Workshop Report: Evaluation of Epidemiological Data Consistency for Application in Regulatory Risk Assessment[§]

Ronald H. White[*,1], Mary A. Fox[2], Glinda S. Cooper[3], Thomas F. Bateson[3], Thomas A. Burke[2] and Jonathan M. Samet[4]

[1]*R.H. White Consultants, Silver Spring, MD, USA*

[2]*Department of Health Policy and Management, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA*

[3]*National Center for Environmental Assessment, U.S. Environmental Protection Agency, Washington DC, USA*

[4]*Department of Preventive Medicine, University of Southern California, Los Angeles, CA, USA*

Abstract: Epidemiological study results have a key role in the assessment of health risks associated with exposures to chemicals and pollutants, and often serve as the basis for the development of regulatory limits for environmental and occupational health. A key uncertainty in the application of epidemiological study results in risk assessments stems from variability in defining and operationalizing the concept of consistency of findings across studies, with assessments of consistency often a controversial component of risk assessments. Although assessment of consistency of findings across a diverse collection of epidemiological studies is central to evaluating that body of evidence for supporting causal inferences, the variability in definition and formal evaluation methods strongly suggest the need for constructive approaches to consistently and transparently evaluate data consistency.

In response to the need to improve approaches to assessing consistency in epidemiological study results, the Johns Hopkins Risk Sciences and Public Policy Institute organized a workshop held in Baltimore, Maryland in September 2010 to identify and discuss key methodological issues, and to develop recommendations for qualitative and quantitative approaches to addressing those issues. A multi-disciplinary approach was utilized for the workshop, involving invited experts from a variety of fields, and the invited participants were drawn from academia, industry, government, and the public interest sectors. This report provides a summary of selected epidemiology methodological issues discussed by the workshop participants and provides the workshop's key findings and recommendations for future approaches to addressing this issue.

Keywords: Consistency, epidemiology, heterogeneity, regulation, risk assessment, workshop report.

INTRODUCTION

Epidemiological studies play a key role in the assessment of risks associated with exposures to chemicals and pollutants and for development of regulatory standards covering environmental and occupational settings. The strengths and weaknesses of epidemiological methodology, as well as the overall value of the use of epidemiological evidence to support regulatory standards, have been widely discussed in the scientific and public health policy literature (e.g., [1-8]). Issues related to the presentation of epidemiological results that inform risk assessments, the need to apply modern biostatistical techniques to epidemiological data, and methodological challenges in the use of epidemiological data in quantitative risk assessment have also been noted [9-12].

When evaluating epidemiological findings in support of causal inference, a key uncertainty often stems from apparent inconsistency across studies. Evaluations of consistency are often controversial, and contradictory determinations may result from varying stakeholder perspectives. The evaluation of consistency in epidemiological results has been discussed for more than 50 years (e.g., [13-15]), expanding and become more nuanced as the field of epidemiology (and specifically environmental epidemiology) has matured. For example, gender-based differences in susceptibility to a potential endocrine-disrupting chemical can explain differences in observed effects among studies that include varying proportions of males and females – i.e., there may be biological reasons not to expect to see the same effect. Similarly, differences in exposure metrics and the range of exposures could lead to differences in observed estimates among studies. The evaluation of consistency of findings across a diverse collection of epidemiological studies is

[*]Address correspondence to this author at the R.H. White Consultants, LLC, 12900 Tourmaline Terrace, Silver Spring, MD, USA; Tel: (240) 381-4075; Fax: (301) 384-8876; E-mail: ronaldhwhite@comcast.net

[§]The views expressed in this article are those of the authors and do not necessarily represent the views or policies of the U.S. Environmental Protection Agency, the National Institute of Environmental Health Sciences, the National Institute of Occupational Safety and Health, or Health Canada.

1874-2971/13 2013 Bentham Open

central to evaluating that body of evidence for supporting causal inferences for hazard identification, one of the core components of environmental health risk assessment [16]. There remains a need for approaches to objectively and transparently evaluate consistency. This workshop report provides recommendations regarding approaches to assist in the evaluation of consistency in epidemiological study results.

