UNITED STATES ENVIRONMENTAL PROTECTION AGENCY
WASHINGTON, D.C. 20460

OFFICE OF
PREVENTION, PESTICIDES, AND
TOXIC SUBSTANCES

April 22, 2010

MEMORANDUM

SUBJECT: Transmittal of Meeting Minutes of the FIFRA Scientific Advisory Panel Meeting on the 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.

TO: Steven Bradbury, Ph.D.
Acting Director
Office of Pesticide Programs

FROM: Myrta R. Christian
Designated Federal Official
FIFRA Scientific Advisory Panel
Office of Science Coordination and Policy

THRU: Laura Bailey
Executive Secretary
FIFRA Scientific Advisory Panel
Office of Science Coordination and Policy

Frank Sanders
Director
Office of Science Coordination and Policy

Please find attached to this memorandum the meeting minutes of the FIFRA Scientific Advisory Panel open meeting held in Arlington, Virginia on February 2 – 4, 2010. This report addresses a set of scientific issues being considered by the Environmental Protection Agency pertaining to the "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."

Attachment

cc:

Steve Owens
James J. Jones
Steven Bradbury
Vickie Dellarco
William Jordan
Donald Brady
William Diamond
Margie Fehrenbach
Joan Harrigan-Farrelly
Jack Housenger
Richard Keigwin
Steve Knizner
Tina Levine
Lois Rossi
Betsy Shaw
Vanessa Vu (SAB)
Enesta Jones
Douglas Parsons
Anna Lowitt
OPP Docket

FIFRA Scientific Advisory Panel Members

Steven G. Heeringa, Ph.D. (FIFRA SAP Chair)
John R. Bucher, Ph.D., D.A.B.T.
Janice E. Chambers, Ph.D., D.A.B.T., Fellow A.T.S.
Gerald A. LeBlanc, Ph.D.
Carey N. Pope, Ph.D.
Kenneth M. Portier, Ph.D.

FQPA Science Review Board Members

John C. Bailar, III, M.D., Ph.D.
Frank J. Bove, Sc.D
Richard Greenwood, Ph.D.
Ellen B. Gold, Ph.D.
Shelley A. Harris, Ph.D.
William L. Hayton, Ph.D.
Chensheng Lu, Ph.D.
Bette Meek, Ph.D.
Nu-may Ruby Reed, Ph.D., D.A.B.T.
John S. Reif, D.V.M., M.Sc. (med)

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

NOTICE

These meeting minutes have been written as part of the activities of the Federal Insecticide, Fungicide, and Rodenticide Act (FIFRA) Scientific Advisory Panel (SAP). The meeting minutes represent the views and recommendations of the FIFRA SAP, not the United States Environmental Protection Agency (Agency). The content of the meeting minutes does not represent information approved or disseminated by the Agency. The meeting minutes have not been reviewed for approval by the Agency and, hence, the contents of these meeting minutes do not necessarily represent the views and policies of the Agency, nor of other agencies in the Executive Branch of the Federal government, nor does mention of trade names or commercial products constitute a recommendation for use.

The FIFRA SAP is a Federal advisory committee operating in accordance with the Federal Advisory Committee Act and established under the provisions of FIFRA as amended by the Food Quality Protection Act (FQPA) of 1996. The FIFRA SAP provides advice, information, and recommendations to the Agency Administrator on pesticides and pesticide-related issues regarding the impact of regulatory actions on health and the environment. The Panel serves as the primary scientific peer review mechanism of the EPA, Office of Pesticide Programs (OPP), and is structured to provide balanced expert assessment of pesticide and pesticide-related matters facing the Agency. FQPA Science Review Board members serve the FIFRA SAP on an *ad hoc* basis to assist in reviews conducted by the FIFRA SAP. Further information about FIFRA SAP reports and activities can be obtained from its website at http://www.epa.gov/scipoly/sap/ or the OPP Docket at (703) 305-5805. Interested persons are invited to contact Myrta R. Christian, SAP Designated Federal Official, via e-mail at christian.myrta@epa.gov.

In preparing the meeting minutes, the Panel carefully considered all information provided and presented by EPA, as well as information presented by public commenters. This document addresses the information provided and presented by these groups within the structure of the charge.

TABLE OF CONTENTS

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

Steven G. Heeringa, Ph.D.
FIFRA SAP Chair
FIFRA Scientific Advisory Panel
Date: April 22, 2010

Myrta R. Christian, M.S
Designated Federal Official
FIFRA Scientific Advisory Panel
Date: April 22,2010

Federal Insecticide, Fungicide, and Rodenticide Act
Scientific Advisory Panel Meeting
February 2 – 4, 2010

PARTICIPANTS

FIFRA SAP Chair

Steven G. Heeringa, Ph.D., Research Scientist & Director for Statistical Design, University of Michigan, Institute for Social Research, Ann Arbor, MI

Designated Federal Official

Myrta R. Christian, M.S., FIFRA Scientific Advisory Panel, Office of Science Coordination and Policy, EPA

FIFRA Scientific Advisory Panel Members

John R. Bucher, Ph.D., D.A.B.T., Associate Director, National Toxicology Program, NIEHS, Research Triangle Park, NC

Janice E. Chambers, Ph.D., D.A.B.T., Fellow A.T.S., William L. Giles Distinguished Professor & Director, Center for Environmental Health Sciences, College of Veterinary Medicine, Mississippi State University, Mississippi State, MS

Gerald A. LeBlanc, Ph.D., Professor and Department Head, Department of Environmental & Molecular Toxicology, North Carolina State University, Raleigh, NC

Kenneth M. Portier, Ph.D., Program Director, Statistics, American Cancer Society, National Home Office, Atlanta, GA

Carey N. Pope, Ph.D., Professor, Head & Sitlington Chair of Toxicology, Department of Physiological Sciences, Oklahoma State University College of Veterinary Medicine, Stillwater, OK

FQPA Science Review Board Members

John C. Bailar, III, M.D., Ph.D., Scholar in Residence, The National Academies, Washington, DC, Professor Emeritus, University of Chicago

Frank J. Bove, Sc.D, Senior Epidemiologist, Division of Health Studies, Agency for Toxic Substances and Disease Registry, Public Health Service, Atlanta, GA

Richard Greenwood, Ph.D., Professor of Environmental Science, School of Biological Sciences, University of Portsmouth, Portsmouth, United Kingdom

