EXHIBIT 53

Center for Chemical Regulation and Food Safety

Exponent®

Design of Epidemiologic Studies for Human Health Risk Assessment of Pesticide Exposures

Exponent®

Design of Epidemiologic Studies for Human Health Risk Assessment of Pesticide Exposures

Prepared for

CropLife America

Prepared by

Exponent
1150 Connecticut Avenue, NW
Suite 1100
Washington, DC 20036

January 4, 2016

1507316.000 - 5694

Contents

Page

List of Figures

Introduction

Epidemiology is the study of the distribution and determinants of health and disease in populations. In human health risk assessment, which aims to estimate the nature and probability of adverse health effects in humans of adverse health effects in humans who may be exposed to environmental hazards (U.S. EPA, 2015b), information is often based on extrapolation of laboratory animal toxicology studies. However, high-quality epidemiologic studies can also provide valuable information about the quantitative exposure-response relationship in humans who have experienced actual, directly relevant exposures (Burns et al., 2014; Calderon, 2000; Hertz-Picciotto, 1995; Lavelle et al., 2012; Vlaanderen et al., 2008). As opposed to laboratory animal data, epidemiologic data are not subject to major uncertainties related to species extrapolation to humans. Epidemiologic studies can also encompass heterogeneous populations and real-world variability in the duration, intensity, route, and level of exposure, as well as mixtures of exposures. By contrast, laboratory studies are typically restricted to genetically inbred strains and controlled, high-dose exposures that may not reflect realistic conditions. However, epidemiologic studies are usually poorly suited for detecting small increases in risk, and study design limitations can permit bias and confounding that undermine the validity of results for causal inference. In particular, poor exposure assessment is largely responsible for the limited use of epidemiologic data in human health regulatory risk assessments.

In 2010, the Office of Pesticide Programs (OPP) at the U.S. Environmental Protection Agency (EPA) released a draft proposed framework for incorporating human epidemiologic and incident data (i.e., case reports and short-term poisoning incidents surveillance studies) into pesticide risk assessments (U.S. EPA, 2010). The proposed approach was designed to impart scientific rigor, consistency, and transparency to the Agency's evaluation of epidemiologic data in pesticide risk assessments, thereby taking advantage of the increased availability of large, prospective epidemiologic studies. The integration of epidemiologic studies into risk assessment of pesticides was also intended to be conceptually consistent with the National Research Council (NRC)'s 2007 vision and strategy on 21st-century toxicity testing, with an emphasis on using systems biology, bioinformatics, and high-throughput technologies to better understand adverse outcome pathways (AOPs) (NRC, 2007). In particular, the draft framework used problem formulation, as routinely used in ecological risk assessments, as a method to define exposure pathways and potential health outcomes of interest, along with appropriate scientific methods for characterizing risk in the context of addressing risk management questions and risk mitigation options. The draft framework also used the concepts of mode of action (MOA) and AOP to guide the integration of information from different lines of scientific evidence and across different levels of biological organization, from the initiating molecular event to tissue- and organ-level responses, extending out to whole-organism and population levels of effect. This approach was intended to build a biological systems-level approach to increase scientific confidence in risk management decision making based on potential causal relationships between chemical exposures and disease outcomes.

A Federal Insecticide, Fungicide, and Rodenticide Act (FIFRA) Scientific Advisory Panel subsequently met to review the draft framework and make recommendations for revision (FIFRA Scientific Advisory Panel, 2010). One of the panel's recommendations was to devote particular attention to the quality of epidemiologic studies, including consideration of the validity of the exposure assessment, sample size and statistical power, the definition and assessment of the outcome, possible sources of bias, consideration of and control for confounding and effect modification, and external validity or generalizability. The FIFRA Scientific Panel also recommended consideration of prospective cohort studies, historical cohort studies, case-control studies, cross-sectional studies, and hybrid designs in the weight of evidence regarding an exposure-outcome association, although it recommended separate treatment of ecologic studies due to their inherent limitations for risk estimation at the individual level.

Many epidemiologic studies are not useful for risk assessment, often due to the lack of valid, specific, quantitative measures of exposure, especially etiologically relevant exposure in the past. Other common limitations are poor control for confounding and other biases, constrained evaluation of effect modification based on small subgroups, and limited generalizability to the entire population. Due to these shortcomings, some scientists believe that the role of epidemiology in risk assessment should be limited mainly or exclusively to "hazard identification," i.e., an early phase of risk assessment in which the overall weight of relevant scientific evidence is identified and reviewed to determine the types of health outcomes that can be caused by an exposure (NRC, 1983) and further explored in mechanistic research. However, others view epidemiology as potentially contributing to later phases of risk assessment, including "dose-response assessment" (i.e., quantitative estimation of the incidence of a health outcome as a quantitative function of the amount of exposure) and "risk characterization" (i.e., integration of the exposure assessment and dose-response assessment components of a risk assessment to synthesize an overall conclusion about risk that is complete, informative, and useful for decision makers) (U.S. EPA, 2000).

The potential role of epidemiology in informing all phases of risk assessment can be substantially enhanced through more rigorous study design and conduct. For example, epidemiologic studies with quantitative, specific, and accurate measurements of internal biological dose and associated health outcomes can be used for dose-response assessment, while those that demonstrate changes in population health due to modification of a causal exposure can provide some proof-in-principle of observed risk and offer input for the regulatory impact assessment of the potential benefits of increased regulatory actions to mitigate risk. Therefore, to provide guidance for the design and interpretation of future epidemiologic studies, this document draws from previous discussions of the strengths and limitations of epidemiologic data for human health risk assessment (Burns et al., 2014; Calderon, 2000; Hertz-Picciotto, 1995; Lavelle et al., 2012; Vlaanderen et al., 2008), the importance of integrating observational human and experimental animal data for this purpose (FIFRA Scientific Advisory Panel, 2010; U.S. EPA, 2010), and specific examples of the use of human epidemiologic data for pesticide risk assessment to organophosphate pesticides (U.S. EPA, 2015a; U.S. EPA, 2011) to provide recommendations on key attributes for any epidemiologic study to increase its likelihood of providing informative results for human health risk assessment.

Study Design

Standard epidemiologic study designs include ecologic, cross-sectional, case-control, retrospective/historical cohort, and prospective cohort studies, as well as variants and hybrids of these designs (e.g., case-control, case-crossover, and case-only studies). Ecologic studies estimate exposure crudely at the group level, and group-level associations cannot validly be assumed to hold at the individual level. Therefore, although ecologic studies can be helpful to identify potential hazards and formulate causal hypotheses (i.e., "problem formulation"), they typically are not useful for quantitative human health risk assessments.

Cross-sectional, case-control, and cohort studies benefit from the collection of detailed individual-level data on exposures, outcomes, and potential confounders or effect modifiers. Cross-sectional studies assess exposure and disease status simultaneously, often making it impossible to demonstrate temporal concordance, show whether the exposure preceded the outcome chronologically. Cross-sectional studies are thus susceptible to information bias (if exposure ascertainment or reporting differs systematically between cases and non-cases/controls) and reverse causality (if the disease condition itself affects the measured exposure). Selection bias can also threaten the validity of these studies if study participation or completeness of data collection varies by exposure and disease status. Other key limitations of cross-sectional studies are the inability to distinguish between incident (newly developed) and prevalent (pre-existing) health conditions, and the possible enrichment of prevalent cases with a relatively longer duration or better prognosis than is typical.

Case-control studies are efficient for the investigation of rare outcomes and those with a long putative latency period between exposure and disease onset. Well-conducted case-control studies can be conceptualized as more efficient versions of corresponding cohort studies in which the cases are the same as those who would have been included in the cohort study, and the controls are a sample of the remaining cohort (source population). Thus, rigorous case-control studies yield relative risks that are valid estimates of the rate ratios that would be obtained from cohort studies. However, retrospective case-control studies collect exposure information after disease onset, making it difficult under some circumstances to establish the temporal sequence between the exposure and the outcome, and raising the possibility of information bias and reverse causation. Selection bias can occur due to differentially incomplete study participation or data collection, or if inappropriate control identification and selection lead to an exposure distribution that is not representative of that in the study base that gave rise to the cases. Case-control studies are generally inefficient for the investigation of rare exposures, and they usually focus on one or a limited number of disease conditions at a time, except when nested within a cohort study.

Retrospective or historical cohort studies are efficient for the investigation of uncommon exposures and can enable the evaluation of associations with a large number of health outcomes, although cohort studies are often underpowered for rare outcomes. Selection bias due to differential enrollment by disease status is generally unlikely because the study population is usually defined independently of the outcome—often based on place of employment or residence— although it can still occur. Selection bias can also arise if follow-up varies by exposure and disease status. Because exposures are ascertained after at least some outcomes have already occurred, information bias is a possibility, although it can be minimized by the use of existing exposure information that was recorded independently of disease status. The reliance on previously collected exposure records, however, often limits detailed assessment of exposures, confounders, and effect modifiers, since existing data typically have not been collected for research purposes. Identification of an appropriate comparison (unexposed) group can also be challenging in retrospective cohort studies.

Thus, for the purposes of informing all three phases of human health risk assessment, the prospective cohort study design is generally the most likely to yield useful results. Advantages include the ability to examine multiple health outcomes, the opportunity to collect detailed information on exposures and other covariates, the possibility of establishing the temporal sequence between exposure and outcome, and the low probabilities of information bias due to differential exposure misclassification and selection bias due to differential participation, since disease status is unknown at the time of study initiation and data collection. Selection bias due to differential follow-up, however, can still occur, particularly if study attrition is substantial. By enrolling individuals from a broad age range, prospective cohort studies enable investigation of the potential health effects of an exposure over the life course. With repeated exposure measurement, these studies can potentially capture whether risk of the outcome varies by pattern of exposure over time, and they can also evaluate whether the exposure-outcome relationship varies with temporal aspects of exposure, such as duration and age at initiation, cessation, or peak exposure.

In summary, well-conducted cross-sectional, case-control, and cohort studies can all yield valid and informative results for risk assessment. However, the generally low probability of biased exposure misclassification and the possibility of assessing repeated exposures and multiple health outcomes across the life span are major advantages of prospective cohort studies, if properly conducted.