WORKSHOP ON EVALUATING CONSISTENCY IN EPIDEMIOLOGICAL DATA FOR APPLICATION IN REGULATORY RISK ASSESSMENT

The Johns Hopkins Risk Sciences and Public Policy Institute organized a workshop to develop recommendations for qualitative and quantitative approaches to assessing consistency in epidemiological results. The workshop, held in Baltimore, Maryland on September 23-24, 2010, was co-sponsored by the U.S. Environmental Protection Agency (U.S. EPA), the National Institute of Environmental Health Sciences, and the National Institute for Occupational Safety and Health, with additional support provided by Health Canada. A multi-disciplinary approach was utilized, involving invited experts from the fields of epidemiology, risk assessment, exposure assessment, biological sciences, biostatistics, and science policy, drawn from academia, industry, government, and the public interest sectors (Table **1**). Following an opening plenary session in which several individual perspectives on evaluation of consistency of epidemiological results were presented and discussed, the participants were divided into three groups to discuss the key issues identified below. Group deliberations were reported and discussed in a concluding plenary session. This report summarizes discussions of key issues identified for the workshop, and presents the findings and recommendations from the workshop. While there was general consensus among participants regarding the findings and recommendations discussed below, except where explicitly noted, no formal process (e.g., voting) regarding unanimity of views was undertaken.

Workshop Discussions, Findings and Recommendations

General Issues

Consideration of statistical testing, power, precision, and interpretation of study results has been a core part of epidemiology for more than 30 years [17, 18]. There was general agreement that the results of a study should not be characterized on the basis of the presence or absence of statistical significance, and that the consistency of epidemiological study results cannot be assessed by counting the number of studies that have “positive” or “negative” findings. Lack of statistical significance is not synonymous with “no effect”, and it is important to distinguish between studies that demonstrate no effect and studies that would be better described as being “inconclusive” or “uninformative.” For example, a study with a relative risk estimate close to 1.0 with narrow confidence intervals could reasonably be described as showing no effect, but a small study with a similar point estimate but much wider confidence intervals would be better described as being “inconclusive.”

The distinction between heterogeneity and inconsistency within the context of evaluation of study results was another general issue addressed in the workshop. Heterogeneity and inconsistency were considered to be two distinct concepts. Heterogeneity represents variation in observed effects that can be expected based on differences in populations (e.g., different effects of an endocrine-disrupting chemical in men and women) or other study design attributes (e.g., a very weak or null effect seen with an “ever/never” exposure classification compared with a stronger effect seen with a more focused and specific exposure measure). Inconsistency, however, implies unexplained variation that may be reflective of spurious findings. Distinguishing between these concepts is central to an evaluation of the consistency of study results. It was noted that a variety of factors may affect the observed heterogeneity of a group of studies, including: the types and ranges of exposure levels and circumstances; the exposure measurement methods including its accuracy in reflecting exposure during a critical period with respect to the outcome of interest; the length of follow-up for disease incidence or mortality; the extent of misclassification of outcomes due to differences or changes in disease detection or definitions over time; the degree to which results may be influenced by measured and unmeasured confounding, and the statistical power of a study and potential imprecision of effect estimates. Workshop participants considered that only when the effects of these considerations have been evaluated can any remaining differences be interpreted as potentially representing inconsistency.

Specific Issues

Specific issues considered at the workshop with respect to evaluating consistency among epidemiological studies included variation in outcome definition, exposure assessment methodology, the definition and identification of an exposure-response trend, as well as critical periods of exposure and follow-up period, and how these variations should be considered when comparing results among studies. Finally, the workshop also considered approaches for evaluating large bodies of epidemiological evidence with respect to determining consistency of findings. While other issues may also relate to potential sources of epidemiological study result inconsistency, given time constraints the workshop was limited to the topics discussed in more detail below.

Considering Variation in Outcome Definition in Interpreting the Consistency of Results Across Studies

Summary of Issues: 1) How should variation in study findings among potentially related health outcomes be evaluated? 2) How should the quality of the disease definition (i.e., reliability and validity, or refinement by subtype) be considered when evaluating consistency (or heterogeneity) in effect measures among studies?

Some types of diseases and early states of disease in particular may be difficult to define or measure. Some studies may assess functional tests or disease markers, which may or may not be considered adverse outcomes from clinical or public health perspectives. With improved understanding of the etiologic pathways and overall biological basis for disease, epidemiological studies may use upstream markers of the disease process rather than apical endpoints. For example, in assessing the relationship of air pollution exposure to exacerbation of cardiovascular disease

Table 1. State of the Science Workshop: Evaluation of Consistency of Epidemiological Results for Application in Regulatory Risk Assessment, September 23-24, 2010 Baltimore, MD, USA