Ellen B. Gold, Ph.D., Professor and Chair, Department of Public Health Sciences, School of Medicine, University of California – Davis, Davis, CA

Shelley A. Harris, Ph.D., Scientist, Population Studies and Surveillance, Cancer Care Ontario, Scientist, Occupational Cancer Research Centre, Associate Professor, Dalla Lana School of Public Health, University of Toronto, Toronto, Ontario, Canada

William L. Hayton, Ph.D., Professor, Division of Pharmaceutics, The Ohio State University, Columbus, OH

Chensheng Lu, Ph.D., Mark and Catherine Winkler Assistant Professor of Environmental Exposure Biology, Department of Environmental Health, Harvard School of Public Health Boston, MA

Bette Meek, Ph.D., Associate Director, Chemical Risk Assessment, McLaughlin Centre for Population Health, University of Ottawa, Ottawa, Ontario, Canada

Nu-may Ruby Reed, Ph.D., D.A.B.T., Staff Toxicologist, Department of Pesticide Regulation, Health Assessment Section, Medical Toxicology Branch, California EPA, Sacramento, CA

John S. Reif, D.V.M., M.Sc. (med), Professor and Head, Section of Epidemiology, Department of Environmental Health, Colorado State University, Fort Collins, CO

INTRODUCTION

The Federal Insecticide, Fungicide, and Rodenticide Act (FIFRA) Scientific Advisory Panel (SAP) has completed its review of Scientific Issues Associated with the "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." Advance notice of the meeting was published in the *Federal Register* on November 18, 2009. The review was conducted in an open Panel meeting held in Arlington, Virginia, from February 2 – 4, 2010. Dr. Steven G. Heeringa chaired the meeting. Myrta R. Christian served as the Designated Federal Official.

Data from epidemiology studies and human incident reports contain valuable information, and contributes to a weight of evidence analysis in the characterization of human exposure, response to pesticides, and human health risks. Epidemiology and incident data do, however, pose challenges with respect to characterizing human health risks. EPA convened this meeting of the FIFRA Scientific Advisory Panel (SAP) to discuss science issues related to using epidemiology and human incident data in human health risk assessment. The Office of Pesticide Programs (OPP) solicited comment on a draft framework for implementing the use of such data into human health risk assessment in conjunction with several case studies. OPP's draft framework prescribes a weight of the evidence approach using the best available science for mode of action, exposure, pharmacokinetics, animal and human data determined by in vivo or in vitro studies, and physiologically-based pharmacokinetic models. Three case studies evaluated by the SAP were intended to illustrate the draft framework and to highlight key science challenges with incorporating epidemiology or human incident data into a risk assessment. The Agency solicited comment on the weight of the evidence approach for evaluating and integrating the exposure, laboratory animal data, and human incident information.

One case study presented an evaluation of several ecological and retrospective cohort epidemiology studies for atrazine. OPP, in collaboration with EPA's Offices of Water and Research and Development (ORD), solicited comment on the strengths and weaknesses of these types of epidemiology studies, and sought advice on the appropriate use of such studies in the atrazine human health risk assessment. This case study is also the first step in EPA's atrazine science re-evaluation plan as described previously at the November 3, 2009 FIFRA SAP (http://www.epa.gov/scipoly/sap/meetings/2009/110309meeting.html).

A second case study illustrated the analysis of reported human incident cases. This case study was based on diazinon, a pesticide used historically in residential settings.

A third case study represented the collaborative work by scientists from OPP, ORD, the National Cancer Institute (NCI), and the National Institute of Environmental Health Sciences (NIEHS) on the Agricultural Health Study (AHS). This case study compared the exposure algorithms used by OPP and the AHS, and considered the temporal relationships for multi-chemical exposure in the AHS. The purpose of the comparison of the OPP and AHS exposure algorithms is to better understand the differences and similarities in how the two approaches estimate worker exposure. Temporal relationships for multi-chemical exposure in the AHS involved the timing and combined uses of pesticides, with particular emphasis on pesticides sharing common modes of

action. The Agency solicited advice on the types of evaluations conducted to date and those being proposed with the third case study.

The FIFRA SAP will advise the Agency on approaches for integrating diverse types of experimental toxicology and epidemiology data. The SAP input will be considered for characterizing atrazine's human health risks to be presented to the SAP in September 2010.

Drs. Steven Bradbury, Acting Director, Office of Pesticide Programs, EPA, and Tina Levine, Director, Health Effects Division, OPP, provided opening remarks at the meeting.

The agenda for this SAP meeting included presentations from the Health Effects Division in the OPP, the National Cancer Institute and public comments.

PUBLIC COMMENTERS

Written statements were provided by:

Mark Schultz of Land Stewardship Project, Kathryn Gilje of Pesticide Action Network, Lorette Picciano of Rural Coalition/Coalición Rural and members of other organizations
Kenneth Racke, on behalf of Dow AgroSciences
Dan Campbell, on behalf of Syngenta Crop Protection, Inc
Erik Janus, on behalf of CropLife America
Jere White on behalf of Kansas Corn Growers Association and other organizations
Michele Marcus, Ph.D., Emory University
Lisa Kelley, on behalf of National Council of Farmer Cooperatives
Wayne Clifford, Washington State Department of Health

Oral statements were presented by:

Jennifer Sass, Ph.D., on behalf of Natural Resources Defense Council
Gerard Swaen, Ph.D., on behalf of Dow AgroSciences
Erik R. Janus, M.S., on behalf of CropLife America
Dominik Alexander, Ph.D., on behalf of Exponent Inc.
James Swenberg, Ph.D., Noel Weiss, Ph.D., Sir Colin Berry, Tim Pastoor, Ph.D., Charles Breckenridge, Ph.D., on behalf of Syngenta Crop Protection
Mr. Tyler Wegmeyer, on behalf of the American Farm Bureau Federation®
Mr. Scott Slaughter, on behalf of the Center for Regulatory Effectiveness
Jessica Johnson Bennett, on behalf of the National Corn Growers Association
Gary J. Burin, Ph.D., on behalf of The Triazine Network
R.L. Silken Jr., Sielken Associates, on behalf of Syngenta

SUMMARY OF PANEL DISCUSSION AND RECOMMENDATIONS

This report of the FIFRA Scientific Advisory Panel (SAP) addresses several key issues relating to the integration of data from epidemiology studies and human incident reports into the human health risk assessment process. The report is based on the draft framework prepared by the Office of Pesticide Programs (OPP) and responses to specific questions framed by the Agency. The draft framework included three case studies selected to identify key issues and scientific challenges in incorporating the results of epidemiology studies and human incident data for pesticide exposure characterization into the risk assessment process.