Exposure Assessment

An essential component of any human health risk assessment is characterizing the exposure-response relationship, that is, evaluating the likelihood and severity of specific health outcomes at different levels and conditions of exposure (U.S. EPA, 2015b). In risk assessment, the critical effect is the adverse health outcome that occurs at the lowest level of exposure among all available studies, based on the assumption that prevention of that critical effect would also prevent other adverse effects. Ideally, a study population should experience a sufficiently wide range of exposure, including low and high levels, such that extrapolation outside of the observed exposure range is not necessary to estimate health risks at low exposures.

Quantitative assessment of valid, reliable, and etiologically relevant exposures is arguably the most formidable challenge in conducting epidemiologic studies to inform risk assessment. In this context a distinction should be made between "exposure," which is used here to refer to the concentration of an agent in the external environment, and "dose," which here refers to the biologically absorbed internal concentration of the agent. Because some exposures may not result in any appreciable dose (e.g., as in the case of substances with low vapor pressure and low skin penetrability, resulting in low absorbed dose from inhalation and dermal exposures, respectively) (Acquavella et al., 2004), and because the same exposure can result in substantially different doses (e.g., due to inter-individual differences in metabolism) (Garfitt et al., 2002), internally measured dose is more likely to be etiologically relevant than externally measured exposure.

Epidemiology is fundamentally an observational science, meaning that naturally occurring situations are observed by investigators to understand the patterns and causes. Because exposures are not assigned in a fixed manner—and, indeed, it would be considered unethical to expose study subjects to agents at concentrations that are known or suspected to cause adverse health effects—current and especially past exposures can be challenging to measure accurately and reliably.

If available, biomarkers of exposure, such as concentrations of a pollutant or its metabolites in tissues or bodily fluids and exposure-induced molecular changes, generally provide a better measure of internal dose than other exposure metrics, such as self-reported exposure or environmental sampling data (Schmidt, 2006). Biomarkers must be validated in large, representative populations to ensure their utility and relevance. First and arguably foremost for regulatory decision making, they must be specific to the exposure of interest, i.e., able to reflect the dose of a particular agent or group of agents, but not others. Without specificity to the chemical of interest, a valid association cannot be identified. Second, biomarkers must be shown to be accurate, i.e., able to measure what they aim to measure. Third, they must be reliable, i.e., able to yield the same results upon repeated testing. Finally, the analytical method used to detect the biomarker must be sensitive, i.e., able to detect low doses of the agent of interest. Most biomarkers cannot distinguish among sources or routes of exposure, which can sometimes be differentiated on the basis of questionnaires and environmental measures.

With respect to pesticide exposure, the advantages of assessment using biomarkers are especially evident when contrasted with the limitations of self-reported exposure data. Directly analogous biomarkers may exist for humans and animals, thereby facilitating the integration of human epidemiology and animal toxicology results, whereas human self-reported exposure data have no animal equivalent. In addition, self-reported information is prone to various types of error, especially inaccuracy in the recollection of specific pesticides and amounts used (Blair and Zahm, 1990). Any given day of pesticide use can entail highly variable amounts of pesticides used and numbers of mixing operations, and urinary biomarkers of pesticide exposure have been shown to be poorly correlated with estimated exposure intensity scores based on farmers' self-reported data (Acquavella et al., 2006).

In the Agricultural Health Study cohort, for example, the reliability (reproducibility) of self-reported information on ever having mixed or applied specific pesticides was evaluated by comparing responses to two questionnaires completed one year apart by nearly 4,000 pesticide applicators (Blair et al., 2002). Agreement on ever/never use ranged from approximately 70% to 90%, and the kappa statistic value for inter-rater agreement ranged between approximately 50% and 70%. However, for more detailed questions about duration, frequency, and decade of first use of specific pesticides, agreement was lower (mostly 50–60%, with kappa 30–80%). In other validation studies that compared urinary pesticide biomarker levels with estimated exposure intensity based on an expert-derived algorithm using self-reported or directly observed exposure, data, Spearman correlation coefficients ranged between 0.4 and 0.8, depending on the type of pesticide (Blair et al., 2011; Coble et al., 2011). Correlations were poorer (between -0.4 and 0.2) for self-reported determinants of pesticide exposure such as kilograms of active ingredient, hours spent mixing and applying, and number of acres treated. These results underscore the limitations of self-reported pesticide exposure data and highlight the importance of using validated biomarker levels when possible.

A key limitation of reliance on biomarkers of exposure is that such measures usually reflect only current or recent past exposures. In some circumstances, such as studies of *in utero* exposures, measures of relatively short-term biomarker levels may not be problematic. To identify causal relationships between exposures and health outcomes with long presumed or known latency periods, biological markers of distant past exposure would be ideal, but such measures may be difficult or impossible to obtain (Chang et al., 2014). For example, valid biological markers of an exposure may not yet exist, or the agent may be rapidly metabolized such that measured biological levels reflect only recent exposure. In addition, biomarkers may be unable to provide information on duration of exposure, which may have an equal or greater impact on a given health outcome compared with cumulative exposure or intensity of exposure. Under such circumstances, the observed association between current or recent exposures and outcome risk may not reflect the association with more etiologically relevant exposure. The ideal way to overcome this limitation, as well as to manage uncertainty about critical exposure window(s), is to enroll cohort members early in life and collect biomarkers repeatedly over an extended time so that exposures at different ages and calendar periods can be documented for many years prior to outcome onset. However, this approach is extremely resource-intensive and often infeasible, especially in a mobile population in which a large proportion of cohort members would become lost to follow-up over time.

Instead, a more practical method to assessing distant past exposures may be to develop models that incorporate information such as historical exposure sources, environmental sampling levels in air, soil, water, and other media, individual locational history (e.g., places of residence, employment, and education), individual exposure opportunities and pathways, and physiology. For example, to estimate local residents' and employees' past exposure to perfluorooctanoic acid (PFOA) released from a manufacturing facility in the Mid-Ohio River Valley, investigators used information on the environmental fate and transport of PFOA in air, surface water, and groundwater; participant-reported demographic characteristics, residential histories, drinking water sources, tap water consumption rates, workplace histories, and body weight; and a single-compartment absorption, distribution, metabolism, and excretion model to estimate the facility's contribution to PFOA in serum (Shin et al., 2011a; Shin et al., 2011b).

An essential step in building such models is to validate estimated exposure levels against directly measured values, ideally over time and in a range of subgroups to ensure model robustness (Chang et al., 2014). Only models that have been rigorously validated and shown to yield predicted values that are highly correlated with measured values should be used in epidemiologic research. Even highly accurate models may rely on unrealistic assumptions such as a fixed residential address, constrained daily mobility, temporal stability in exposure patterns, and randomly distributed measurement error and missing data. Thus, as novel, real-time sensing technologies enable more direct, detailed, and accurate exposure measurement, the NRC envisions a shift in 21st-century exposure science toward less emphasis on models and interpolation, and more emphasis on use of massive exposure datasets (NRC, 2012).

Outcome Assessment

For characterization of an exposure-response relationship, outcome assessment is equally as important as exposure assessment. In some cases, as when study subjects can be linked to a well-established, population-based disease registry that uses medical records for diagnostic confirmation, outcome assessment is relatively straightforward. For example, high-quality cancer and cause-of-death registries exist in many areas and are commonly used for epidemiologic research. However, disease registries are not available for most health outcomes, particularly those that correspond to toxicological endpoints often studied in experimental animals, such as systemic, immunological, neurological, reproductive, and developmental effects (ATSDR, 2015).

When linkage to existing population-based disease registries is not available, outcome ascertainment can be challenging. As with exposures, outcome measures should be accurate, reliable, specific, and sensitive, and ideally derived from objective, clinically confirmed sources or based on standardized, validated tools rather than self-reported, unconfirmed data. Even when the outcome measure fulfills all of these requirements, it is impractical for use in epidemiologic research unless it can readily be ascertained in all study subjects. For example, a disease registry that excludes large segments of the target population, a medical record notation that is missing for many patients, or an invasive diagnostic test that is refused by a substantial proportion of study subjects is not a practical basis for outcome ascertainment in epidemiologic research, because substantial selection bias is likely to occur under such circumstances.

When determining the appropriate outcome(s) to measure in epidemiologic studies designed to inform risk assessment, consideration should be given toward comparability with outcomes of interest in laboratory animal studies. Insight into the MOA/AOP can inform the selection of outcomes that are physiologically comparable between humans and animals, or intermediate or surrogate outcomes reflecting disease processes that precede clinically recognizable diseases or adverse health outcomes. Such outcomes along the pathway between exposure and frank disease may include preclinical health indicators (e.g., peripheral blood counts or serum lipid levels) or cellular biomarkers of toxicological effects (e.g., DNA adducts or gene expression profiles). As with exposures, integration of results from epidemiology and toxicology studies can be facilitated by the use of outcome biomarkers that are relevant and measurable in both humans and animals.

However, unless a surrogate outcome is perfectly predictive of the true clinical endpoint, false positive or false negative results will arise. If possible, studies that use outcome biomarkers should also evaluate the clinical conditions of interest to ensure outcome validity (Strimbu and Tavel, 2010). Moreover, an exposure may act through a biological pathway other than the expected MOA/AOP, especially given that the pathophysiology of a given disease is often incompletely known. Thus, biomarkers as surrogate endpoints should be constantly re-evaluated as additional information emerges from all relevant fields—including data from *in vitro*, *in silico*, animal, and human studies—to help build a biologically plausible MOA/AOP pathway.

Confounder and Effect Modifier Assessment

Confounding is defined as distortion of the estimated association between an exposure and an outcome due to the presence of a common causes or causes of the exposure and the outcome (Figure 1).

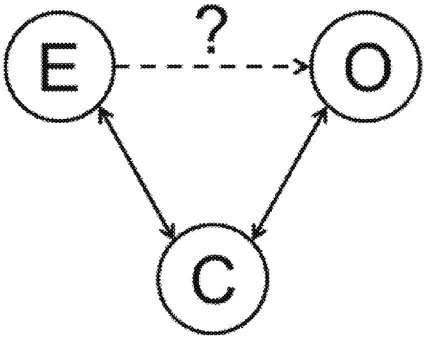

Figure 1. Illustration of confounding. A confounder, C, is a variable that is associated with the exposure (E), independently associated with the outcome (O), and not on the causal pathway between the two. In the presence of uncontrolled confounding, a spurious relationship is detected between the exposure and the outcome.