Participants
Thomas Burke, Ph.D., Johns Hopkins University (co-chair)*
Jonathan Samet, MD, MS, University of Southern California (co-chair)*
Thomas Bateson, Sc.D., MPH, U.S. Environmental Protection Agency*
Aaron Blair, Ph.D., MPH, National Cancer Institute
David Coggon MD, Southampton General Hospital
Glinda Cooper, Ph.D., U.S. Environmental Protection Agency*
Elizabeth Delzell, D.Sc., University of Alabama at Birmingham
Kay Dickersin, Ph.D., Johns Hopkins University
Elizabeth Fontham MPH, Dr.PH, Louisiana State University
Bruce Fowler, Ph.D., Agency for Toxic Substances and Disease Registry, U.S. Centers for Disease Control and Prevention
Mary Fox, Ph.D., Johns Hopkins University*
Freya Kamel, Ph.D., National Institute of Environmental Health Sciences
Ellen Kirrane, Ph.D., U.S. Environmental Protection Agency
Daniel Krewski, MHA, MSc, Ph.D., U of Ottawa
Germaine Buck Louis, Ph.D., National Institute for Child Health and Human Development
Charles Poole, Ph.D., University of North Carolina
Ruthann Rudel, MS, Silent Spring Institute
Jennifer Sass, Ph.D., Natural Resources Defense Council
Cheryl Siegel Scott, MSPH, U.S. Environmental Protection Agency
Jack Siemiatycki, Ph.D., U Montreal
Thomas Smith, Ph.D., MPH, Harvard University
Leslie Stayner, Ph.D., U of Illinois
Patricia Stewart, Ph.D., National Cancer Institute
J. Morel Symons, Ph.D., DuPont Haskell Laboratory for Health and Environmental Sciences
Elizabeth A. Whelan, Ph.D., National Institute of Occupational Safety and Health
Ronald White, MST, Johns Hopkins University*
Michael Wright, Sc.D., MPH U.S. Environmental Protection Agency
Observers
Krista Christensen, Ph.D., MPH, U.S. Environmental Protection Agency
Barbara Glenn, Ph.D., MPH, U.S. Environmental Protection Agency
Karen Hogan, Ph.D., U.S. Environmental Protection Agency
Jennifer Jinot, Ph.D. U.S. Environmental Protection Agency
Patricia Murphy, Ph.D., MPH, U.S. Environmental Protection Agency
Molini Patel, Ph.D., MPH, U.S. Environmental Protection Agency

*Workshop Organizing Committee

and cardiovascular mortality, epidemiological studies have examined endpoints such as heart rate variability, dysrhythmic susceptibility, and cardiac repolarization as well as than the more “downstream” outcomes of cardiovascular-related mortality, incidence of myocardial infarction, or non-fatal cardiovascular-related hospitalizations [19]. The interpretation of consistency across study results is complicated when different findings are seen across a range of outcomes. In some situations, there may be evidence of an abnormality across studies, but there is variation in what specific abnormality is associated with a pollution exposure (even if some of the same tests are used across studies). Are the data consistent because there is evidence of damage across the studies, or inconsistent because the results across a range of related outcomes may differ with respect to magnitude? A further challenge to consistency assessment arises when the definitions or classification criteria for a disease differ across studies or change over time (e.g., become more refined), as has occurred for leukemia [20, 21].

Workshop Discussion, Findings and Recommendations: Workshop participants recognized that the selection of outcome measurements is often driven by feasibility for assessment of population-level health outcomes, as well as the purpose of the study. For example, a study designed to measure reproductive mechanisms of action may use a measure of sperm damage or sperm concentration, but a study focusing on broader population impact issues may use a measure of fertility or time to pregnancy.

In general, workshop participants noted that more information is often needed regarding the sensitivity and specificity of outcome measurements, including biomarker data, and validation information, if available, can offer a basis for comparing outcome assessment methods or outcome scales. However for historical publications such validation information is often not available and professional judgment is required to assess validity.

There was a recognition that larger or more robust relationships may be found with more sensitive "upstream" markers of disease when compared to endpoints reflecting clinical expression of a defined disease state, but that differences between study results examining clinically manifest disease and those based on preclinical disease outcomes may not necessarily indicate inconsistent results. The issue of whether the upstream endpoints represent an adverse health effect or might be a reversible, transient effect was noted, though this issue was not discussed in detail as it was deemed beyond the scope of this workshop.

When considering a series of study results for a given health outcome with several related outcome measurement methods (e.g., lung function or kidney function) or histological sub-types (certain cancers), an important consideration involves the issue of "lumping" together of study results as opposed to "splitting" or stratifying the data. If possible, the decision on how to analyze study results for a specific outcome or subcategory of outcome should depend on the plausible or known biological mechanism of the hazard. A "splitting" approach is based on the assumption that there is a clear delineation between the categories, but this assumption may not hold, particularly under different stages of disease development and progression.

Consideration of Variation in Exposure Measurement in Interpreting the Consistency of Results Across Studies

*Summary of Issue*s: 1) How should differing exposure assessment methods be accounted for in a formal and transparent manner, particularly with respect to effect estimate attenuation that is expected with non-differential misclassification, when evaluating the consistency of study results? 2) What criteria should be applied in selecting specific data points (e.g., exposure groups) for the evaluation of the consistency of data among studies?