The Panel commends EPA for its effort to incorporate human data in human health risk assessments in a transparent manner, based on the draft framework. Overall, the Panel was impressed with the documentation presented in the draft framework and the oral presentations by OPP staff that complemented the written information. The Panel commended OPP staff for its work in developing the highly detailed draft framework. Because of the extensive information presented in the draft framework and the clarity of the charge questions, the Panel was able to conduct indepth discussions and provide EPA with specific recommendations for difficult scientific challenges.

Much work remains to be done before the changes in toxicity testing proposed by National Research Council of the National Academy of Science can be implemented. During the transition period from the current practice to that of measuring perturbations of molecular pathways due to exposure to environmental agents, it will be necessary to evaluate all available information. Epidemiologic studies have the potential to inform both the experimental toxicologist, and the regulatory manager of possible sources of harm in human populations. However, like all information considered in risk assessments, the quality and reliability of the information provided by epidemiologic studies needs to be closely scrutinized. This SAP report is intended to provide specific guidance to OPP with respect to incorporation of epidemiologic data into risk assessment.

The Panel recommended that OPP conduct a broader analysis and revision of the framework regarding the extent and nature of reliance on epidemiologic data in risk assessment to include a range of environmental chemicals including arsenic, disinfection byproducts, benzene and solvents such as trichloroethylene and perchloroethylene in addition to those described in the draft framework and associated case studies. Review of the nature and extent of reliance upon human data in risk assessments more broadly within the Agency and as a basis for development of relevant guidance in other regulatory programs is encouraged. Acquisition of epidemiologic expertise for the review and interpretation of human data, including assessment of exposure in relevant studies is also strongly recommended.

Early recognition of the likely contribution of human data in scoping/problem formulation will increase transparency, facilitate peer engagement and conserve resources. For example, lack of adequate characterization of exposure-response relationships in epidemiological studies may preclude the need to do an extensive weight of evidence analysis for these data, since they cannot contribute significantly to dose-response analysis and hence, risk characterization.

The value of framework analysis in coordinating assessment and research has not been emphasized fully in the documentation. For example, there is repeated reference to problem formulation in the draft framework without indication of how the toxicological and epidemiological databases might be considered in an integrated fashion to identify uncertainties and critical data gaps. This would be an appropriate way to identify limitations of available human data to focus additional research. Targeted additional data might include, for example, *in vitro* studies in human tissues or cell lines and focused epidemiological studies to address specific questions in potentially susceptible subgroups by evaluation of early biomarkers of effect.

Based on experience in studies of toxicity of industrial chemicals and environmental contaminants, weighting of different types of human data varies and depends on the nature of the studies and on the results (i.e., the weight of evidence of an effect in humans). Where sufficient weight of evidence for causality of an adverse outcome in humans exists based on robust epidemiologic study designs and data with well-characterized exposures, these data would be preferred to those obtained from laboratory animals for dose-response characterization. Results of epidemiologic studies in which exposure was well-characterized for a relevant effect in humans could be used as a basis to "bound" dose-response estimates from animal studies.

However, incorporation of data from epidemiologic studies into risk assessment clearly poses challenges to assure that the quality of the data is adequate for risk assessment. Therefore, particular attention must be paid to the quality of epidemiologic studies. Relevant considerations include, but are not limited to: 1) the quality of the exposure assessment, including validation measures if available; 2) sample size and statistical power to assure that a meaningful effect could be detected with reasonable probability if one exists; 3) careful definition of the outcome and assessment of the accuracy of classifying persons as diseased or not diseased; 4) attention to possible sources of bias including selection bias, information bias and confounding; 5) adequate consideration of and control for confounding and identification of effect modifying factors; and 6) external validity or the potential for the study to be generalized to other populations. As discussed further in the body of the report, the Panel recommends that OPP consider the full range of epidemiologic study designs that are used for hypothesis testing and not limit consideration to prospective cohort studies. Historical cohort studies, case-control studies, cross-sectional studies and hybrid designs all have potential to be used in the weight of evidence approach in a quantitative manner when they meet the requisite criteria, particularly for exposure assessment.

The Agency proposes a weight of the evidence approach (WOE) for evaluating human and experimental animal data. The approach described in the draft framework incorporates the modified Bradford Hill criteria as a tool for integrating a variety of types of data, including human epidemiologic data, in risk assessment. The Panel recommends that the modified Bradford Hill criteria are highly appropriate for framing the likelihood of a specified consequence of exposure to a particular chemical in humans. These criteria are well-accepted for assessing evidence of causation in the field of epidemiology and in public health. It is important, however, that the criteria not be viewed as a checklist, but rather as characteristics that collectively provide a systematic way to evaluate the evidence, aggregate observations and guide assessments and conclusions. They require flexible interpretation. Particular attention should be

focused on the criteria of strength of association, dose-response, temporal relation, consistency of findings and "biological plausibility", but with the realization that the biology often has not been established when the epidemiologic evidence becomes available. The draft framework should include a clear discussion of the meaning of biologic plausibility because several interpretations are possible. Emphasis should also be placed on evidence of dose-response relationships in epidemiologic studies, as well as on the temporal sequence of events, the strength of the associations and the consistency of findings across studies and populations.

The Panel commends the Agency for their efforts in developing the draft framework for integration of animal and human data. Future application of the framework for evaluation of weight of evidence has substantial potential for improving risk assessment and for the transparent and appropriate integration of human and toxicological data. The use of the "source to adverse outcome pathway" can also be important in identifying critical data gaps and as a basis to focus additional research. The approach moves the focus in toxicology and risk assessment from late adverse effects to earlier biomarkers of exposure and effect, so that more informative human data at relevant dose levels can be collected. Further, the framework is helpful in directing attention to dose-response relationships for early key events in the assessment of available data; it is these dose-response relationships that are critical in the subsequent risk characterization. However, for epidemiologic data to be most useful in risk characterization, exposure must be robustly and quantitatively addressed, preferably with inclusion of appropriate biomarkers of exposure and effect based on identification of key events in a mode of action context.