Effect modification, also referred to as interaction or heterogeneity, is defined as (real or spurious) variation in the exposure-outcome association across levels of another factor (Figure 2).

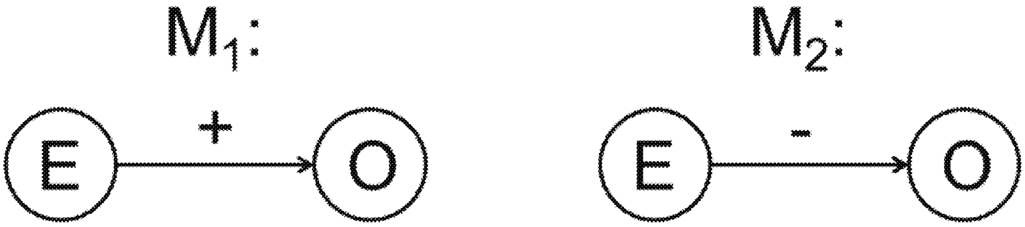

Figure 2. Illustration of effect modification. In the presence of effect modification, the association between the exposure (E) and the outcome (O) varies across levels of the effect modifier (M). Here, the exposure-outcome association is positive (+) within stratum 1 of the effect modifier, whereas the association is negative (-) within stratum 2 of the effect modifier.

Valid, reliable, and thorough assessment of major confounders and effect modifiers is essential for accurate characterization of exposure-outcome relationships. When associations are not reported separately across heterogeneous subgroups of an effect modifier, then the overall association will not apply to all (or even any) segments of the target population. In human health risk assessment, potential effect modification by basic population characteristics such as age, sex, and race/ethnicity are important to consider, as is effect modification by other key environmental chemical exposures. However, investigation of effect modification should be driven by biologically motivated, preferably *a priori* hypotheses, rather than exploratory searches for statistically significant results.

In the absence of adequate measurement and control for confounding, the magnitude and possibly the direction of the observed exposure-outcome association will be incorrect. Even when a confounder is included in a statistical model, residual confounding can occur if the confounder is not classified with sufficient detail or accuracy. For example, adjustment for ever/never smoking status cannot fully control for the confounding effect of tobacco smoking on many exposure-outcome associations. In the case of pesticides, for which occupational exposures usually involve multiple agents as well as other agriculture-related exposures, it can be especially difficult to estimate the independent effect of a single chemical.

When uncontrolled or residual confounding is unavoidable—for instance, when information on a confounder is unavailable or sparse—investigators should not only acknowledge the potential for bias, but also assess the possible magnitude and direction of bias. Assessment of the direction of confounding can be accomplished through the use of directed acyclic graphs to represent causal relationships between variables (VanderWeele et al., 2008). In the case of a dichotomous exposure, outcome, and confounder, the magnitude of confounding is bounded by (i.e., cannot exceed) the strength of the associations of the confounder with the exposure and the outcome, and it also depends on the prevalence of the confounder (Rothman et al., 2012). Knowledge about plausible relationships between these factors can be used to conduct sensitivity analysis of the potential impact of unmeasured but known confounders. However, the magnitude of confounding can greatly exceed these bounds in the presence of several confounders or a multi-level confounder. Moreover, even modest confounding can have a relatively large impact on weak exposure-outcome associations.

Selection Bias

As discussed earlier, the threat of selection bias is one justification for prioritizing prospective cohort studies, where subjects are enrolled prior to the onset of health outcomes of interest, over case-control studies, where subjects are enrolled after disease onset. Selection bias distorts the estimated exposure-outcome association as a result of the procedures used to select subjects into the study or the analysis, or factors that influence study participation. Bias occurs when the exposure-outcome association differs between study participants and all theoretically eligible study subjects—that is, when study selection or participation differs by exposure and outcome status.

In a prospective cohort study, selection bias can occur if data completeness or study attrition—that is, loss to follow-up—is related to exposure and outcome status. Collecting analyzable data from all subjects and maintaining a high rate of follow-up are thus integral to ensuring study validity. Selection bias at initial study enrollment is less likely to occur because subjects are identified prior to the onset of the outcome, although it can arise, for example, if participation is influenced by knowledge of individual risk (e.g., based on family history of a disease) or by certain common health risk factors, such as socioeconomic status.

In practice, adjusting for selection bias can be difficult because detailed data are often lacking for quantitative analysis of differences between participants and non-participants (Porta et al., 2014; Rothman et al., 2012). However, control for selection bias is sometimes possible if the selection factors act (and can therefore be adjusted) like confounders, or if the selection probabilities within each level of the factors affecting selection can be obtained or estimated under plausible assumptions.

Generalizability

Generalizability, or external validity, refers to whether inferences drawn from a study subpopulation can validly be applied to people outside of the study source population. For national regulatory standards, study findings should be applicable to the general population. Internal validity, or accurate measurement of effects among study subjects, is a prerequisite for external validity; results cannot be generalized to other populations if they are not valid even for the source population under study. Whether study results are generalizable to other populations is determined based on theory, expert judgment, and integration of external scientific knowledge, such as understanding of the mechanism underlying an observed exposure-outcome association (Porta et al., 2014). Generalizability can be enhanced by study representativeness—that is, similarity of the study subjects, setting, exposures, and outcomes to other broader populations—especially if the study subjects are representative of the general population of a country or other large geographic region.

In the context of human health risk assessment conducted to inform regulatory decision-making, the generalizability of epidemiologic studies takes on greater importance than in an academic research context. Ideally, studies should enroll sufficient numbers of diverse subjects to enable characterization of variability in risk across different susceptible populations, and to determine whether effect modification occurs by age, sex, race/ethnicity, socioeconomic status, and other characteristics. Subjects with a wide range of exposure levels, including individuals with no exposure, should be studied to enable broad characterization of exposure-response patterns. Direct evaluation of risk in various populations reduces uncertainty resulting from extrapolation or undue generalization from a narrowly defined study population.

Statistical Error and Analysis

In classical statistical hypothesis testing, two types of error can occur: type I error (also called alpha error), which refers to incorrect rejection of a hypothesis, or a "false-positive" test result; and type II error (also called beta error), which refers to failing to reject a false hypothesis, or a "false-negative" test result. Statistical power is the converse of type II error (i.e., 1 – beta); it is the probability that a hypothesis will be rejected if it is false or, roughly, the ability of a study to detect a statistical association if one exists. By convention, power of at least 80% is considered acceptable for a statistical test.

Factors that influence the power of a given test include the study design, the number of subjects, the variability of the outcome, the effect size, and the alpha level. The effect size and outcome variability are fixed, the alpha level is conventionally set at 5%, and the study design may be constrained by resources and data availability. Therefore, the number of subjects is usually the factor most amenable to modification by investigators. More subjects generally lead to greater statistical power (as long as the subjects are not all in a single comparison group), so investigators aim to enroll a sufficient number of subjects to detect the primary hypothesized association(s) with at least 80% power. Especially when power is lower than 80%, but even when it is greater, any statistically nonsignificant result should be interpreted in light of whether the null hypothesis was not rejected due to type II error.

On the other hand, consideration should also be given to whether statistically significant results are due to type I error. When multiple hypotheses are tested simultaneously, the probability of rejecting at least one true null hypothesis is typically high. That is, the probability of making at least one type I error in n tests is $1 - (1 - \alpha)^n$; for $n = 20$ tests, this probability is 64%. The problem of multiple testing is exacerbated when one tests a large number of hypotheses, then focuses *a posteriori* on tests with statistically significant p-values. Most procedures for adjusting statistically for multiple comparisons, such as the methods to control the family-wise error rate (the probability of at least one type I error) or the false discovery rate (the expected proportion of type I errors among the rejected hypotheses), involve adjustment of the alpha level (Glickman et al., 2014). These procedures are based on the assumption that the observed distribution of p-values is unbiased—an assumption that is seldom strictly true in epidemiologic research. At a minimum, authors should report all analyses that were conducted, not only those yielding statistically significant results, and they should clearly differentiate between *a priori* and *a posteriori* hypotheses. In general, the tendency of epidemiologists to highlight significant associations and downplay null results (Kavvoura et al., 2007) contributes to a high ratio of false-positive to false-negative results in the published literature

Other assumptions that underlie any statistical analysis should be tested for validity. For example, the Cox proportional hazards regression model (Cox, 1972), which was originally developed for the analysis of data from clinical trials, is used in epidemiology as a standard approach to the analysis of prospective cohort studies. As opposed to clinical trials in which exposures are fixed and uniform, follow-up is relatively short, and confounding is eliminated by randomization, prospective cohort studies often feature heterogeneous, imprecisely measured, and time-varying exposures, relatively long follow-up, and confounding by many measured and unmeasured risk factors. A fundamental assumption of the Cox proportional hazards regression model—namely, that the relative risk (hazard ratio) is constant over the time scale (age, in many prospective cohort studies)—may frequently be violated, such that the average hazard ratio is uninformative (Hernán, 2010). Instead, directly estimated hazard functions may be the more natural target of estimation in cohort studies.

Communication of Uncertainty

Besides random error, uncertainty in epidemiology can arise from sources of systematic error, including exposure and outcome measurement error, confounding, selection bias and other types of bias, and the need for extrapolation due to insufficient information on exposure levels or susceptible populations of interest. According to the NRC, uncertainty in risk assessment refers to "lack of information, incomplete information, or incorrect information," and it "depends on the quantity, quality, and relevance of data and on the reliability and relevance of models and inferences to fill data gaps" (NRC, 2009). For epidemiologic studies to contribute to the quantification of uncertainty in risk assessment, each study should better describe and, ideally, quantify the magnitude and impact of various sources of uncertainty.

