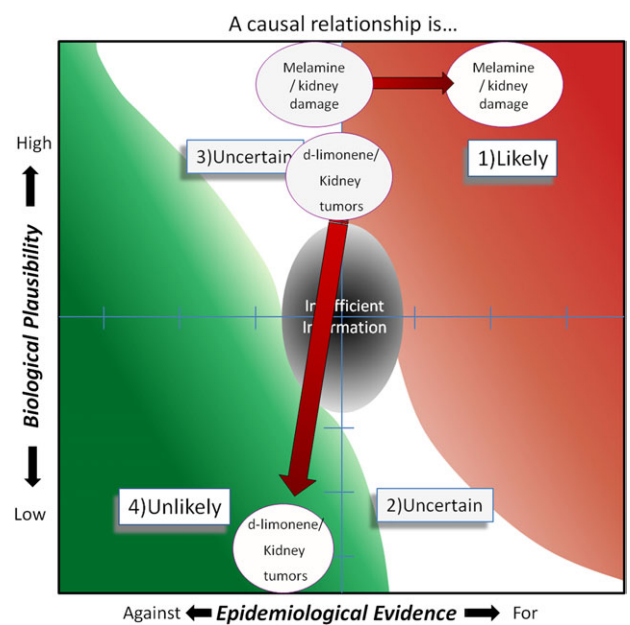


FIG. 6. Applications of the Epid-Tox framework: melamine and d-limonene.

assessment. It will benefit from application and critique and will undoubtedly require some modification. This formalized set of steps, in and of itself, provides a structure for challenging both disciplines and how they can and should be brought together. Each step of the process invites improvement, including the availability of studies, the determination of quality, the proper metric for assigning the degree or strength of evidence, and the appearance and utility of the two-dimensional grid.

Availability of evidence will always be an Achilles Heel in any evaluation that seeks scope and completeness. Unpublished data that languish in the drawer or is only contained in official submissions to regulatory agencies is an unfortunate omission that currently is unavoidable. Furthermore, for both disciplines, the lack of “negative” studies (showing no effects) are usually judged to be of lesser value either by the investigator seeking a new finding or a journal editor requiring impactful research. Such evidence rarely appears in the literature.

One area that continues to plague both toxicology and epidemiology is measurement of quality. Whereas poor and excellent studies can often be identified and categorized, there is no consistent and agreed method for those that fall in between poor and excellent. For example, the Strobe statement (von Elm *et al.*, 2007) details criteria for judging the quality and reliability of epidemiology studies, but the method is not routinely used or cited with studies or reviews. In experimental biology, the criteria

for quality are less well defined and are often the product of where the work was done and where it was published.

Another area that will need debate and refinement is the degree of detail one needs to complete the two-dimensional grid proposed in the Epid-Tox Framework. It may be sufficient to declare general degrees of confidence in the scaling of the two axes. However, with more precise scaling, one could imagine dividing the grid into four quadrants, with four quadrants within each quadrant, thereby subdividing and giving greater granularity to the overall conclusions. This may add greater detail to the analysis but may also spark fruitless debates about precisely where the biological plausibility or epidemiological point should be on each axis.

Nonetheless, a framework can provide the logic and disciplined thinking that promotes open discourse and leads to evidenced-based decisions. Furthermore, decisions about what epidemiological or toxicological study should be done can be facilitated by using the framework. For example, clear indications from animal studies for a particular effect should inform the data collected in an epidemiology study. Likewise, epidemiological findings should spur the design of *in silico*, *in vitro*, or *in vivo* studies that could corroborate observations in human populations. Important decisions about human safety should rely on the cohesive appreciation of both epidemiology and toxicology and the synergistic