The draft framework provides a description of the strengths and limitations of several epidemiologic study designs. The Panel recommends that the framework clearly separate ecologic studies from other designs, since their inherent limitations and inability to estimate risk at the level of the individual render them inherently weak for quantitative purposes. Further, the Panel recommends that the term "retrospective studies" as used in the Agency report be replaced with a more complete and accurate description of the study designs that are used in epidemiology to assess risk, including prospective and historical cohort studies, case-control studies, cross-sectional studies and hybrid designs.

In response to the Agency charge, the Panel identified several criteria for "robust, well-designed epidemiology studies. It is inappropriate to attempt to use these criteria for ecologic studies due to their inherent limitations as described in detail in the body of the report. For the hypothesis-testing designs, the paramount requirement in environmental epidemiology is a well-characterized, quantitative exposure assessment that minimizes exposure measurement error and decreases the likelihood of introducing misclassification in categorical or continuous data analyses. The exposure assessment should be evaluated for accuracy, precision and reliability and should include validation where feasible. Incorporation of biomonitoring and biomarkers of exposure is particularly helpful in this regard. Exposure metrics can represent dose estimates (for example, average daily dose or peak dose), duration of exposure or a combination of these in a cumulative exposure metric.

Other characteristics of robust, well-designed studies identified by the Panel are those that are applicable to all types of epidemiologic studies and include the general criteria of study quality described in a preceding paragraph. The study should have a well-defined population that

includes exposed persons with a wide range of exposures as well as unexposed persons to be maximally informative. Investigators should recognize and attempt to control for selection bias, information bias and confounding and to identify effect modification, particularly as relevant to susceptible subsets of the population, including children. The use of explicit, well-defined criteria for ascertainment of outcome is also important, with recognition of sources of error in data bases such as birth certificate files.

Prospective and historical cohort studies, case-control and cross-sectional studies have significant potential for use in a WOE approach for risk assessment. The extent to which data from these epidemiologic designs can be applied in a quantitative context will depend on the methods used for exposure assessment and the ability of the investigators to make relatively accurate and precise measurements of dose and outcome. The potential weaknesses of each study design are well-described in the literature; thus, the Agency needs to remain cognizant of these when considering the use of data from any single study or an aggregation of studies for a specific pesticide. Well-designed case-control and historical cohort studies may have quantitative value in the risk assessment process, depending on their ability to establish dose-response relationships. Well-designed and carefully executed prospective cohort studies, such as the Agricultural Health Study (AHS), provide maximal opportunity for incorporating epidemiologic data into risk characterization. Prospective cohort studies collect exposure data prior to the onset of disease thus minimizing several forms of potential bias. Further, the depth and extent of exposure assessment, including the potential for incorporating biomonitoring measurements and biomarkers of exposure and effect in selected samples of the cohort, present unique advantages for dose estimation and assessment of multiple biologically and toxicologically relevant endpoints.

The Panel recommends that the Agency consider study quality as the primary factor in selection of studies for incorporation in the atrazine WOE analysis. An important issue is how the Agency decides whether to use particular sets of data. In the interests of transparency, the Panel recommends that the Agency establish a set of criteria for determining the acceptability of epidemiologic studies. These criteria may be based on quantitative criteria, scientific judgment, or a combination of these. Inevitably, it will be necessary to exercise some degree of scientific judgment in this assessment. The Panel recommends that epidemiologists participate actively in the process. Observational research is subject to potential error due to the nature of the science. However, the presence of uncertainty in epidemiologic research does not necessarily imply that the study cannot be used. Epidemiologists routinely make judgments about the extent of potential biases, such as an inability to measure exposure precisely or to arrive at valid estimates of dose, as well as the probable effects of these uncertainties on the risk estimates. In particular, epidemiologists consider the possibility that exposure misclassification may have biased a dichotomous categorization of exposure toward the null and distorted an exposure-response relationship or that differential misclassification, e.g., recall bias, biased the risk estimate away from the null.

The Panel also recommends that novel observations derived from epidemiologic studies should not be dismissed on the basis of a lack of concordant findings in animals. The potential for variation in the MOA across species suggests that such findings be subjected to further exploration. At the same time, the possibility that some statistically significant associations may be due to chance should be borne in mind by Agency reviewers. Similarly, the Panel recognizes

the importance of null findings in epidemiologic research, despite the potential for publication bias. The Panel recommends that a comprehensive literature search be performed, followed by the interpretation of findings by well-trained epidemiologists and others with specific expertise for the relevant exposures and outcomes. With respect to the reproductive outcomes evaluated in the draft framework and case study A, the Panel recommends that future assessments include the full suite of adverse reproductive outcomes including potential effects on the male.

Case Study A

A second goal of the SAP's work was to assess the evidence for an association between exposure to atrazine and adverse reproductive outcomes in five papers published since the 2003 IRED decision on atrazine. This case study is also the first step in EPA's atrazine science re-evaluation plan as described in the November 3, 2009 FIFRA SAP report.

In general, OPP performed an accurate and thorough analysis of the five published studies included in Case Study A and has captured most of the limitations of these studies. The descriptions of potential bias, methods of exposure assessment and statistical analysis were accurate. The Panel made several additional observations relevant to specific studies that are included in the body of the report.

The first weakness of Case Study A was limiting the criteria for inclusion of studies published since 2003 as the Panel was unable to examine the full weight of the epidemiologic evidence for atrazine and reproductive outcomes. All reports in the case study used either an ecological or a retrospective cohort design. Thus the Panel did not have the opportunity to explore how other epidemiologic designs that used cross-sectional, case-control, or other approaches might be incorporated in the risk assessment process. Second, the overall quality of these studies was relatively low, thus limiting their applicability to the upcoming review of atrazine or the more general issue of incorporating epidemiology in risk assessment. Third, two of the five published studies used an ecologic design (Mattix et al., 2007; Winchester et al., 2009). At best, these studies might contribute to hazard identification and problem formulation, but better studies of atrazine and reproductive outcomes are available to meet this goal, and the Panel recommended a comprehensive literature review be undertaken to include all epidemiologic studies of the potential effects of atrazine.

The Panel recommended that the draft framework be expanded to include: 1) background information on the target health effects (definitions, overall incidence rates and rates by age, sex and race, established risk factors for each reproductive outcome, including tobacco and alcohol use, body mass index, nutritional status, reproductive history, medications and recreational drug use); 2) a summary of accuracy and completeness of reporting of endpoints from birth certificates; 3) a discussion of temporal and spatial variations in incidence for reproductive endpoints; 4) a discussion of methods used to handle analyses below the limit of detection; 5) an assessment of whether the endpoints are specific and whether their definitions are precise enough to distinguish specific endpoints resulting from different modes of action; 6) an analysis of whether the observed endpoints were compatible with the known reproductive effects of atrazine observed in animal studies; 7) a discussion about feasible modes of action for chemicals in inducing adverse reproductive outcomes; and 8) examples of well-characterized associations

between environmental exposures and reproductive outcomes from epidemiologic studies of other chemicals. Specific comments regarding each of the five studies are presented in the body of the report.