Concerns have long been raised regarding the appropriate characterization and communication of uncertainty in epidemiologic studies for use in human health risk assessment and regulatory decision making (Briggs et al., 2009; Burns et al., 2014; Byrd and Barfield, 1989; Spiegelman, 2010; Stayner et al., 1999). Recommended approaches to improving the delineation of uncertainty in epidemiologic studies, thereby enabling quantitative assessment of and adjustment for potential sources of bias, include the following:

- Describing and discussing in detail all sources of uncertainty, as well as the degree of potential impact on effect estimates; these sources include measurement error, confounding, selection bias, other biases, allowance for latency period, statistical power, choice of data set, choice of statistical model, variation in disease susceptibility, and generalizability, among others;
- Documenting the direction and magnitude of confounder associations with exposures and outcomes;
- Conducting validation studies to test the accuracy of surrogate measures against gold-standard measures;
- Using validation data to quantify the source, type, direction, magnitude, and likelihood of measurement errors;

- Using statistical techniques to adjust for the impact of measurement errors on estimated associations;
- In the absence of validation data, conducting sensitivity or uncertainty analyses to assess the degree of potential bias, based on reasonable assumptions;
- Communicating the extent of uncertainty, including assumptions that underlie key decisions.

More systematic, thorough, and transparent evaluation of uncertainty, as well as collaborative development of standardized approaches to uncertainty assessment across scientific disciplines, would strengthen the utility and application of epidemiologic research for risk assessment and regulatory policy decision making.

Case Studies

Strengths and limitations of specific study design characteristics for human health risk assessment of pesticide exposures can be illustrated through examination of actual epidemiologic studies described in detail in published papers. Two studies that are used as examples in this section are a pair of prospective cohort studies, the Agricultural Health Study and the Columbia Center for Children's Environmental Health (CCCEH) cohort study. Because hundreds of papers have been published from each cohort (AHS, 2015; CCCEH, 2015), this section focuses on the published analyses from each cohort investigating exposure to organophosphate insecticides.

Exposure Assessment

The Agricultural Health Study is a prospective cohort study of 89,656 private pesticide applicators, their spouses, and commercial pesticide applicators recruited in Iowa and North Carolina in 1993–1997 (Alavanja et al., 1996). Exposure to organophosphate insecticides was measured using self-reported data from written questionnaires completed at study enrollment and, to some extent, shortly after enrollment and during follow-up. The initial questionnaire for private pesticide applicators assessed whether subjects had ever personally mixed or applied specific pesticides (including terbufos, fonofos, and trichlorfon, among others), along with duration, days per year, and initial year of use of each pesticide (AHS, 1996). Other pesticides (including malathion, ethyl or methyl parathion, and diazinon, among others) were assessed in less detail based on a question assessing lifetime use. The initial questionnaire for commercial pesticide applicators used similar questions but evaluated different specific pesticides (e.g., malathion, ethyl or methyl parathion, and diazinon with detailed questions; azinphos methyl, phosmet, and tetrachlorvinphos with simplified questions). Additional questions were asked about application methods and use of personal protective equipment for all pesticides or pesticide classes in general. A supplemental take-home questionnaire, completed by 44% of enrolled pesticide applicators, ascertained more detailed information on some pesticides and other covariates.

The Agricultural Health Study questionnaires were highly detailed, thorough, and thoughtfully designed. Few, if any, other epidemiologic studies have conducted more exhaustive questionnaire-based assessment of pesticide exposures. Nevertheless, as discussed earlier, self-reported pesticide use data have substantial drawbacks. These include limited accuracy and reliability of recollected detailed exposures, crude summary measures of exposure that fail to capture important heterogeneity, and only modest correspondence between self-reported exposures and measured biomarker levels, as demonstrated in validation studies conducted in this cohort (Coble et al., 2011). In the context of risk assessment, self-reported pesticide use information also is not readily comparable with controlled doses in laboratory animals. Although self-reported information has the advantage of being able to address distant past exposures, and collecting repeated, validated biomarker data in the entire Agricultural Health Study cohort population would probably have been infeasible in terms of costs and logistics, the reliance on self-reported pesticide exposure data substantially limits the potential for results from this study to be used in dose-response assessment and risk characterization for human health risk assessment.

The CCCEH cohort study measured exposure to several organophosphate insecticides and other pesticides or their metabolites in maternal peripartum and umbilical cord plasma, as well as maternal ambient air samples collected by personal air monitors used during the third trimester of pregnancy (Whyatt et al., 2003; Whyatt et al., 2002). Only chlorpyrifos and diazinon were detected in at least 5% of cord plasma samples or 48-hour maternal ambient air samples, so subsequent analyses of health risk associations focused on these two organophosphate insecticides. A validation study showed stable 2-week integrated indoor air levels of chlorpyrifos within homes (8% of variance explained by within-home variability) and diazinon (6%), and strong correlations were observed between 2-week indoor and 48-hour maternal personal air levels of chlorpyrifos ($\rho = 0.85$) and diazinon ($\rho = 0.90$) (Whyatt et al., 2007). However, chlorpyrifos levels in blood were not associated with personal and indoor air chlorpyrifos levels (Whyatt et al., 2009). Acetyl cholinesterase (AChE) and/or butyl cholinesterase levels were not measured, preventing direct comparisons between observed adverse effect exposure levels and levels producing AChE inhibition in this study population.

The use of specific biomarkers and environmental sampling to quantify exposure to individual organophosphate pesticides in the CCCEH cohort study has obvious advantages, especially in a non-occupationally-exposed population in which self-reported exposure to specific pesticides would be expected to be highly inaccurate. The reliance on one-time measurements taken in the third trimester (maternal ambient air) or shortly after delivery (maternal and umbilical cord plasma) is problematic if exposure changed during pregnancy. For example, changes in dietary patterns could have affected maternal and cord blood levels of organophosphate insecticides over time, and seasonal variability was observed in indoor air levels of pesticides, probably due to use for residential pest control (Whyatt et al., 2007). If exposure changed over time, the measured values might not be representative of earlier exposure levels that could be more etiologically relevant (e.g., with respect to fetal growth and certain other perinatal outcomes). Also, postnatal exposure levels, which might be etiologically relevant to health outcomes later in childhood, were not measured.

However, it is conceivable that exposures were relatively constant during the approximately 40 weeks of pregnancy, making perinatal plasma biomarker levels both etiologically relevant (under the assumption that unmeasured maternal preconception and postnatal exposures are not important) and comparable to experimental doses used in animal studies of developmental toxicity. A validation study in a subset of CCCEH cohort participants showed substantial within-individual variability in maternal prenatal levels of 3,5,6-trichloro-2-pyridinol (TCPy, a metabolite and a primary environmental degradate of chlorpyrifos, chlorpyrifos-methyl, and triclopyr), but the authors did not evaluate intra-individual variability in maternal prenatal blood chlorpyrifos levels (which were not appreciably associated with urinary TCPy levels) (Whyatt et al., 2009).

Outcome Assessment

The Agricultural Health Study used a variety of methods to ascertain health endpoints, depending on the outcome of interest. For example, cancer incidence was ascertained via linkages to statewide cancer registries (Beane Freeman et al., 2005); all-cause and cause-specific mortality was ascertained via linkages to state and national death registries (Lee et al., 2007; Mills et al., 2009); and neurological symptoms, respiratory outcomes, diabetes, and other nonfatal health outcomes were ascertained based on self-report (Hoppin et al., 2006; Kamel et al., 2005; Mills et al., 2009; Montgomery et al., 2008). Well-established population-based cancer registries are generally accepted as providing highly valid and nearly comprehensive ascertainment of incident cancer cases (with some exceptions, such as nonmelanoma skin cancer) in a geographic area. Likewise, death registries are generally accurate and complete for identifying vital status, although specific causes of death are prone to misclassification that can be severe (Kircher et al., 1985; Percy et al., 1981; Smith Sehdev and Hutchins, 2001). Self-reported health outcome data, however, are highly susceptible to misclassification, and are particularly problematic when exposures are also self-reported, such that outcome misclassification may differ systematically by exposure status. In the absence of validation data establishing that the accuracy of self-reported information for specific health outcomes, results based on such data should be interpreted conservatively.

In the CCCEH cohort study, information on birth outcomes was obtained from medical records (Perera et al., 2003; Whyatt et al., 2004), which are objective and generally valid sources, while most neurodevelopmental, cognitive, and behavioral outcomes were assessed based on standardized, validated tools that are widely used in research and medicine (i.e., the Bayley Scales of Infant Development, the Child Behavior Checklist, and the Wechsler Intelligence Scale for Children) (Horton et al., 2012; Lovasi et al., 2011; Rauh et al., 2011; Rauh et al., 2006). Childhood tremor was also assessed using a validated screening tool (hand-drawn spirals) that has been used in other research studies, but population screening tools for tremor are not yet well established (Louis, 2015). Additionally, brain morphology was assessed using high-resolution, T1-weighted magnetic resonance imaging (Rauh et al., 2012), which is a well-established technology, but the interpretation of results with regard to potential neurotoxic effects is not standardized and lacks comparability with other studies.

In general, limitations of standardized assessment tools include the possibility that test administrators can influence the results, that they may not be sufficiently sensitive to capture subtle effects, and that their findings may not correspond to clinically recognizable deficits. With respect to neurological outcomes, it is unclear whether the measures assessed in children using these tools are comparable to those measured in rodent studies, such as cognitive outcomes (e.g., radial arm maze, passive avoidance, conditioned avoidance, novel object recognition), motor activity outcomes (e.g., open field locomotion, figure-eight maze), behavioral outcomes (e.g., time in the open arm in an elevated plus maze, time to start eating in novel environment, chocolate milk preference, forced swim test, time actively interacting in conspecific pairs), sensory function outcomes (e.g., responses to tactile, auditory, olfactory, or visual stimuli), and neuromotor functions (e.g., time to cling to a rod, ability to stay on an increasingly inclined plane or a rotarod) (U.S. EPA, 2015a). In particular, it may not be scientifically justifiable to conclude that positive results for any neurodevelopmental outcome in animals are consistent with or provide toxicological support for results related to another neurodevelopmental outcome in humans.