Differences in exposure assessment techniques across studies may create heterogeneity in effect estimates. This issue also commonly arises when considering occupational exposure studies, given the variety of exposure measures that can be used, ranging from categories based on job title or employment in a particular plant, to individual measurements reflecting differences in worker tasks, time periods, and location. In addition, an "ever exposed" category disregards ranges of exposure, which may be biologically useful in explaining results.

Workshop Discussion, Findings and Recommendations: The accuracy of the exposure assessment methodology used in an epidemiological study is a key determinant of overall study quality. For most exposures, the "gold standard" of having data across the biologically relevant time window is not achievable and that window may not be known. Exposure measurements are often based on proxies for this gold standard measurement, (e.g., a biomarker may not have been measured concurrent with disease ascertainment period, or ambient pollutant measurements representing most, but not all, sources of exposure) which do not capture individual-level variation in exposure or response. This reliance on proxies introduces uncertainty into the analysis in terms of the extent to which a proxy is a valid substitute for a validated assessment of actual exposure. The participants noted that a variety of other factors, including changes in workplace or environmental standards, changes in economic conditions that impact emissions, and changes in manufacturing processes/controls, can impact exposure levels and therefore can affect the exposure-response relationship over time. Given this context, variation in exposure measurement should not focus only on measurement error, but should also consider the contribution of these other dimensions of exposure assessment to variation in observed results.

A major discussion topic focused on the feasibility of assessing the extent to which exposure assessment methodologies contribute to heterogeneity of results among studies. A variety of approaches, of varying degrees of complexity, could be used. Researchers can incorporate into studies analyses that take into account measurement error in the exposure estimates. When evaluating published studies, other options need to be considered, such as stratifying studies by key characteristics of the exposure assessment approach to evaluate the impact of potential exposure misclassification. Adjustment for the observed attenuation may also be possible, and modeling could assist in determining the extent to which different misclassifications or measurement errors may influence risk estimates.

Another important concept with respect to exposure and variability in results concerns the need to clearly understand the exposure range considered within, and between, studies: different exposure-response relationships can be reasonably expected to be seen among different exposure ranges, particularly if the exposure-response relationship is nonlinear. A stronger effect may be estimated in a study that incorporates a wider range of exposures (and thus a greater contrast between the "exposed" and referent categories) than in a study with a more limited exposure range. Thus, comparison of "high" exposure categories across studies (e.g., a meta-analysis of "high" versus "low" comparisons across studies) can be problematic since the different studies may incorporate different range values for their exposure categories. A more valid comparison may be to use this type of comparison to evaluate whether the "high" versus "low" comparison gives a stronger effect estimate than an "ever" versus "never" comparison among the same set of studies.

Definition and Identification of Trends

Summary of Issues: 1) Should a statistical test be the basis for deciding if a trend is present? If so, what considerations should be used in choosing the test and the level of statistical significance to be used? 2) How can differences among studies in the quality of the exposure assessment be transparently and reasonably incorporated into the evaluation of the presence/strength/shape of the observed exposure-response trend?

The presence of an exposure-response gradient is one consideration within the Hill framework for evaluating causality [15]. If risk increases at higher levels of exposure, alternative explanations other than causality become less tenable. Results from several occupational cohort mortality studies suggest that under certain circumstances, exposure-response function/gradients may be nonlinear [22]. Indeed, Hill specifically notes that a nonlinear exposure response gradient may reflect complexities in the relationship, rather than no relationship [15]. In addition, the observed form of the exposure-response relationship may be affected in complicated ways by exposure measurement error, population selection, and modeling approaches.

Workshop Discussion, Findings and Recommendations: Workshop participants agreed that an expectation of monotonic increasing risk with increasing exposure is a reasonable consideration with respect to assessment of a causal association. They also stressed, however, that the underlying (true) exposure-response curve can have a variety of shapes, even within the general category of a monotonic increasing curve. Each of these curves may make biological sense (e.g., a hockey stick pattern reflecting a threshold type of response, a plateau reflecting saturation of a key metabolic activation step, and a flattening or downturn at high exposure reflecting a significant competing risk). An additional difficulty in interpreting trends in epidemiological studies is that because of sources of bias and error, the observed exposure-response may differ from the underlying exposure-response, and thus the absence of a linear exposure-response within a study is not in itself strong evidence for the absence of a causal association. Last, participants noted that population-level exposure-response relationships may differ substantially from exposure-response relationships in individuals.