Case Study C

A case study on integrating human incident data into regulatory risk assessment was included to illustrate an analysis of reported human incident cases using diazinon, a pesticide that has historically been used in residential settings (Attachment C). The Panel's consensus was that little weight should be placed on self-reported incident data in risk assessment. Although human incident data can sometimes be useful in providing information on trends or differences in the frequency and severity of symptoms and whether human effects are consistent with those observed in toxicological experiments or epidemiologic studies, the limitations of using human incident data for risk characterization and risk assessment outweigh the advantages. The major limitations include: 1) likely under-reporting of cases due to the lack of mandatory reporting other than for registrants; 2) uncertainty regarding the exact exposure conditions and amount of the chemical to which the individual was exposed; 3) the fact that human incident data largely capture only acute events and not events with long latent periods or those associated with long-term exposures; 4) the lack of specific training for persons recording information; 5) the non-specific nature of some symptoms and the possibility that these are associated with other illnesses or physiological responses to stress, and 6) the utility of self-reported human incident data only being applicable to pesticides with notable acute toxicity.

The recognized strength of this type of data, in contrast to information from animal toxicity studies, is that responses in humans are detected under real-life situations, with conditions of differential individual sensitivity, modifying factors and other influences possible in the human population. Surveillance for unanticipated effects in incident reports could be useful in identifying alternative mechanisms of action not previously described.

However, the Panel felt that the limitations of the incident data for diazinon out-weigh the possible application of such data for risk characterization. The diazinon case study, as presented by the Agency, is unique because of the distinct symptoms resulting from cholinesterase inhibition and because of the risk mitigation measure of removing diazinon from residential use and the consequent reduction in incidents. Other pesticide groups, such as the triazine herbicide family, that do not produce symptoms of acute toxicity would probably not generate usable incident data for the following analyses.

In conclusion, the Panel recommended that incident reporting data, such as those considered in the diazinon Case Study, could be used qualitatively for problem formulation and hazard identification in the risk assessment process, but their application in risk characterization is very limited unless follow-up information and or laboratory data from individual incident cases become available.

Case Study B

A case study concerning the Agricultural Health Study (AHS), a prospective cohort study of approximately 89,000 licensed pesticide applicators and their wives in Iowa and North Carolina

conducted by the National Cancer Institute (NCI) [Michael Alavanja, principal investigator] was also developed as part of the draft framework. Overall, the Panel concluded that the use of data from a well-designed and carefully executed prospective cohort study, such as the AHS, provides the best opportunity for incorporating epidemiologic data into risk assessment. Further, they also may be useful in comparing dose-response data between humans and laboratory animals for some outcomes. The AHS has the potential to extend the use of epidemiology to risk characterization for agricultural chemicals. Such advances in risk assessment methodology could enhance the usefulness of the risk assessment paradigm for the eventual protection of public health.

However, substantial challenges exist in incorporating exposure data from even a well-conducted prospective cohort study such as the AHS into the risk characterization paradigm. The eventual resolution of large discrepancies between epidemiologic and animal studies in apparent dose-response relationships, the form in which the data are used (categorical assignment of exposure in the AHS versus dose estimates on a continuous scale in OPP) or substantial differences in types of responses between animal and epidemiologic studies, is unclear from the draft framework. If epidemiologic data are to be used to derive quantitative values in risk assessment, determining a process for decision-making in cases in which wide differences are observed in dose-response relationships between animal and epidemiologic studies could clarify the framework and its implementation. At present, it can be argued that in most cases, animal studies can more "robustly" describe dose-response relationships and therefore may currently provide a more reasonable approach for characterizing dose-response relationships and for evaluating mode of action than studies in humans. This does not rule out the possibility that in specific situations epidemiologic data could play an important role in either defining or directly contributing to estimates of departure points and risk.

Therefore, the Panel supports the efforts of OPP and ORD to collaborate with scientists at NCI and the National Institute of Environmental Health Sciences to: a) compare the exposure algorithms used by OPP and the AHS, and b) consider temporal relationships for multi-chemical exposure in the AHS. This work is at a relatively early stage of development but appears to be on track with respect to integrating the exposure data obtained from the AHS with the Agency approaches to quantifying worker exposures.

The Panel recognized the merit of finding commonalities between the Agency's exposure assessment methodology and the AHS exposure metrics. The Agency's method results in estimates of workers' exposure as an input into risk assessment that form the basis for setting exposure limits for workers. Finding commonalities to the AHS exposure metrics would provide a way to extend the usefulness of this large and growing database for the protection of pesticide users.

However, the Panel recommends that this section of the document be revised and streamlined. The AHS exposure metrics and scores, and methods of calculating these values are available in the literature (Dosemeci et al., 2002). However, the Agency's method is less clear. The calculation of exposure (dermal, as expressed in mg a.i./day) using the PHED database could be clarified. Further discussion of the unit exposure parameters can also be added. The Panel also recommended that discussions of the variability and uncertainty associated with the foundational

databases be included for each method, e.g., PHED, AHS's self-reporting and input parameter values.

PANEL DISCUSSION AND RESPONSE TO CHARGE

Agency Charge

1. Draft Framework for Incorporating Human Epidemiologic & Incident Data in Risk Assessment

OPP's draft framework describes a proposed weight of the evidence (WOE) evaluation that integrates science on exposure, pharmacokinetics, and mode of action derived from experimental animal and human *in vivo* and *in vitro* studies. This proposed WOE uses the "source-to-adverse outcome pathway" and the modified Bradford Hill criteria like that in the Mode of Action (MOA) Framework (Section IV of Draft Framework) as tools for organizing and evaluating these diverse types of data to determine the evidence available on the potential human health consequences of pesticide exposures.

Question 1.1 Section II of the draft framework describes the major types of *epidemiology studies* along with their strengths and limitations, factors to consider when reviewing epidemiology studies, and ways to use epidemiology data in risk assessment. Please comment on the soundness and completeness of these discussions. If appropriate, please include comments on additional factors for OPP to consider when evaluating the quality and weighing the utility of epidemiology studies in risk assessment/characterization.