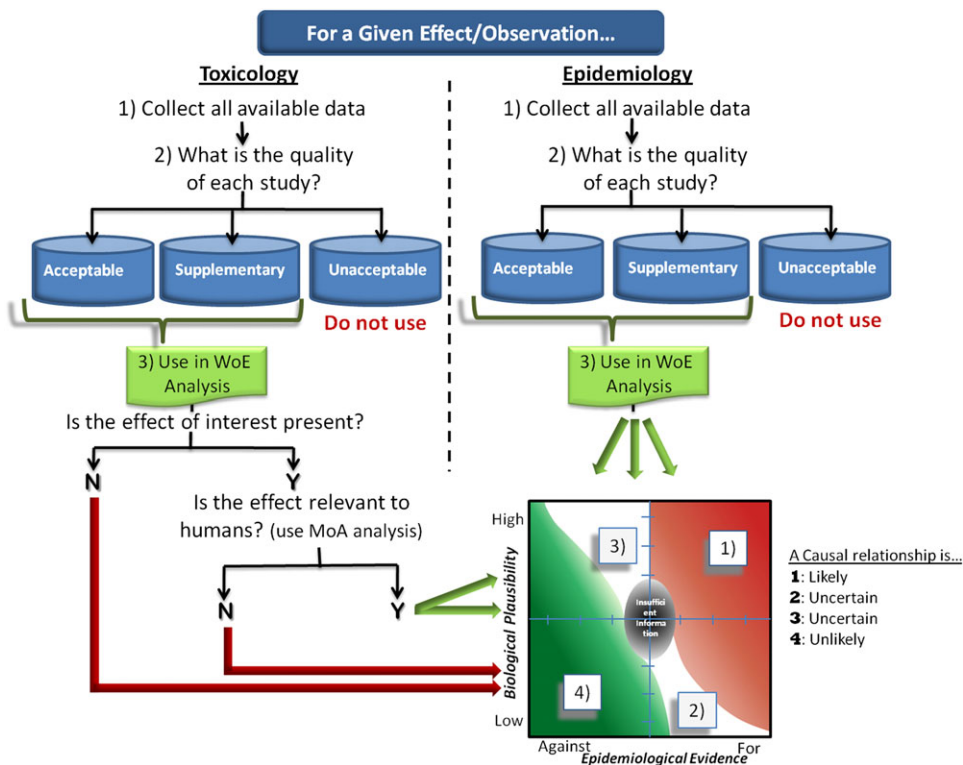


FIG. 7. Schematic representation of the framework for causal inference based upon weight of evidence of animal and epidemiological data.

value that their combination brings to a comprehensive evaluation.

The refinement of any method occurs by working examples through it. In order to take that first step toward refinement, Simpkins *et al.* (2011) provides a case study that utilizes the framework to collect, evaluate, and integrate epidemiological and toxicological evidence for causal inference. Hopefully, more environmental agents will be worked through the Epid-Tox Framework to test and improve its utility.