Confounder and Effect Modifier Assessment

Questionnaires used in both the Agricultural Health Study and the CCCEH cohort study were used to gather detailed information on demographic, environmental, behavioral, and other characteristics that might act as potential confounders or effect modifiers. The Agricultural Health Study questionnaires were particularly extensive (AHS, 1996) and the cohort was sufficiently large as to enable simultaneous statistical adjustment for several potential confounders. For example, associations between diazinon use and cancer risk were adjusted for age, state of residence, education, smoking history and pack-years, alcohol consumption, family history of cancer, and lifetime days of any pesticide application (Beane Freeman et al., 2005), while associations between pesticide use and neurological symptoms were adjusted for age, state, education, smoking pack-years, and alcohol use (Kamel et al., 2007). Effect modification was not systematically examined, but the authors lacked compelling *a priori* hypotheses regarding interactions.

In the CCCEH cohort study, associations with birth outcomes were restricted to nonsmokers (plasma cotinine ≤ 25 ng/mL) and adjusted for measures of maternal body size, parity, newborn sex, and gestational age (Perera et al., 2003), and later additionally for ethnicity, environmental tobacco smoke in the home, and season of delivery (Whyatt et al., 2004). Associations with neurodevelopmental outcomes were adjusted for age, sex, race/ethnicity, maternal IQ, maternal education, quality of the home care-taking environment, and prenatal environmental tobacco smoke exposure (Rauh et al., 2011; Rauh et al., 2006), and later additionally for various neighborhood-level sociodemographic and housing characteristics (Lovasi et al., 2011). Effect modification was examined by race/ethnicity (Perera et al., 2003), calendar period (Whyatt et al., 2004), quality of the home environment (Horton et al., 2012), child sex, and other covariates (Rauh et al., 2011), but statistical power was limited in subgroup analyses.

Selection Bias

Over 80% of eligible pesticide applicators and 75% of spouses of married private applicators enrolled in the Agricultural Health Study during the initial recruitment phase, which took place at licensing facilities for application of restricted-use pesticides (AHS, 1996). However, only 44% of enrolled pesticide applicators completed the detailed take-home questionnaire shortly after enrollment, and participation in follow-up questionnaires was also highly incomplete (64% of private applicators, 59% of commercial applicators, and 74% of spouses in phase 2; 46% of private applicators and 62% of spouses in phase 3) (AHS, 1996). Thus, considerable selection bias could have occurred if nonparticipation was related to exposure and health status. A formal analysis of bias due to study drop-out does not appear to have been conducted. However, an analysis of bias due to missing data—another form of selection bias—revealed that subjects with complete covariate data were substantially different from those with missing data (Lash, 2007). Thus, in analyses relying on follow-up questionnaires or relying on covariates with a high degree of missing data, selection bias is a major concern in the Agricultural Health Study.

For the CCCEH cohort study, eligible pregnant women were identified from prenatal clinics at two New York City hospitals, and about 70% agreed to participate in the study (Whyatt et al., 2002). Participants were somewhat younger and more likely to be African American than nonparticipants; other differences, if measured, were not described. Of 314 mother-newborn pairs eligible for the analysis of birth outcomes, umbilical cord blood chlorpyrifos or diazinon levels or pesticide levels in paired maternal air and blood samples were available for 82% of subjects (Whyatt et al., 2004). Reasons for missing data in the remaining 18% were not provided, and the incomplete exposure data raise the possibility of modest selection bias. At 3 years the retention rate in the full cohort was 83%, with no significant differences between participating and nonparticipating pairs in terms of maternal age, ethnicity, marital status, education, income, newborn gestational age, or newborn birth weight (Rauh et al., 2006). However, analyses were further restricted to 90% of children with the requisite data at 12, 24, or 36 months, and 74% with data at all three time points. At 7 years the retention rate in the full cohort was essentially unchanged at 82%, with no significant sociodemographic differences between participants and nonparticipants, but the proportion with complete data was not reported (Rauh et al., 2011). Even though the authors reported that the included subjects did not differ from the full cohort with respect to demographic characteristics (Horton et al., 2012), it is unknown whether they differed in terms of exposures and outcomes. Thus, although loss to follow-up was reasonable, additional exclusion due to missing data could have introduced additional selection bias.

Generalizability

The Agricultural Health Study was restricted to licensed private and commercial pesticide applicators and spouses of private pesticide applicators residing in Iowa and North Carolina at study entry (Alavanja et al., 1996). Exposures to organophosphate insecticides are anticipated to be higher in this occupational cohort than in the general population, and results therefore may not be generalizable to lower-level, nonoccupational exposures. Numerous types of organophosphate insecticides were used by study subjects, with a sufficient range in annual days and lifetime days of use to provide informative exposure variability (Hoppin et al., 2012).

A comparison of farmers enrolled in the Agricultural Health Study with data from the 1992 and 1997 Censuses of Agriculture for Iowa and North Carolina revealed that study participants were younger, lived or worked on larger farms, more frequently applied herbicides, insecticides, and fungicides, and were more likely to raise beef cattle and swine and grow corn, soybeans, hay, and oats (in Iowa) or more likely to grow crops commonly seen in the state (in North Carolina) (Lynch et al., 2005). Thus, cohort members probably experienced heavier pesticide usage relative to farmers statewide, and results from this study may not apply even to lower-level occupational exposures in the agricultural industry. Results also cannot reliably be generalized to other subpopulations not represented by the study subjects.

Eligible subjects for the CCCEH cohort study were women aged 18–35 years who had resided in northern Manhattan (Central Harlem or Washington Heights/Inwood) or the South Bronx for at least 1 year before pregnancy, self-identified as either African American or Dominican, did not smoke cigarettes, use other tobacco products, or illicit drugs during pregnancy, did not have diabetes, hypertension, or known HIV, and had their first prenatal visit by the 20th week of pregnancy (Whyatt et al., 2002). These restrictive eligibility criteria may limit the generalizability of results from this study, which may not apply to exposure levels or scenarios found in other settings and populations, such as people with occupational pesticide exposures, those living in rural areas, other racial/ethnic groups, people with less residential stability or access to health care, or high-risk pregnancies. The detection of only chlorpyrifos and diazinon, but not other organophosphate insecticides, in an appreciable proportion of plasma and personal ambient air samples among study subjects, along with the relatively narrow range of low-level exposures for the majority of participants, also limits generalizability to more highly exposed individuals. Given that the authors found an association between pre-2001 but not post-2001 cord plasma chlorpyrifos or diazinon levels and fetal growth, it is possible that associations vary by the absolute level of exposure.

Statistical Error and Analysis

In general, statistical methods appeared to be appropriate and were described in detail, and a few papers described the use of alternative methods to evaluate the robustness of the primary statistical models (Beard et al., 2014; Rauh et al., 2011; Starks et al., 2012). Testing of basic regression assumptions—e.g., that the disease rate (for linear regression) or its logarithm (for Poisson regression) changes linearly with equal increment increases in the exposure variable; that changes in the rate from combined effect of different covariates are additive (for linear regression) or multiplicative (for Poisson regression); that the variance of the errors is constant (for linear regression) or the variance of the number of cases is equal to the mean (for Poisson regression) at each level of the covariates—generally was not described, although this is not unusual for epidemiologic papers.

Authors of epidemiologic studies typically not report statistical power to detect specific associations, in part because power calculations are based on assumptions that may not apply, and they assume the absence of bias and confounding. Thus, the general lack of discussion of statistical power in papers from the Agricultural Health Study and the CCCEH cohort study was not exceptional. Nevertheless, insufficient power could have contributed to some statistically nonsignificant associations with rare exposures and/or outcomes in the Agricultural Health Study or any associations in the CCCEH cohort study, in which most analyses were based on 200–300 subjects and subgroup analyses were based on fewer.

Many hypotheses were tested within and among papers derived from these cohorts, and chance was mentioned as a possible explanation for significant findings in a minority of papers. Some studies distinguished between primary and secondary or exploratory analyses, whereas others did not. Although virtually all studies reported at least some statistically nonsignificant findings, it is impossible to determine whether additional analyses yielding null results were not reported.

Communication of Uncertainty

Exposure reliability and validation studies were conducted in both cohorts to evaluate measurement error (e.g., (Blair et al., 2002; Blair et al., 2011; Coble et al., 2011; Hoppin et al., 2002; Whyatt et al., 2009; Whyatt et al., 2007)), and several papers included a limited set of sensitivity analyses that examined the potential impact of certain assumptions or types of bias (Bonner et al., 2007; Jones et al., 2015; Lovasi et al., 2011; Rauh et al., 2011; Starks et al., 2012). Potential exposure and outcome measurement error, information bias, selection bias, and confounding were often discussed, although rarely with estimates of the direction and magnitude of influence. In general, quantitative approaches to correct for measurement error were not implemented.

Very few epidemiologic studies to date have rigorously evaluated and communicated quantitative measures of uncertainty. The Agricultural Health Study and CCCEH cohorts went farther than most in terms of conducting validation studies and sensitivity analyses, acknowledging sources of error and bias, and documenting exposure assessment approaches. Additional efforts from these high-profile studies would help to lead the way toward more thorough, transparent, and quantitative characterization of uncertainty in the field of epidemiology in general.

Feasibility

Large, rigorous epidemiologic studies are generally very complex and resource-intensive to conduct. Prospective cohort studies in particular are the most expensive and time-consuming studies to implement, typically requiring millions or billions of dollars, thousands of participants, hundreds of staff, and years or decades of follow-up time (except in the case of birth cohorts examining neonatal or early childhood outcomes). The Framingham Heart Study, the nation's longest-running prospective cohort study of cardiovascular disease that has been ongoing since 1948, was until recently receiving approximately $9 million per year from the National Institutes of Health (Barlow, 2013). The National Children's Study, a planned prospective cohort study that would have followed 100,000 U.S. children from before birth to age 21 years, was dissolved after more than $1.2 billion had already been spent on trying to launch the study, which was halted due to study design failures (Kaiser, 2014).

Designing and adhering to standardized protocols, enrolling a sufficient number and range of participants, repeatedly collecting questionnaires, biospecimens, medical records, and other data from a large proportion of participants over time, sustaining long-term staff, maintaining data integrity, and obtaining continued funding support are all major challenges to conducting a successful prospective cohort study. The demands of these prerequisites probably explain the relatively infrequency of prospective cohorts in epidemiologic research, compared with more readily implemented study designs such as case-control and retrospective cohort studies. However, the rewards of large investments in prospective cohort studies are seen in the hundreds of major scientific papers generated from single cohorts, the profound public health impact of their findings, and the tendency of health and regulatory agencies to rely preferentially on the results of prospective cohort studies whenever available.