Panel Response 1.1

General Responses for Section II

The Panel commends EPA for its effort to incorporate human data in human health risk assessments in a transparent manner, based on the draft framework. An expanded analysis and rewrite of the extent and nature of reliance upon epidemiologic or incident data within OPP for a range of pesticides in addition to the draft framework and associated case studies on individual pesticides would be informative. Review of the nature and extent of reliance upon human data in risk assessments more broadly within the Agency and as a basis for development of relevant guidance in other regulatory programs is also encouraged. Acquisition of epidemiologic expertise for the review and interpretation of human data, including assessment of exposure in relevant studies, is also strongly advised.

Based on experience in studies of toxicity of industrial chemicals and environmental contaminants, weighting of different types of human data varies and depends on the nature of the studies and on the results (i.e., the weight of evidence of an effect in humans). For example, if sufficient weight of evidence for causality of an adverse outcome in humans exists, and this is based on robust epidemiologic data with well-characterized exposure data, these data would be preferred to those obtained from laboratory animals for dose-response characterization. Further, results of epidemiologic studies in which exposure was well-characterized for a relevant effect in humans could be used as a basis to "bound" dose-response estimates from animal studies.

The lack of pesticide-specific context within the text of the draft framework document (with the possible exception of reference to pesticide-specific exposures) results in rather a general (non-context specific) overview of the strengths and weaknesses of various types of epidemiologic studies. This often involves referencing of generic sources of information on the advantages and disadvantages of various types of epidemiologic studies. However, more recent editions of many of the references and other frequently used sources are available (e.g., Rothman, Greenland and Lash, 2008; Gordis, 2008; Rothman, 2002; Szklo and Nieto, 2007).

Presentation and interpretation of study results need to be considered in evaluating evidence from epidemiologic studies. Authors often base their interpretation on whether a finding is statistically significant, regardless of the magnitude of the association as measured by the odds ratio, prevalence ratio, relative risk, or regression coefficient and regardless of the statistical power for evaluation of each adverse outcome. Therefore, very small but statistically significant associations may be emphasized while larger, but non-significant effects may be ignored. Variations based solely on whether a finding is statistically significant may be emphasized, rather than on the basis of the magnitude of the effect. For example, in at least one investigation in the Case Study B included with the draft framework, an effect measure (correlation coefficient) was not reported because it was not statistically significant. This is inappropriate; focusing solely on statistical significance has the potential to lead to misinterpretation of consistency of reported associations between exposure and effect.

Sensitivity analyses are not uniformly conducted in most epidemiologic studies. Sensitivity analysis can be used to estimate the impact of biases, such as exposure misclassification and potential confounding by known but unmeasured risk factors. EPA should incorporate sensitivity analyses in its list of criteria for reviewing epidemiologic data for risk assessment purposes. When reviewing studies, it is not sufficient to state simply that the study may have suffered from unmeasured confounders or the failure to control for measured confounders. One must first make the case that the risk factors in question are actually confounders, and if it is likely that a confounder may have affected the study results, and that the impact is important enough to affect the interpretation of the findings. In general, the Panel felt that substantial confounding, of a magnitude adequate to change a moderately strong or strong risk estimate to a null finding, is rare in epidemiology.

Specific Responses for Section II: Reviewing Epidemiology Studies for Use in Pesticide risk Assessment

EPA should consider the following issues in reviewing epidemiologic studies for use in risk assessment:

a. Was the epidemiologic study conducted primarily in a hypothesis generating or a hypothesis testing mode?

b. Was the method of assessing exposure accurate and reliable?

c. Were inclusion and exclusion criteria clearly stated and reasonable to provide a representative sample with regard to exposure and health outcome so as to provide a relatively unbiased and representative estimate of effect?

d. Was the method of assessing and the criteria for determining the health outcome clearly stated and valid and reliable; e.g., confirmed with histopathology, and were they designed to detect newly diagnosed (rather than prevalent) cases so that it was reasonably possible to determine that exposure preceded disease?

e. Was appropriate information on potential confounding factors, such as socio-demographic, behavioral and dietary factors collected for both exposed and unexposed groups or for cases and controls in the same way, and were they appropriately controlled in the analyses of the data? Were data on co-morbid conditions collected? (i.e., factors that are associated with the health condition of interest as well as factors associated with exposure)

f. Did the study sample the population or individuals of interest? (i.e., was selection bias minimized and generalizability optimized?) How does the study population relate to the universe of potentially exposed populations?

g. Did the study examine individuals with a wide range of exposures? (i.e., ability to detect a dose-response and to generalize to other populations) Did the study include unexposed populations or individuals?

h. Did the exposures examined in the study relate to past or current situations? (e.g., acute versus chronic exposures and the targeted health end points)

i. Did the study have adequate statistical power to detect meaningful differences for outcomes between the different groups of exposed and unexposed or less exposed individuals while controlling for important confounding factors? Does the sample size take into account the expected incidence of the target health effect in the study populations? (e.g., Page 13, 7th bullet of the EPA Draft Framework for Incorporating Human Epidemiologic & Incident Data in Health Risk Assessment [January 7, 2010] – specify the statistical power of the sample size to detect an effect after adjusting for confounders). Was the study powerful enough to detect as statistically significant meaningful differences while adjusting for confounding variables and exposure measurement error which typically reduce statistical power?

Comments Relevant to Types of Epidemiologic Studies

Case-control studies – These studies generally involve fewer participants than studies of cohorts or dynamic populations. Controls are not always "without a particular disease" (page 14 of draft document). Some control series consist of participants with other diseases not of interest, and some control series consist of a random sample of the study population without regard to status of the disease of interest (so that it is possible for a case to also be a control).