Given these issues, participants recommended a variety of approaches to the assessment of trends within a study. These approaches ranged from “describe, don’t test”, to use of formal statistical tests assuming linearity on a particular scale across all exposures, to decomposing curves into linear and nonlinear components. The advantage of statistical tests is that they provide quantitative support to qualitative and subjective descriptions. The advantage of a descriptive approach is that it increases the information provided to the reader, and can be used as a framework to address potential explanations such as the biological understanding of the disease process or potential bias introduced by exposure measurement error. The sparseness of the data should also be considered; it may be more appropriate to say “these data do not provide a basis for describing the exposure-response relationship” than to say “these data indicate there is [or is not] a trend.”

Another issue concerns the comparison of trends across studies. Observed exposure-response patterns can differ among studies, particularly among studies with different exposure ranges. For studies in which the exposure range is relatively narrow, or when the shape of the exposure-response function within a study is relatively flat, a trend in the exposure-response function may only be observed when studies spanning a wider exposure range are combined. Depending on the details of the exposure measures used in the various studies, it may be possible to use meta-analysis and meta-regression approaches to obtain an overall estimate of trend and to understand differences among studies.

Consideration of Varying Lengths of Follow-Up or Exposure Windows in Interpreting the Consistency of Results

Summary of Issues: 1) When two or more analyses of data from the same cohort are available, with different lengths of follow-up, what considerations (i.e., type of disease, mechanism of disease, age-interactions) should be used to determine the most relevant follow-up window? 2) How can differences among studies in the length of follow-up or exposure windows be transparently and reasonably incorporated into the evaluation of consistency of observed effects?

Often, data from occupational (or other) cohorts are analyzed at multiple points during follow-up. There is a potential for risk estimates to vary over follow-up, reflecting changing patterns of exposure and underlying exposure-response time dynamics, with effects that are seen earlier not observed later, or effects only emerging after the passage of a greater period of time [23, 24]. Trends may be explored in one or more time dimensions: time since follow-up began, time since exposure, chronological age, and calendar time. Risks might plausibly vary across each of these scales and such variation might be relevant in the development of models for exposure-response relationships. An example of the complexity that can occur with time-related measures can be seen in the analysis of radon-induced lung cancer. The National Academy of Sciences Biological Effects of Ionizing Radiation (BEIR) VI Committee had access to a large and rich data set created by merging data from 11 cohorts of underground miners [25], allowing for analysis of time-varying risk estimates that varied with age, time since exposure, and exposure.

Workshop Discussion, Findings and Recommendations: Workshop participants agreed that trends may plausibly change with time without detracting from the validity of an exposure-response gradient observed at one time. When analyzing an exposure-response relationship over time, variability in relative risk estimates due to a dilution effect from increased person-years observed or from depletion of the susceptible population or from the biological mechanisms underlying disease production, may be seen with increasing lengths of follow-up.

There was considerable discussion concerning the ways in which different types of biological mechanisms would result in different observed effects in situations with different lengths of follow-up. An understanding of the underlying biological mechanisms involved in the specific exposure-disease relationship under study could assist in explaining trends over time. This information may not be available, however. Participants also noted that epidemiological observations can be a source of important insights into the nature of the underlying biological mechanisms involved in disease pathogenesis, so that there may be circularity in looking to a mechanistic framework as a basis for interpreting epidemiological results.

Substantial datasets may be needed for analyzing time-varying exposure-response relationships (see BEIR VI example above). In systematically evaluating exposure-response relationships for different follow-up periods across studies, it would be important to distinguish the study follow-up period from the outcome latency period. Sometimes necessary time-related details are lacking; publications may not adequately document key information on length of follow-up, exposure windows and related changes over time, particularly when reporting on later follow-up periods.

Workshop participants also noted a need for further exploration of key concepts related to length of follow-up through case examples based on existing literature or through development of simulation modeling approaches. Example issues include the effect of depletion of susceptible populations on effect estimates over time, and the effect of a specific form of time-varying exposure-response on the observed results under different lengths of follow-up. Participants noted that consideration should be given to the uncertainty associated with such estimates and to the utilization of time-dependent models.

Approaches to Evaluating Large Bodies of Epidemiological Studies in Interpreting the Consistency of Results

Summary of Issues: 1) What criteria could be applied in selecting studies for inclusion and for selecting specific data points (e.g., subgroups or exposure groups) in assessments of consistency of epidemiological results? 2) How should factors such as variation in study design, study population, differing exposures to pollutant mixtures (ambient and occupational exposures), and mode(s) of action information be considered?

One approach for summarizing large amounts of information for a causal assessment is referred to as a weight of the evidence approach. This approach considers results across the available studies, but gives greater "weight" to those considered to have the greatest reliability and validity [26, 27]. Formal meta-analysis with weighting of studies by size (i.e., inverse of study variance) could be considered an example of a weight of evidence approach, but this type of quantitative summary is not a necessary component of this approach. A second approach for summarizing information selects a relatively limited number of studies for inclusion in the review and/or causal assessment based on the quality of the study; these high quality studies are sometimes referred to as "informative studies".