Weight-of-evidence Assessment of Causation

In the "hazard identification" phase of the risk assessment paradigm, the relevant scientific evidence is identified and reviewed to determine whether exposure to a specific agent can cause a particular health outcome (NRC, 1983). As a basis for concluding whether observed associations are likely to be causal, the results of any individual epidemiologic study must be interpreted in the context of the complete body of pertinent epidemiologic literature, combined with supporting evidence from other scientific fields, including toxicology and mode-of-action studies.

The guidelines put forth by Sir Austin Bradford Hill in 1965 for evaluating the causality of an exposure-outcome association (Hill, 1965) are commonly cited and implemented in epidemiology, sometimes in slightly modified form, and they are broadly accepted in the scientific community (Federal Judicial Center and National Research Council of the National Academies, 2011; Gordis, 2000; Hennekens and Buring, 1987; Lilienfeld and Stolley, 1994; Mausner and Kramer, 1985; Rothman et al., 2012; Schlesselman, 1982). These nine guidelines for evaluating the current state of knowledge regarding an exposure-outcome association are strength, consistency, specificity, temporality, biological gradient, plausibility, coherence, experiment, and analogy. Hill referred to these as nine "features to be specially considered" or "viewpoints" on the basis of which one should evaluate associations before declaring them causal. He did not assert that all guidelines must be met to establish causality; rather, he stated: "None of my nine viewpoints can bring indisputable evidence for or against the cause-and-effect hypothesis and none can be required as a *sine qua non*" (Hill, 1965). Nevertheless, these guidelines provide a useful, logical, and widely used framework for evaluating the weight of scientific evidence in favor of causality.

When evaluating the overall body of literature, the quality of each epidemiologic study—that is, its likelihood of yielding valid results—should also be taken into account. In general, results from studies that are more likely to be valid should be carry greater weight in a weight-of-evidence assessment. Guidelines for assessing the quality of individual epidemiologic studies also exist (Downs and Black, 1998; Genaidy et al., 2007; Guyatt et al., 2011; Johnson et al., 2014; Viswanathan and Berkman, 2011; Wells et al., 2014), but no set of criteria is widely accepted and used. Standard considerations for evaluating epidemiologic study quality are those discussed in this document, namely, study design, validity and reliability of the assessment of exposures, outcomes, and other covariates, confounder adjustment, potential for selection bias and other types of bias, generalizability, statistical power, approach to multiple testing, and appropriateness of the statistical analysis. Several existing scales for epidemiologic study quality assessment also take clarity and completeness of reporting into account, but these criteria do not directly affect study validity. For example, if a prospective cohort study has substantial loss to follow-up, the risk of selection bias will high regardless of whether the loss to follow-up is clearly described. Thus, although poorly described studies are often difficult to interpret, thoroughly described studies should not necessarily be treated as more likely to be valid.

A framework for evaluating human epidemiologic studies for quantitative risk assessment has been formulated by the European Union Network of Excellence Environmental Cancer Risk, Nutrition and Individual Susceptibility (ECNIS) Integrated Risk Assessment Group (Vlaanderen et al., 2008). The framework is based on three tiers: the first tier consists of criteria to exclude studies that are not suitable for quantitative risk assessment, the second tier consists of criteria to exclude studies that have an inappropriate study design for quantitative risk assessment, and to select appropriate criteria for further evaluation in the third tier, which consists of design-specific criteria to rank and ultimately select the studies for inclusion in quantitative risk assessment. From the first tier, a set of minimum criteria can be formulated for epidemiologic studies to inform quantitative risk assessment (Box 1). After categorization based on study design (second tier), the considerations in the third tier are comparable to those used to evaluate epidemiologic study quality in general, namely, potential for selection bias (response rate, loss to follow-up), outcome assessment (minimum follow-up time, blinded health outcome assessment), exposure assessment (quality of the exposure measurement methods, application of exposure measurements in exposure assessment, type of exposure metric, specificity of the exposure indicator, blinded exposure assessment, quality of the exposure assignment strategy), generalizability (insight in the variability of exposure; also, response rate and loss to follow-up), and potential for other types of bias (potential for information bias, insight in the potential for systematic error in study results; also, blinded exposure and outcome assessment). By providing a systematic and transparent method to select and rank epidemiologic studies based on quality and relevance, these guidelines aid in the assessment of the epidemiologic weight of evidence for quantitative risk assessment.

Box 1. Minimum parameters for the design of epidemiologic studies to inform risk assessment

- Cohort, case-control, cross-sectional, or hybrid design
- Individual-level exposure, outcome, and covariate data
- Specific, quantitative exposure assessment
- Specific outcome assessment based on accepted norms or standards
- Measurement and adjustment of all relevant strong confounders
- Sufficient description of study selection criteria to enable assessment of potential selection bias
- Sufficient description of statistical analysis to enable assessment of assumptions and appropriateness
- Sufficient description of results to enable assessment of hypothesis tests conducted
- Sufficient description of subject characteristics and exposure variability to enable assessment of generalizability

Bridging Toxicology and Epidemiology

An important challenge in incorporating epidemiology into risk assessment is facilitating better understanding of the strengths and limitations of epidemiology, including both the overall science and individual research studies, among toxicologists and other laboratory scientists. Whereas animal toxicologists are accustomed to standard study design specifications for risk assessment, epidemiology is not amenable to fixed design parameters. Observational research tends to focus on people in their natural settings, where they largely choose their own exposures (either directly or indirectly, for example, by choosing where to live or work), as well as whether to participate in studies. Consequently, observational epidemiology studies are often more realistic and generalizable than, for example, randomized clinical trials based on highly restricted patient populations. However, the observational nature of epidemiology also makes this field of research susceptible to uncertainty (e.g., regarding environmental and behavioral exposures, especially in the past) and bias due to systematic differences between exposed and unexposed, and diseased and nondiseased groups.

In epidemiology, there is no universal "ideal study design." The appropriate study design for a given exposure-outcome association depends, for example, on the prevalence and variability of the exposure, the rate of the outcome, the anticipated strength of the association, the latency period between exposure and outcome, the window(s) of susceptibility, and the presence of key effect modifiers. A prospective study design is often preferred, but not for rare outcomes, especially those with a long latency period during which study attrition might be high. Forty weeks of follow-up might be ideal for studies of birth outcomes, but woefully inadequate for studies of cancer incidence. Two hundred study subjects might be sufficient to detect a strong association with a common risk factor, but inadequate to detect a weak association. None of these factors are under the investigator's control, and epidemiologists often have little empirical guidance upon which to base assumptions regarding these factors when they design a study. As a result, an epidemiologist's answer to questions about the choice of study design, sample size/statistical power, duration of follow-up, and other study characteristics is generally that "it depends."

Understandably, this response can be frustrating to laboratory scientists who are more familiar with clear-cut study requirements. To bridge this gap, greater communication and cross-disciplinary education are needed so that non-epidemiologists can distinguish between high- and low-quality studies and given appropriate weight to their results. A checklist approach cannot be used to evaluate the quality of epidemiologic studies. Instead, well-informed judgment is required—yet it is not enough for epidemiologists simply to say that they know a good study when they see one. Instead, EPA and other risk-informed regulatory agencies should foster interaction and collaboration between epidemiologists and toxicologists so that scientists in both fields can interpret and use the full spectrum of relevant studies for the purposes of risk assessment.

Future of Epidemiology

Epidemiologic research on disease etiology is increasingly incorporating tools from molecular biology, genomics, and other high-throughput “omic" technologies to measure numerous exposures and/or outcomes simultaneously. This line of research into the cellular and molecular underpinnings of exposure-outcome associations is highly compatible with the National Research Council (NRC)’s visions for 21st-century toxicity testing (NRC, 2007) and exposure science (NRC, 2012), and gives promise to the idea of illuminating MOAs and AOPs by integrating information from epidemiology, toxicology, and related scientific fields. With growing capacity to rapidly and inexpensively measure a wide array of exposures or outcomes, important considerations will be how to properly collect, process, and store biospecimens for future research use, and how to evaluate the validity and reliability of new technologies, as well as the reproducibility of results that they generate.

As part of its vision for cancer epidemiology in the 21st century, the U.S. National Cancer Institute (NCI) recommends expanding cohort studies to collect exposure, clinical, and other information across the lifespan and to include multiple health outcomes (Khoury et al., 2013). In making this recommendation, the NCI recognizes that assembling a cohort with documented medical histories and exposure information and appropriate biospecimens is a daunting endeavor, especially within the U.S. health care system. One suggested approach to expanding cohort studies is to foster collaboration, replication, and translation by increasing data and biospecimen access and sharing, harmonization, and joint analyses based on already existing cohorts. Cohorts are being leveraged to incorporate data on additional exposures and health endpoints through linkages to existing sources such as federal, state, and local environmental data systems, electronic health records, Medicare/Medicaid, and disease registries. In addition, investigators responsible for cohort studies are participating in large-scale efforts to conduct parallel and pooled analyses that require large numbers, detailed data, diverse populations, and independent replication (NCI, 2015). These collaborative efforts will help to bolster molecular epidemiology research, in which associations are often modest in magnitude and discoveries concerning disease mechanisms can be incremental.

Part of extending the reach and impact of epidemiology to have a greater impact on public health policy—another goal of the NCI for the 21st century (Khoury et al., 2013)—is to increase the relevance and utility of epidemiologic studies to human health risk assessment. Large, rigorous prospective cohort studies with validated exposure biomarker data, confirmed health outcomes analogous to animal endpoints, thorough adjustment for confounding and investigation of effect modification, minimal opportunity for selection bias and other biases, and a broad range of susceptible populations and exposures offer the greatest potential for integration with laboratory data from animal studies to serve as a solid scientific foundation for effective, evidence-based regulatory action.

References

Acquavella JF, Alexander BH, Mandel JS, Burns CJ, Gustin C. (2006). Exposure misclassification in studies of agricultural pesticides: insights from biomonitoring. Epidemiology 17:69-74

Acquavella JF, Alexander BH, Mandel JS, Gustin C, Baker B, Chapman P, Bleeke M. (2004). Glyphosate biomonitoring for farmers and their families: results from the Farm Family Exposure Study. Environ Health Perspect 112:321-6

AHS. (2015). Publications. News & Findings. Agricultural Health Study (AHS). Available at: http://aghealth.nih.gov/news/publications.html. Last accessed: 26 October 2015.