The appropriate selection of cases and controls is important to avoid selection bias. Specific considerations in the assessment of case-control studies include:

a. Determine whether a clear case definition and inclusion and exclusion criteria for the case group have been provided. Case definitions and criteria provide a basis for the minimization of bias and must be sufficiently specific so that cases of specific diseases that have disparate etiologies are not included. For example, although cervical cancer and uterine cancer are both cancers of the reproductive system, they have different etiologies and risk factors so that in a study of reproductive system cancers, the lack of specificity will result in the decrease in risk estimates for a cancer when a true association exists when other cancers for which no association exists are included in the analysis.

b. Similarly, determine whether inclusion and exclusion criteria and sampling techniques result in a control group that is representative of the population that gave rise to the cases.

c. In the selection of cases for inclusion, determine whether the case group includes newly diagnosed (incident) cases or existing (prevalent) cases. Incident cases will be more useful in establishing the likelihood that exposure preceded disease onset and was not related to survival.

d. Determine whether the collection of data on exposures and confounding factors is identical for cases and controls. Where possible, the data collected should enhance the likelihood that the exposure preceded disease onset, e.g., collecting biological samples from cases and controls to ascertain exposure levels may reflect recent exposures that may be due to cases modifying their behaviors because of illness, treatment or altered metabolism due to disease and would thus be less relevant. Validation of exposures and measurement of potential confounding variables to minimize recall bias strengthens case-control studies.

e. Case-control studies, as well as historical cohort studies and cross-sectional studies, may estimate exposures in the past using monitoring data combined with sophisticated modeling techniques ("historical exposure reconstruction"), or in the case of occupational studies, individual, area or plant monitoring data integrated into a job-exposure matrix.

f. Determine whether participation rates of both cases and controls were sufficiently high to minimize the likelihood of participation bias. Eligibility and participation rates and comparison of important covariates (e.g., socioeconomic status, demographics, employment, comorbidity, etc.) should be compared between cases and controls to assess the potential for selection bias.

Cohort Studies – Cohort studies can be conducted retrospectively (historical) or prospectively. Occupational cohort studies most generally take the form of the historical cohort study. The Agricultural Health Study (AHS) is a longitudinal, prospective cohort study in which members are followed over time until they die or the study ends.

An important distinction must be recognized between prospective, population-based and occupational cohort studies. In the former instance, the cohort participants are drawn from a sample of the general population, ideally representative at the time of selection, and then followed over time. In the latter case, members of an occupational cohort are drawn from worker records and do not represent the general population.

"Population-based cohort studies" are those in which that the study group is obtained from a dynamic population such as the population of a state or municipality. Some cohort studies evaluate cohort members only during the period of their residency, leading to ascertainment error. For example, cancers occurring in former residents of a state are not captured by that state but by the state in which they now live. If they leave and are not followed up, this may be a source of selection bias.

Some of the considerations relevant to the assessment of cohort studies include:

a. Determine the extent of attrition or loss to follow-up (especially differential attrition which could lead to selection bias). Determine whether a high rate of participation was maintained longitudinally with minimal attrition so as to minimize participation bias.

b. Potential observer bias can occur if observers are not masked to the exposed/unexposed status of individuals and/or the hypotheses under study, particularly when the outcome(s) involve(s) some subjectivity. Determine whether observers or those who determined the health outcomes of interest were masked as to exposure status of individuals so as to minimize observer bias.

c. Determine whether the criteria for diagnosis/identification of cases are applied consistently in exposed and unexposed populations, assessed with the same frequency in both groups and handled consistently over time.

d. Determine whether appropriate analyses that maximize the use of data for participants who withdraw, move or are otherwise lost to follow-up have been conducted.

e. Determine whether the assumptions of longitudinal analytic approaches are actually met and that these analytic approaches are used appropriately.

f. Determine whether exposed and unexposed individuals were monitored or measured at the same intervals and in the same ways.

The deletion of 'for rare diseases' in the description of cohort studies is suggested. Prospective cohort studies are not efficient for studying rare diseases even when they contain large study populations such as the Agricultural Health Study (AHS), which will have adequate power to evaluate risks for frequently occurring or less rare cancers but substantially less power for rare cancers and other uncommon outcomes.

On page 14 of the EPA Draft Framework for Incorporating Human Epidemiologic & Incident Data in Health Risk Assessment (January 7, 2010), the text states that groups of exposed and unexposed cohorts are studied over time. However, some cohort studies compare the exposed cohort to U.S. (or state) populations rather than unexposed individuals to calculate standardized mortality or morbidity ratios.

Cross-sectional studies – Exposure information is not necessarily obtained "at the same point in time" as the outcome nor is it always just a characterization of exposures around the time that the outcome is measured (page 14). Many cross-sectional studies use historical information on

exposure, which is relevant in relation to temporality. The key distinguishing feature of a cross-sectional study is the measurement of prevalence of disease (e.g., birth defects, SGA), symptoms, biological/physical and physiological response measurements (e.g., pulmonary function tests, blood pressure, chest x-ray, clinical examinations, liver and kidney biomarkers). The ability to conduct screening to make measurements which would not be routinely collected is an advantage of the cross-sectional study.

Prevalence is the proportion of individuals in a population that has disease (Rothman, 2002). Prevalence can be determined as "point prevalence" (e.g., measuring PFTs at a particular workplace at a particular time), or as "period prevalence (e.g., the proportion of cases of autism among those residing in a town during 1998). Cross-sectional studies do not involve a follow-up period (prevalence is a proportion not a rate). Cross-sectional studies are not longitudinal unless one does a series of cross-sectional studies of the same population. In contrast, incidence (the number of new cases divided by the total number of persons at risk for that disease in a population during a specified time period) involves a follow-up period (i.e., it is longitudinal).

A cross-sectional study includes only those who are present when the event takes place (e.g., the pregnancy must result in a live birth; the worker must be currently employed; people who moved out of the town prior to 1998 are not evaluated). A major drawback is that such studies involve "survivor populations" and do not evaluate those who, for example, have left a workplace because they became ill from the workplace exposure. Another major drawback involves the nature of prevalence. Prevalence is a function of incidence and duration. For diseases of long duration, it may not be clear whether the exposure of interest increases the risk of disease or prolongs the duration without increasing the risk of the disease. Incidence is usually more relevant for the types of effects that are generally considered in relation to long term effects of chemical exposures.

Hybrid Designs – Some study designs in which case and control samples are drawn from cohorts as they proceed over time ("case-cohort") or at the end of follow-up ("nested case-control") are not mentioned in the text. These designs can be very efficient and have considerable potential for detailed exposure assessment. Investigators may have collected biologic specimens for evaluation of biomarkers before cases occur. Because these designs are nested within cohorts, exposures can be measured and biologic samples collected at baseline and repeatedly over time before follow-up for disease occurrence. Therefore, there is the potential for less bias in assessing exposure. Another study design that may become important for pesticide research is the case-crossover study design.