The application of quality criteria has historically been utilized in systematic reviews of clinical questions [28], and has been less commonly used in evaluation of risk of environmental exposures. Defining, *a priori*, criteria for a "good" study, as well as possibly weighting those criteria, can be challenging. The use of quality criteria and a scoring framework in Turner *et al.* [29] and Wigle *et al.* [30] for assessment of epidemiological study data were discussed as case study examples, although it was noted that there was a movement away from use of quantitative scoring based on qualitative criteria in systematic reviews of clinical trials [31]. In practice, a single study rarely fulfills all of the chosen criteria, and it can be difficult to distinguish the failure of a study to fulfill a specific criterion from the failure of a report to provide enough details to allow the correct scoring of a specific criterion.

Workshop Discussion, Findings and Recommendations: Participants indicated a preference for an inclusive approach to study selection for use in assessing relatively large bodies of studies, rather than excluding certain studies. Concern was raised that exclusion of studies could be perceived as manipulation of the evidence to favor a particular hypothesis or position. Consideration of study quality should be part of the review process, but studies of all designs should be included since each study type has its own set of strengths and weaknesses. Including different study types may assist in balancing potential limitations; stratification may also be used as a way to assess the influence of methodological differences on study results.

Workshop participants discussed the potential application of study quality criteria in the context of qualitative systemic reviews, as well as the basis for quality criteria scoring. Participants supported the use of qualitative weights reflecting study design features (e.g., subject selection criteria, exposure assessment methods, statistical analysis approaches) that are methodologically more accepted or validated. Stratification based on these types of details may be more useful than a "scoring" system for the purpose of examination of consistency and evaluation of sources of heterogeneity in results. Participants noted that it is critical to document and communicate the criteria used to weigh studies, and to provide justification for the criteria. Emphasis should be placed on core principles applied with best judgment to the studies, rather than on set rules designed to apply to all situations. These core principles address outcome ascertainment, exposure measures, and other sources of bias. It was noted that some studies may be informative for hazard assessment but may not provide the basis for development of exposure-response relationships for use in quantitative risk assessment.

Summary of Workshop Discussions

There were some common themes across the discussion of issues relating to evaluation of consistency of results of epidemiologic studies. These themes relate to tools and approaches for assessment of consistency, need for improvements in exposure assessment and data reporting, and consideration of biologically-based frameworks for the assessment of study results.

Qualitative and Quantitative Analytic Tools and Approaches for Assessment of Consistency

Workshop participants discussed the potential utility of a variety of tools and approaches for application in assessing the consistency of epidemiological results. Qualitative approaches include analysis of the features of individual studies as well as systematic across studies. The qualitative approaches build on the framework described in the previous section. Specific questions that could be considered include the following study attributes study attributes:

a) What is the health outcome under investigation? Are there differences in case definition, case mix or type of outcome data (incidence *vs* mortality; preclinical *vs* clinical disease state) that might affect risk estimates compared with other studies?

b) What are the exposures of the groups under comparison (in term of routes, levels and timing), what exposure monitoring data are available, how is exposure assigned to subjects, and how would sources of bias associated with exposure assignment affect observed risk estimates in comparison with other studies?

c) What potential confounders may be important, how well have these been taken into account by study design or statistical analysis and to what extent might residual confounding result in over- or underestimation of risk? The focus of this evaluation should be on known causes of the disease (especially those carrying high relative risks) and factors that are strongly associated with the exposure under study (might these factors be causes of the disease?).

d) What other potential major sources of bias might have caused the risk estimate to be over- or underestimated, and to what extent?

Considering the above attributes, how different are the risk estimates among the studies? The 2006 U.S. EPA Criteria Document for Ozone [32], for example, notes that "consideration of consistency and heterogeneity of effects are appropriately understood as an evaluation of the similarity or general concordance of results, rather than an expectation of finding quantitative results within a very narrow range." To what extent could differences be attributable to differences in identified effect modifiers (i.e., heterogeneity of effects due to population differences)? Is there a level of heterogeneity in risk estimates that is unlikely to be attributed to differences in study design or population characteristics/subject selection (i.e., true inconsistency)? These are the types of questions that should be considered in an evaluation that attempts to distinguish between heterogeneity and inconsistency of study results.