AHS. (1996). Questionnaires & Study Data. Scientific Collaboration. Agricultural Health Study (AHS). Available at: http://aghealth.nih.gov/collaboration/questionnaires.html. Last accessed: 26 October 2015.

Alavanja MC, Sandler DP, McMaster SB, Zahm SH, McDonnell CJ, Lynch CF, Pennybacker M, Rothman N, Dosemeci M, Bond AE, Blair A. (1996). The Agricultural Health Study. Environ Health Perspect 104:362-9

ATSDR. (2015). Toxicological Profiles. Agency for Toxic Substances & Disease Registry (ATSDR). Available at: http://www.atsdr.cdc.gov/toxprofiles/index.asp. Last updated: 27 January 2015

Barlow R. (2013). Framingham Heart Study carries on, despite budget cuts. 1 August 2013. Available at: http://www.bu.edu/today/2013/framingham-heart-study-carries-on-despite-budget-cuts/. BU Today. Boston, MA

Beane Freeman LE, Bonner MR, Blair A, Hoppin JA, Sandler DP, Lubin JH, Dosemeci M, Lynch CF, Knott C, Alavanja MC. (2005). Cancer incidence among male pesticide applicators in the Agricultural Health Study cohort exposed to diazinon. Am J Epidemiol 162:1070-9

Beard JD, Umbach DM, Hoppin JA, Richards M, Alavanja MC, Blair A, Sandler DP, Kamel F. (2014). Pesticide exposure and depression among male private pesticide applicators in the agricultural health study. Environ Health Perspect 122:984-91

Blair A, Tarone R, Sandler D, Lynch CF, Rowland A, Wintersteen W, Steen WC, Samanic C, Dosemeci M, Alavanja MC. (2002). Reliability of reporting on life-style and agricultural factors by a sample of participants in the Agricultural Health Study from Iowa. Epidemiology 13:94-9

Blair A, Thomas K, Coble J, Sandler DP, Hines CJ, Lynch CF, Knott C, Purdue MP, Zahm SH, Alavanja MC, Dosemeci M, Kamel F, Hoppin JA, Freeman LB, Lubin JH. (2011). Impact of pesticide exposure misclassification on estimates of relative risks in the Agricultural Health Study. Occup Environ Med 68:537-41

Blair A, Zahm SH. (1990). Methodologic issues in exposure assessment for case-control studies of cancer and herbicides. Am J Ind Med 18:285-93

Bonner MR, Coble J, Blair A, Beane Freeman LE, Hoppin JA, Sandler DP, Alavanja MC. (2007). Malathion exposure and the incidence of cancer in the agricultural health study. Am J Epidemiol 166:1023-34

Briggs DJ, Sabel CE, Lee K. (2009). Uncertainty in epidemiology and health risk and impact assessment. Environ Geochem Health 31:189-203

Burns CJ, Wright JM, Pierson JB, Bateson TF, Burstyn I, Goldstein DA, Klaunig JE, Luben TJ, Mihlan G, Ritter L, Schnatter AR, Symons JM, Yi KD. (2014). Evaluating uncertainty to strengthen epidemiologic data for use in human health risk assessments. Environ Health Perspect 122:1160-5

Byrd DM, Barfield ET. (1989). Uncertainty in the estimation of benzene risks: application of an uncertainty taxonomy to risk assessments based on an epidemiology study of rubber hydrochloride workers. Environ Health Perspect 82:283-7

Calderon RL. (2000). Measuring risks in humans: the promise and practice of epidemiology. Food Chem Toxicol 38:S59-63

CCCEH. (2015). Center Scientific Papers. Columbia Center for Children's Environmental Health (CCCEH). Mailman School of Public Health, Columbia University. Available at: http://ccceh.org/our-research/scientific-papers. Last accessed: 26 October 2015.

Chang ET, Adami HO, Bailey WH, Boffetta P, Krieger RI, Moolgavkar SH, Mandel JS. (2014). Validity of geographically modeled environmental exposure estimates. Crit Rev Toxicol 44:450-66

Coble J, Thomas KW, Hines CJ, Hoppin JA, Dosemeci M, Curwin B, Lubin JH, Beane Freeman LE, Blair A, Sandler DP, Alavanja MC. (2011). An updated algorithm for estimation of pesticide exposure intensity in the agricultural health study. Int J Environ Res Public Health 8:4608-22

Cox DR. (1972). Regression models and life-tables. J Roy Stat Soc Ser B 34:187-220

Downs SH, Black N. (1998). The feasibility of creating a checklist for the assessment of the methodological quality both of randomised and non-randomised studies of health care interventions. J Epidemiol Community Health 52:377-84

Federal Judicial Center, National Research Council of the National Academies. (2011).Reference Manual on Scientific Evidence. Third Edition. Washington, D.C.: National Academies Press

FIFRA Scientific Advisory Panel. (2010). SAP Minutes No. 2010-03. A Set of Scientific Issues Being Considered by the Environmental Protection Agency Regarding: Draft Framework and Case Studies on Atrazine, Human Incidents, and the Agricultural Health Study: Incorporation of Epidemiology and Human Incident Data into Human Health Risk Assessment. February 2–4, 2010, FIFRA Scientific Advisory Panel Meeting, Held at the Environmental Protection Agency Conference Center, Arlington, Virginia.

Garfitt SJ, Jones K, Mason HJ, Cocker J. (2002). Exposure to the organophosphate diazinon: data from a human volunteer study with oral and dermal doses. Toxicol Lett 134:105-13

Genaidy AM, Lemasters GK, Lockey J, Succop P, Deddens J, Sobeih T, Dunning K. (2007). An epidemiological appraisal instrument - a tool for evaluation of epidemiological studies. Ergonomics 50:920-60

Glickman ME, Rao SR, Schultz MR. (2014). False discovery rate control is a recommended alternative to Bonferroni-type adjustments in health studies. J Clin Epidemiol 67:850-7

Gordis L. (2000).Epidemiology. Philadelphia: WB Saunders

Guyatt GH, Oxman AD, Vist G, Kunz R, Brozek J, Alonso-Coello P, Montori V, Akl EA, Djulbegovic B, Falck-Ytter Y, Norris SL, Williams JW, Jr., Atkins D, Meerpohl J, Schunemann HJ. (2011). GRADE guidelines: 4. Rating the quality of evidence--study limitations (risk of bias). J Clin Epidemiol 64:407-15

Hennekens CH, Buring JE. (1987).Epidemiology in Medicine. Boston: Little Brown and Company

Hernán MA. (2010). The hazards of hazard ratios. Epidemiology 21:13-5

Hertz-Picciotto I. (1995). Epidemiology and quantitative risk assessment: a bridge from science to policy. Am J Public Health 85:484-91

Hill AB. (1965). The environment and disease: association or causation? Proc R Soc Med 58:295-300

Hoppin JA, Long S, Umbach DM, Lubin JH, Starks SE, Gerr F, Thomas K, Hines CJ, Weichenthal S, Kamel F, Koutros S, Alavanja M, Beane Freeman LE, Sandler DP. (2012). Lifetime organophosphorous insecticide use among private pesticide applicators in the Agricultural Health Study. J Expo Sci Environ Epidemiol 22:584-92

Hoppin JA, Umbach DM, London SJ, Lynch CF, Alavanja MC, Sandler DP. (2006). Pesticides and adult respiratory outcomes in the agricultural health study. Ann N Y Acad Sci 1076:343-54

Hoppin JA, Yucel F, Dosemeci M, Sandler DP. (2002). Accuracy of self-reported pesticide use duration information from licensed pesticide applicators in the Agricultural Health Study. J Expo Anal Environ Epidemiol 12:313-8

Horton MK, Kahn LG, Perera F, Barr DB, Rauh V. (2012). Does the home environment and the sex of the child modify the adverse effects of prenatal exposure to chlorpyrifos on child working memory? Neurotoxicol Teratol 34:534-41

Johnson PI, Sutton P, Atchley DS, Koustas E, Lam J, Sen S, Robinson KA, Axelrad DA, Woodruff TJ. (2014). The Navigation Guide - evidence-based medicine meets environmental health: systematic review of human evidence for PFOA effects on fetal growth. Environ Health Perspect 122:1028-39

Jones RR, Barone-Adesi F, Koutros S, Lerro CC, Blair A, Lubin J, Heltshe SL, Hoppin JA, Alavanja MC, Beane Freeman LE. (2015). Incidence of solid tumours among pesticide applicators exposed to the organophosphate insecticide diazinon in the Agricultural Health Study: an updated analysis. Occup Environ Med 72:496-503

Kaiser J. (2014). NIH cancels massive U.S. children's study. 12 December 2014. Available at: http://news.sciencemag.org/funding/2014/12/nih-cancels-massive-u-s-children-s-study. Science

Kamel F, Engel LS, Gladen BC, Hoppin JA, Alavanja MC, Sandler DP. (2007). Neurologic symptoms in licensed pesticide applicators in the Agricultural Health Study. Hum Exp Toxicol 26:243-50

Kamel F, Engel LS, Gladen BC, Hoppin JA, Alavanja MC, Sandler DP. (2005). Neurologic symptoms in licensed private pesticide applicators in the agricultural health study. Environ Health Perspect 113:877-82

Kavvoura FK, Liberopoulos G, Ioannidis JP. (2007). Selection in reported epidemiological risks: an empirical assessment. PLoS Med 4:e79

Khoury MJ, Lam TK, Ioannidis JP, Hartge P, Spitz MR, Buring JE, Chanock SJ, Croyle RT, Goddard KA, Ginsburg GS, Herceg Z, Hiatt RA, Hoover RN, Hunter DJ, Kramer BS, Lauer MS, Meyerhardt JA, Olopade OI, Palmer JR, Sellers TA, Seminara D, Ransohoff DF, Rebbeck TR, Tourassi G, Winn DM, Zauber A, Schully SD. (2013). Transforming epidemiology for 21st century medicine and public health. Cancer Epidemiol Biomarkers Prev 22:508-16