REFERENCES

- Bayne-Jones, S., Burdette, W., Cochran, W., Farber, E., Fieser, L., Furth, J., Hickam, J., LeMaistre, C., Schuman, L., and SeEVERS, M. (1964). In *Smoking and Health: Report of the Advisory Committee to the Surgeon General of the Public Health Service*. Department of Health, Education, and Welfare Public Health Service. Public Health Service Publication No. 1103 Superintendent of Documents. U.S. Government Printing Office, Washington, DC.
- Boobis, A. R., Cohen, S. M., Dellarco, V., McGregor, D., Meek, M. E., Vickers, C., Willcocks, D. and Farland, W. (2006). IPCS framework for analyzing the relevance of a cancer mode of action for humans. *Crit. Rev. Toxicol.* **36**, 781–792.
- 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.
- Cohen, S. M., Meek, M. E., Klaunig, J. E., Patton, D. E., and Fenner-Crisp, P. A. (2003). The human relevance of information on carcinogenic modes of action: overview. *Crit. Rev. Toxicol.* **33**, 581–589.
- Colditz, G. A., Baer, H. J., and Tamimi, R. M. (2006). Breast cancer. In *Cancer Epidemiology and Prevention* (D. Schottenfeld, Fraumeni J. F. Jr.), 3rd ed. Oxford University Press, New York.
- Cole, P. (1997). Causality in epidemiology, health policy and law. *Environ. Law Rep.* **27**, 10279–10285.
- ECETOC. (2009). *Framework for the Integration of Human and Animal Data in Chemical Risk Assessment*. Technical Report No. 104 ISSN-0773-8072-104. European Centre for Ecotoxicology and Toxicology of Chemicals, Brussels, Belgium.
- Forsyth, B. W., Horwitz, R. I., Acampora, D., Shapiro, E. D., Viscoli, C. M., Feinstein, A. R., Henner, R., Holabird, N. B., Jones, B. A., Karabelas, A. D. E., *et al.* (1989). New epidemiologic evidence confirming that bias does not explain the aspirin/Reye's syndrome association. *J. Am. Med. Assoc.* **261**, 2517–24.
- Graham, J. D. (1995). The role of epidemiology in regulatory risk assessment. In *Proceedings of the Conference on the Proper Role of Epidemiology in Risk Analysis, 13–14 October 1994, Boston, MA*. Elsevier Science Ltd, New York, NY.
- Guzelian, P. S., Victoroff, M. S., Halmes, N. C., James, R. C., and Guzelian, C. P. (2005). Evidence-based toxicology: a comprehensive framework for causation. *Hum. Exp. Toxicol.* **24**, 161–201.
- Halpin, T. J., Holtzauer, F. J., Campbell, R. J., Hall, L. J., Correa-Villasenor, A., Lanese, R., Rice, J., and Hurwitz, E. S. (1982). Reye's syndrome and medication use. *J. Am. Med. Assoc.* **248**, 687–691.
- Hill, A. B. (1965). The environment and disease: association or causation. *Proc. R. Soc. Med.* **58**, 295–300.
- Hunter, D. J., Spiegelman, D., Adami, H.-O., Beeson, L., van den Brandt, P. A., Folsom, A. R., Graser, G. E., Goldbohm, R. A., Graham, S., Howe, G. R., *et al.* (1996). Cohort studies of fat intake and the risk of breast cancer—a pooled analysis. *N. Engl. J. Med.* **334**, 356–361.
- Hurwitz, E. S., and Schonberger, L. B. (1987). Public Health Services study of Reye's syndrome and medication. Report of the main study. *J. Am. Med. Assoc.* **257**, 1905–1911.
- IARC. (2001). *Phenobarbital and Its Sodium Salts. IARC Monographs on the Evaluation of Carcinogenic Risks to Humans*. Vol. 79, 161–288. IARC Press, Lyon, France.
- Julien, E., Boobis, A., and Olin, S. (2009). The key events dose-response framework: a cross-disciplinary mode-of-action based approach to examining dose-response and thresholds. *Crit. Rev. Food Sci. Nutr.* **49**, 682–689.
- Kheifets, L., Bowman, J. D., Checkoway, H., Feychting, M., Harrington, J. M., Kavet, R., Marsh, G., Mezei, G., Renew, D. C., and van Wijngaarden, E. (2009). Future needs of occupational epidemiology of extremely low frequency electric and magnetic fields: review and recommendations. *Occup. Environ. Med.* **66**, 72–80.
- Kheifets, L., Monroe, J., Vergara, X., Mezei, G., and Abdelmonem, A. (2008). Occupational electromagnetic fields and leukemia and brain cancer: an update to two meta-analyses. *J. Occup Environ Med.* **50**, 677–688.
- Klaunig, J. E., Babich, M. A., Baetcke, K. P., Cook, J. C., Corton, J. C., David, R. M., DeLuca, J. G., Lai, D. Y., McKee, R. H., Peters, J. M., *et al.* (2003). PPAR alpha agonist-induced rodent tumors: modes of action and human relevance. *Crit. Rev. Toxicol.* **33**, 655–780.
- Koch, R. (1884). Die Aetiologie der Tuberkulose. *Mitt Kaiser Gesundh* 2:1–88.
- Koch, R. (1893). Über den augenblicklichen Stand der bakteriologischen Choleradiagnose (in German). *Zeitschrift für Hygiene und Infektionskrankheiten* **14**, 319–333.
- Kundi, J. M. (2007). Causality and the interpretation of epidemiological evidence. *Environ. Health Perspect.* **114**, 969–974.
- Lagiou, P., Adami, H., and Trichopoulos, D. (2005). Causality in cancer epidemiology. *Eur. J. Epidemiol.* **20**, 565–574.
- London Principles. (1996). *The London Principles for Evaluating Epidemiologic Data in Regulatory Risk Assessment*. Available at: <http://www.fedfocus.org/science/london.html>.
- Meek, M. E., Bucher, J. R., Cohen, S. M., Dellarco, V., Hill, R. N., Lehman-McKeeman, L. D., Longfellow, D. G., Pastoor, T., Seed, J., and Patton, D. E. (2003). A framework for human relevance analysis of information on carcinogenic modes of action. *Crit. Rev. Toxicol.* **33**, 591–653.
- Pastoor, T., and Stevens, J. (2005). Historical perspective of the cancer bioassay. *Scand. J. Work Environ. Health* **31**(Suppl. 1), 129–140.
- Phillips, C. V., and Goodman, K. J. (2004). The missed lessons of Sir Austin Bradford Hill. *Epidemiol. Perspect. Innov.* **1**, 3.
- Popper, K. (1959). In *The Logic of Scientific Discovery*. First printed in English by Hutchinson & Co.; Republished by Ruteledge, 2006, New York, NY.
- Potter, M. (1962). Percivall Pott's contribution to cancer research. *NCI Monograph.* **10**, 1–13.
- Qian, G. S., Ross, R. K., Yu, M. C., Yuan, J. M., Gao, Y. T., Henderson, B. E., Wogan, G. N., and Groopman, J. D. (1994). A follow-up study of urinary markers of aflatoxin exposure and liver cancer risk in Shanghai, People's Republic of China. *Cancer Epidemiol. Biomarkers Prev.* **3**, 3–10.
- Rothman, K. J. (1976). Causes. *Am. J. Epidemiol.* **104**, 578–592.
- Rothman, K. J., and Greenland, S. (2005). Causation and causal inference in epidemiology. *Am. J. Public Health* **95**, S144–S150.
- Sarid, R., Klepfish, A., and Schattner, A. (2002). Virology, pathogenic mechanisms and associated diseases of Kaposi sarcoma-associated herpesvirus (Human herpesvirus 8). *Mayo Clin. Proc.* **77**, 941–949.
- Seed, J., Carney, E., Corley, R., Crofton, K., DeSesso, J., Foster, P., Kavlock, R., Kimmel, G., Klaunig, J., Meek, E., *et al.* (2005). Overview:

- using mode of action and life stage information to evaluate the human relevance of animal toxicity data. *Crit. Rev. Toxicol.* **35**, 663–672.
- Simpkins, J. W., Swenberg, J. S., Weiss, N. S., Brusick, D., Eldridge, J. C., Stevens, J. T., Handa, R. G., Hovey, R. C., Plant, T. M., Pastoor, *et al.* (2011). Atrazine and breast cancer: a framework assessment of the toxicological and epidemiological evidence. *Tox. Sci.* Advance Access published on July 18, 2011; doi: doi:10.1093/toxsci/kfr176.
- Sonich-Mullin, C., Fielder, R., Wiltse, J., Baetcke, K., Dempsey, J., Fenner-Crisp, P., Grant, D., Hartley, M., Knaap, A., Kroese, D., *et al.* (2001). IPCS conceptual framework for evaluating a mode of action for chemical carcinogenesis. *Regul. Toxicol. Pharmacol.* **34**, 146–152.
- Susser, M. (1986). The logic of Sir Karl Popper and the practice of epidemiology. *Am. J. Epidemiol.* **124**, 711–718.
- Swenberg, J. A., and Lehman-McKeeman, L. D. (1999). α 2 Urinary-globulin-associated nephropathy as a mechanism of renal tubule cell carcinogenesis in male rats. In *Species Differences in Thyroid Kidney and Urinary Bladder Carcinogenesis*. IARC Scientific Publications No. 147 (C. Capen, E. Dybing, J. Rice, and J. Wilbourn, Eds.). Lyon, France, pp. 95–118. WHO Press, Geneva, Switzerland.
- Tabuenca, J. M. (1981). Toxic-allergic syndrome caused by ingestion of rapeseed oil denatured with aniline. *Lancet* **2**, 567–568.
- Tomatis, L., Aitio, A., Wilbourn, J., and Shuker, L. (1989). Human carcinogens so far identified. *Jpn J Cancer Res* **80**, 795–807.
- U.S. Environmental Protection Agency (USEPA) (1993). *Pesticide Reregistration Rejection Rate Analysis—Toxicology* National Service Center for Environmental Publications 738R93004, p. 22.
- U.S. Environmental Protection Agency (USEPA) (2001). *HED Standard Operating Procedure: Executive Summaries for Toxicology Data Evaluation Records (DERs)* SOP 2001.02, p. 7.
- von Elm, E., Altman, D. G., Egger, M., Pocock, S. J., Gotsche, P. C., and Vandenbroucke, J. P. (2007). The strengthening of the reporting of observational studies in epidemiology (STROBE) statement: guidelines for reporting observational studies. *PLoS Med* **4**, 1623–1627.
- Wang, L. Y., Hatch, M., Chen, C. J., Levin, B., You, S. L., Lu, S. N., Wu, M.-H., Wu, W.-P., Wang, L.-W., Wang, Q., *et al.* (1996). Aflatoxin exposure and risk of hepatocellular carcinoma in Taiwan. *Int. J. Cancer* **67**, 620–625.
- Weed, D. L. (2005). Weight of evidence: a review of concepts and methods. *Risk Anal.* **25**, 1545–1557.
- Whysner, J., Ross, P. M., and Williams, G. M. (1996). Phenobarbital mechanistic data and risk assessment: enzyme induction, enhanced cell proliferation and tumor promotion. *Pharmacol. Ther.* **71**, 153–191.
- Wilbourn, J., Haroun, L., Vainio, H., and Montesano, R. (1984). Identification of chemicals carcinogenic to man. *Toxicol. Pathol.* **12**, 397–399.
- World Health Organization. (2009). *Report of a WHO expert meeting in collaboration with FAO supported by Health Canada. Toxicological and Health Aspects of Melamine and Cyanuric Acid*. WHO Press, Geneva, Switzerland.
- Yamagiwa, K., and Ichikawa, K. (1918). Experimental study of the pathogenesis of carcinoma. *J. Cancer Res.* **3**, 1–29.