Quantitative tools include meta-analysis, pooled data analysis, meta-regression, trend tests, and quality criteria scoring. Meta-analysis (in conjunction with a systematic review) and pooled analysis can be useful for deriving summary risk estimates from multiple studies that are considered or can be made substantially similar (homogeneous) [33]. Stratified meta-analysis is used when relevant studies are heterogeneous and obtaining an overall summary estimate is not advised. Stratified meta-analysis further subdivides the study set into homogeneous strata before estimating risk by strata and represents an approach to describing various sources of heterogeneity that may be useful to inform a risk assessment. Meta-regression modeling is another statistical tool that can be used to explore contributors to heterogeneity in terms of study-level covariates [34]. In addition to supporting evaluation of consistency, these tools can address a potential limitation of dose-response assessment, namely selection of a single study for derivation of reference values or cancer slope factors. While workshop participants recommended conducting empirical research to evaluate the utility of the recommended tools, a key issue that was not addressed in detail at this workshop was the feasibility of these types of analyses within the context of risk assessment, and the types of situations in which this effort would be useful.

Improvements in Exposure Assessment and Data Reporting

As evident from the preceding discussion of the importance of exposure measurement issues, the need for continued improvement in both the quantity and quality of exposure measurement data available in occupational and environmental epidemiological studies was highlighted throughout the workshop. The need for improved epidemiological data in general, and particularly for studies with sufficient detail to evaluate consistency, was also noted. One frustration voiced by some workshop participants was that relevant information, such as quantitative exposure measures, stratified analyses, or adequate detail on follow-up procedures for cohort studies, often is missing from peer-reviewed articles. As the use of these details in analysis of consistency of study results becomes more common and the value of this type of analysis grows, these details should become part of the standard practice for reporting.

Biologically-Based Assessment of Study Results

A major area of discussion focused on the developing need for a biologically-based approach to interpreting differences across study results as well as in assessing the quality of study designs and analyses. The participants emphasized the value of understanding the biology of the disease development and progression process in the selection of exposure metrics, identification of related health outcomes, and interpretation of exposure-response relationships. It was also noted that understanding of the biological basis for diseases is evolving and that epidemiologic studies can contribute to this understanding. Such a biologically-based approach requires information from a variety of scientific and medical disciplines, and therefore the concept of utilizing multi-disciplinary teams (including exposure scientists, epidemiologists, clinicians, and others as appropriate) can provide insights in factors that may contribute to "heterogeneity" and "inconsistency" in epidemiological evidence.

CONCLUSION

In summary, it is important to determine how "inconsistency" is defined in practice. Heterogeneity in epidemiological study results can be expected in many situations due to the factors discussed above as well as other methodological issues. Though the direction of effects estimates is important, rather than focusing on a binary assessment of whether data are consistent or inconsistent it may be more useful to focus on the extent, sources, and interpretation of heterogeneity.

CONFLICT OF INTEREST

The author confirms that this article content has no conflicts of interest.

ACKNOWLEDGEMENTS

Primary funding for the workshop and preparation of this article was provided by the U.S. Environmental Protection Agency. Additional funding for the workshop was provided by the National Institute of Environmental Health Sciences (U.S.A.), the National Institute for Occupational Safety and Health (U.S.A.), and Health Canada. The authors wish to acknowledge the contribution of James J. Collins, Ph.D.,

Dow Chemical Co., to the concept and organization of the workshop.

REFERENCES

[1] Whittemore AS. Epidemiology in Risk Assessment for Regulatory Policy. J Chron Dis 1986; 39(12): 1157-68.

[2] Gordis L, Ed. Epidemiology and Health Risk Assessment. New York: Oxford University Press 1988.

[3] Hertz-Picciotto I. Epidemiology and quantitative risk assessment: a bridge from science to policy. Am J Public Health 1995; 85(4): 484-91.

[4] Burke, TA. The proper role of epidemiology in regulatory risk assessment: a regulators perspective. In: The Role of Epidemiology in Regulatory Risk Assessment, Graham JD, Ed. New York: Elsevier Publishers Inc. 1995.

[5] Graham JD, Paustenbach DJ, Butler WJ. Epidemiology and risk assessment: Divorce or marriage? In: The Role of Epidemiology in Regulatory Risk Assessment, Graham, JD, Ed. New York: Elsevier Publishers Inc. 1995.

[6] Gamble JF, Lewis RJ. Health and respirable particulate (PM10) air pollution: a causal or statistical association? Environ Health Perspect 1996; 104(8): 838-50.

[7] Samet JM, Schnatter R, Gibb H. Epidemiology and risk assessment. Am J Epidemiol 1998; 148(10): 929-36.

[8] Nachman KE, Fox MA, Sheehan MC, Burke TA, Rodricks JV, Woodruff TJ. Leveraging epidemiology to improve risk assessment. Open Epidemiol J 2011; 4: 3-29.

[9] Nurminen M, Nurminen T, Corvalán CF. Methodologic issues in epidemiologic risk assessment. Epidemiology 1999; 10: 585-93.