Kircher T, Nelson J, Burdo H. (1985). The autopsy as a measure of accuracy of the death certificate. N Engl J Med 313:1263-9

Lash TL. (2007). Bias analysis applied to Agricultural Health Study publications to estimate non-random sources of uncertainty. J Occup Med Toxicol 2:15

Lavelle KS, Robert Schnatter A, Travis KZ, Swaen GM, Pallapies D, Money C, Priem P, Vrijhof H. (2012). Framework for integrating human and animal data in chemical risk assessment. Regul Toxicol Pharmacol 62:302-12

Lee WJ, Alavanja MC, Hoppin JA, Rusiecki JA, Kamel F, Blair A, Sandler DP. (2007). Mortality among pesticide applicators exposed to chlorpyrifos in the Agricultural Health Study. Environ Health Perspect 115:528-34

Lilienfeld DE, Stolley PD. (1994).Foundations of Epidemiology. New York: Oxford University Press

Louis ED. (2015). Utility of the hand-drawn spiral as a tool in clinical-epidemiological research on essential tremor: data from four essential tremor cohorts. Neuroepidemiology 44:45-50

Lovasi GS, Quinn JW, Rauh VA, Perera FP, Andrews HF, Garfinkel R, Hoepner L, Whyatt R, Rundle A. (2011). Chlorpyrifos exposure and urban residential environment characteristics as determinants of early childhood neurodevelopment. Am J Public Health 101:63-70

Lynch CF, Sprince NL, Heywood E, Pierce J, Logsden-Sackett N, Pennybacker M, Alavanja MC. (2005). Comparison of farmers in the Agricultural Health Study to the 1992 and the 1997 censuses of agriculture. J Agromedicine 10:13-22

Mausner JS, Kramer S. (1985).Epidemiology: An Introductory Text. Philadelphia: WB Saunders Company

Mills KT, Blair A, Freeman LE, Sandler DP, Hoppin JA. (2009). Pesticides and myocardial infarction incidence and mortality among male pesticide applicators in the Agricultural Health Study. Am J Epidemiol 170:892-900

Montgomery MP, Kamel F, Saldana TM, Alavanja MC, Sandler DP. (2008). Incident diabetes and pesticide exposure among licensed pesticide applicators: Agricultural Health Study, 1993-2003. Am J Epidemiol 167:1235-46

NCI. (2015). NCI Cohort Consortium. Epidemiology and Genomics Research Program, Division of Cancer Control and Population Sciences, National Cancer Institute (NCI). Available at: http://epi.grants.cancer.gov/Consortia/cohort.html. Last updated: 5 October 2015.

NRC. (2012). Exposure Science in the 21st Century: A Vision and a Strategy. Washington, DC: Committee on Human and Environmental Exposure Science in the 21st Century, Board on Environmental Studies and Toxicology, Division on Earth and Life Sciences, National Research Council (NRC), National Academies Press

NRC. (1983). Risk Assessment in the Federal Government: Managing the Process Washington, DC: Committee on the Institutional Means for Assessment of Risks to Public Health, National Research Council (NRC), National Academies Press

NRC. (2009). Science and Decisions: Advancing Risk Assessment. Washington, DC: National Academies Press

NRC. (2007). Toxicity Testing in the 21st Century: A Vision and a Strategy. Washington, DC: Committee on Toxicity Testing and Assessment of Environmental Agents, National Research Council (NRC), National Academies Press

Percy C, Stanek E, 3rd, Gloeckler L. (1981). Accuracy of cancer death certificates and its effect on cancer mortality statistics. Am J Public Health 71:242-50

Perera FP, Rauh V, Tsai WY, Kinney P, Camann D, Barr D, Bernert T, Garfinkel R, Tu YH, Diaz D, Dietrich J, Whyatt RM. (2003). Effects of transplacental exposure to environmental pollutants on birth outcomes in a multiethnic population. Environ Health Perspect 111:201-5

Porta M, Greenland S, Hernán M, dos Santos Silva I, Last JM, Eds. (2014). A Dictionary of Epidemiology. New York: Oxford University Press

Rauh V, Arunajadai S, Horton M, Perera F, Hoepner L, Barr DB, Whyatt R. (2011). Seven-year neurodevelopmental scores and prenatal exposure to chlorpyrifos, a common agricultural pesticide. Environ Health Perspect 119:1196-201

Rauh VA, Garfinkel R, Perera FP, Andrews HF, Hoepner L, Barr DB, Whitehead R, Tang D, Whyatt RW. (2006). Impact of prenatal chlorpyrifos exposure on neurodevelopment in the first 3 years of life among inner-city children. Pediatrics 118:e1845-59

Rauh VA, Perera FP, Horton MK, Whyatt RM, Bansal R, Hao X, Liu J, Barr DB, Slotkin TA, Peterson BS. (2012). Brain anomalies in children exposed prenatally to a common organophosphate pesticide. Proc Natl Acad Sci U S A 109:7871-6

Rothman KJ, Greenland S, Lash TL. (2012).Modern Epidemiology. Third Edition, Mid-Cycle Revision. Philadelphia: Lippincott Williams & Wilkins

Schlesselman JJ. (1982).Case-Control Studies: Design, Conduct, Analysis. New York: Oxford University Press

Schmidt CW. (2006). Signs of the times: biomarkers in perspective. Environ Health Perspect 114:A700-5

Shin HM, Vieira VM, Ryan PB, Detwiler R, Sanders B, Steenland K, Bartell SM. (2011a). Environmental fate and transport modeling for perfluorooctanoic acid emitted from the Washington Works Facility in West Virginia. Environ Sci Technol 45:1435-42

Shin HM, Vieira VM, Ryan PB, Steenland K, Bartell SM. (2011b). Retrospective exposure estimation and predicted versus observed serum perfluorooctanoic acid concentrations for participants in the C8 Health Project. Environ Health Perspect 119:1760-5

Smith Sehdev AE, Hutchins GM. (2001). Problems with proper completion and accuracy of the cause-of-death statement. Arch Intern Med 161:277-84

Spiegelman D. (2010). Approaches to uncertainty in exposure assessment in environmental epidemiology. Annu Rev Public Health 31:149-63

Starks SE, Gerr F, Kamel F, Lynch CF, Jones MP, Alavanja MC, Sandler DP, Hoppin JA. (2012). Neurobehavioral function and organophosphate insecticide use among pesticide applicators in the Agricultural Health Study. Neurotoxicol Teratol 34:168-76

Stayner L, Bailer AJ, Smith R, Gilbert S, Rice F, Kuempel E. (1999). Sources of uncertainty in dose-response modeling of epidemiological data for cancer risk assessment. Ann N Y Acad Sci 895:212-22

Strimbu K, Tavel JA. (2010). What are biomarkers? Curr Opin HIV AIDS 5:463-6

U.S. EPA. (2010). Draft Framework for Incorporating Human Epidemiologic & Incident Data in Health Risk Assessment. Washington, DC: U.S. Environmental Protection Agency (EPA), Office of Pesticide Programs

U.S. EPA. (2015a). Literature Review on Neurodevelopment Effects & FQPA Safety Factor Determination for the Organophosphate Pesticides. Washington, DC: U.S. Environmental Protection Agency (EPA), Office of Pesticide Programs

U.S. EPA. (2011). Revised Human Health Risk Assessment on Chlorpyrifos

U.S. EPA. (2015b). Risk assessment. Available at: http://www2.epa.gov/risk. Last updated: 7 October 2015.

U.S. EPA. (2000). Risk Characterization Handbook. Washington, DC: Science Policy Council, U.S. Environmental Protection Agency (EPA),

VanderWeele TJ, Hernan MA, Robins JM. (2008). Causal directed acyclic graphs and the direction of unmeasured confounding bias. Epidemiology 19:720-8

Viswanathan M, Berkman ND. (2011). Development of the RTI Item Bank on risk of bias and precision of observational studies. Research Triangle Park, NC: RTI International–University of North Carolina Evidence-based Practice Center, 77

Vlaanderen J, Vermeulen R, Heederik D, Kromhout H, Ecnis Integrated Risk Assessment Group EUNOE. (2008). Guidelines to evaluate human observational studies for quantitative risk assessment. Environ Health Perspect 116:1700-5

Wells GA, Shea B, O'Connell D, Peterson J, Welch V, Losos M, Tugwell P. (2014). The Newcastle-Ottawa scale (NOS) for assessing the quality of nonrandomized studies in meta-analysis. Available at: http://www.ohri.ca/programs/clinical_epidemiology/oxford.asp. Last accessed: 20 October 2015

Whyatt RM, Barr DB, Camann DE, Kinney PL, Barr JR, Andrews HF, Hoepner LA, Garfinkel R, Hazi Y, Reyes A, Ramirez J, Cosme Y, Perera FP. (2003). Contemporary-use pesticides in personal air samples during pregnancy and blood samples at delivery among urban minority mothers and newborns. Environ Health Perspect 111:749-56

Whyatt RM, Camann DE, Kinney PL, Reyes A, Ramirez J, Dietrich J, Diaz D, Holmes D, Perera FP. (2002). Residential pesticide use during pregnancy among a cohort of urban minority women. Environ Health Perspect 110:507-14

Whyatt RM, Garfinkel R, Hoepner LA, Andrews H, Holmes D, Williams MK, Reyes A, Diaz D, Perera FP, Camann DE, Barr DB. (2009). A biomarker validation study of prenatal chlorpyrifos exposure within an inner-city cohort during pregnancy. Environ Health Perspect 117:559-67

Whyatt RM, Garfinkel R, Hoepner LA, Holmes D, Borjas M, Williams MK, Reyes A, Rauh V, Perera FP, Camann DE. (2007). Within- and between-home variability in indoor-air insecticide levels during pregnancy among an inner-city cohort from New York City. Environ Health Perspect 115:383-9

Whyatt RM, Rauh V, Barr DB, Camann DE, Andrews HF, Garfinkel R, Hoepner LA, Diaz D, Dietrich J, Reyes A, Tang D, Kinney PL, Perera FP. (2004). Prenatal insecticide exposures and birth weight and length among an urban minority cohort. Environ Health Perspect 112:1125-32