[10] Stayner L, Smith RJ, Gilbert S, Bailer AJ. Epidemiological approaches to risk assessment. Inhal Tox 1999; 11: 593-601.

[11] Schwartz J. The Use of Epidemiology in environmental risk assessment. Hum Ecol Risk Assess 2002; 8(6): 1253-65.

[12] Ryan L. Epidemiologically based environmental risk assessment. Stat Sci 2003 18(4): 466-80.

[13] Cornfield J, Haenszel W, Hammond EC, Lilienfeld AM, Shimkin MB, Wynder EL. Smoking and lung cancer: recent evidence and a discussion of some questions. JNCI 1959; 22: 173-203.

[14] Office of the Surgeon General of the United States. Report of the Surgeon General's Advisory Committee on Smoking and Health. 1964. Smoking and Health, Public Health Service Publication No. 1103, Washington, DC. Available from: http://www.surgeongeneral.gov/library/reports/index.html [accessed July 17, 2012].

[15] Hill AB. The environment and disease: association or causation. Proc R Soc Med 1965; 58: 285-300.

[16] National Research Council Committee on the Institutional Means for Assessment of Risks to Public Health. Risk Assessment in the Federal Government: Managing the Process. Washington, D.C.: National Academies Press 1983.

[17] Rothman KJ. A show of confidence. N Engl J Med 1978; 299(24): 1362-63.

[18] Rothman KJKJ. Curbing type I and type II errors. Eur J Epidemiol 2010; 25(4): 223-24.

[19] U.S. Environmental Protection Agency. Integrated science assessment for particulate matter (final report) EPA/600/R-08/139F. Washington, DC 2009.

[20] World Health Organization. Classification of tumours of haematopoietic and lymphoid tissues. IARC Press: Lyon 2008.

[21] American Society of Clinical Oncology. Leukemia –Acute Myeloid. Available from: http://www.cancer.net/patient/Cancer+Types/Leukemia+-+Acute+Myeloid+-+AML?sectionTitle=Subtypes [accessed July 18, 2012].

[22] Stayner L. Attenuation of exposure-response curves in occupational cohort studies at high exposure levels. Scand J Work Environ Health 2003; 29: 317-24.

[23] Beane Freeman LE, Blair A, Lubin JH, *et al*. Mortality from lymphohematopoietic malignancies among workers in formaldehyde industries: the National Cancer Institute Cohort. J Natl Cancer Inst 2009; 101: 751-61.

[24] Mundt KA, Dell LD, Austin RP, Luippold RS, Noess R, Bigelow C. Historical cohort study of 10,109 men in the North American vinyl chloride industry, 1942-72: update of cancer mortality to 31 December 1995. J Occup Environ Med 2000; 57: 774-81.

[25] National Research Council. Health Effects of Exposure to Radon: BEIR VI. Washington, DC: National Academies Press 1999.

[26] U.S. Environmental Protection Agency. Guidelines for carcinogen risk assessment EPA/630/P-03/001B. Washington DC 2005.

[27] International Agency for Research on Cancer. IARC Monographs on the Evaluation of Carcinogenic Risks to Humans: Preamble. 2006. Available from: http://monographs.iarc.fr/ENG/Preamble/CurrentPreamble.pdf [accessed July 17, 2012].

[28] Cochrane Collaboration. Evidence-based health care and systematic reviews. Available from: http://www.cochrane.org/about-us/evidence-based-health-care [accessed July 18, 2012].

[29] Turner MC, Wigle DT, Krewski D. Residential pesticides and childhood leukemia: A systematic review and metaanalysis. Environ Health Perspect 2010; 118: 33-41.

[30] Wigle DT, Turner MC, Krewski D. A systematic review and meta-analysis of childhood leukemia and parental occupational pesticide exposure. Environ Health Perspect 2009; 117: 1505-13.

[31] Whiting P, Harbord R, Kleijnen J. No role for quality scores in systematic reviews of diagnostic accuracy studies. BMC Med Res Methodol 2005; 5:19.

[32] U.S. Environmental Protection Agency. Air Quality Criteria for Ozone and Related Photochemical Oxidants: Volume I of III EPA/600/R-05/004aF. Washington, DC 2006.

[33] Gordon I, Boffetta P, Demers P. A Case study comparing a meta-analysis and a pooled analysis of studies of sinonasal cancer among wood workers. Epidemiology 1998; 9: 518-524.

[34] Morton SC, Adams JL, Suttorp MJ, Shekelle PG. Meta-regression Approaches: What, Why, When, and How? Technical Review 8, AHRQ Publication No. 04-0033. Rockville, MD: Agency for Healthcare Research and Quality 2004.

Received: August 29, 2012 Revised: November 10, 2012 Accepted: November 10, 2012