Ecologic Studies – First, a single ecologic study cannot provide strong enough evidence to establish a causal relationship (page 15 of the EPA Draft Framework for Incorporating Human Epidemiologic & Incident Data in Health Risk Assessment [January 7, 2010]). This is the fundamental premise underlying the Bradford Hill considerations for weight of evidence, particularly the criterion of consistency. Rather, ecologic studies (especially those with good design), suggest hypotheses for further research (i.e., are "hypothesis-generating"). In practice, ecologic studies are sometimes conducted as a crude approach for testing hypotheses by evaluating correlations but their limitations are well-recognized.

Second, an important distinction must be made between ecologic (group-based) and individual-based studies, particularly with reference to the exposure assessment. One can do a study of individuals and use ecologic data (group data for county, town, or census tract) to assess exposures. However, in ecologic studies, no information is available on whether the people who are diseased are the ones with exposure, while in individual studies, that information is available. For example, a study of disease rates by contamination levels in water on a state basis might be ecologic with respect to the exposure assessment, but the outcome has been obtained at the level of the individual. Thus, the term "semi-ecologic" is sometimes used to describe studies in which the outcome is measured at the level of the individual but the exposure is measured at the level of the group.

The situation in which disease is determined at the level of the individual but exposure is defined by characteristics such as water supply may create confusion in determining the study design. For example, in Case Study A of the draft framework, the Villanueva study (2005) was interpreted as ecologic. In an ecologic study, exposure is assigned to population groups, not individuals. The exposure variable is not representative of any individual's exposure. Instead, the exposure variable characterizes an entire population or geographic area. The ecologic exposure variable is usually of the form of "% of the population with a certain characteristic" (e.g., % unemployed) or of the form "average income". In the case of the two ecologic studies in Case Study A, exposure was characterized on the basis of areas. On the other hand, the Villanueva study assigned exposures to individuals by defining unique water distribution systems with known water sources in which all the population served by each system received similar levels of atrazine (and other contaminants). Although exposure misclassification was likely to exist (e.g., due to bottled water consumption by some women), this was an individual level study.

Ecologic studies are not necessarily geographically-based. For example, a study might compare average rates of exposure in persons who have had special training vs. those who have not had such training or disease rates before and after a major change in exposure.

In addition to the ecologic fallacy mentioned in the text, an additional bias arises from the inability to control properly for confounding factors at the individual level when information is available only at the group level. Finally, in ecologic studies it may be difficult to determine whether exposure preceded the disease.

The results of an ecologic study in which a hypothesis is tested will provide only weak evidence for a causal relationship. This is because of possible ecologic biases, weaknesses in the group-level exposure characterization and uncertainty regarding temporality. However, it is important to assess ecologic studies on the basis of the quality of their design (i.e., some ecologic studies can provide much stronger evidence than others, though the designs in Case Study A are relatively weak). Useful information may be gleaned from an ecologic study if it is well-designed.

Cluster Investigations – The text does not describe this form of investigation. Cluster investigations can be the first alert to a major health issue e.g., vinyl chloride and angiosarcoma of the liver. They can also lead to epidemiologic studies that provide important new information

on the health effects of environmental or occupational exposures, though they are notoriously difficult to interpret, without follow-up.

Meta-analyses and pooled analyses – These are also not described but may be conducted when the body of epidemiologic evidence is sufficient. Such collective analysis can make additional contributions, particularly when the statistical power to detect meaningful effects in individual studies is low.

Important factors to consider when evaluating epidemiologic studies for use in risk assessment

Exposure Assessment

It is important to distinguish approaches which can lead to quantitation of exposure from approaches to exposure assessments that are more qualitative in nature. The quality of the exposure assessment is the principal determinant of the impact of epidemiologic studies in dose-response analysis and risk characterization in risk assessment.

This section in the draft framework (page15) appropriately emphasizes the difficulty of evaluating low exposures. However, it is equally important to stress the need for exposure information sufficient to characterize the time period when the exposure would be likely to have its effect on the outcome of interest. In other words, the timing of exposure may be equally if not more important than the level or duration of exposure. In most instances, the relevant exposures occurred in the past (maybe distant past for cancers and other chronic diseases). "Direct" approaches, such as biomonitoring and personal monitoring, are generally not useful for characterizing prior exposures unless the contaminants of interest are very persistent (i.e., bioaccumulate, long half-lives of excretion). However, historical exposure reconstruction based on available data and sophisticated modeling techniques often contribute significantly in estimating past exposures sufficient for risk assessment purposes and additional consideration of these approaches in this section of the draft framework is suggested.

Historical records and questionnaires – Most often, these approaches do not estimate quantitative levels of exposure but rather, serve as the basis to assign categorical levels. Occasionally, however, quantitative levels of exposure are estimated on the basis of proxy measures (e.g., duration of exposure, pounds of pesticide applied over a particular time). It seems important to emphasize that even when quantitative measurements are available (e.g., personal monitoring), these are often categorized in order to avoid use of a continuous variable in an exponential model (i.e., to avoid the strong model assumption that risk increases exponentially with each unit increment over the full range of exposure). This can lead to measurement error and bias, and risk estimates may be biased in unpredictable ways.

Environmental monitoring – These data can be used to estimate quantitative levels of exposure but there should be some understanding of the relationship between potential exposure and dose. Another advantage not mentioned in this paragraph is that often a large amount of monitoring data is available over time, including data obtained in the past when the exposure was relevant to the outcome under evaluation. In contrast, although biomonitoring reflects internal exposure, it

may provide only a "spot" evaluation of current exposure, but may be more closely related to the adverse outcome of interest in the source to outcome pathway.

Personal monitoring – One disadvantage of personal monitoring not considered in the draft framework is the lower likelihood of having information to characterize exposures during the relevant time period, usually in the past. Also, it is unlikely that the full range of exposures over time will be represented, and sampling may not be maintained over a sufficiently long period to capture peaks and fluctuations. People may change their behavior when they wear personal monitors, and there will be behavioral differences between individuals that will affect personal exposure monitoring results. In other words, personal monitoring may need to be supplemented by environmental monitoring, biological monitoring and/or interview or questionnaire information (e.g., daily diaries).

Biomonitoring – The same disadvantages noted for personal exposure monitoring apply here and some of the disadvantages mentioned in this section also apply to personal monitoring (e.g., sampling over short duration or at one point in time, possibly not measuring the right chemical or metabolite, measuring other chemicals or metabolites, past exposures)

Biomarkers– The selection of relevant biomarkers of exposure and/or effect, based on the toxicological properties of the pesticide (particularly on mode of action) enables much greater likelihood of meaningfully integrating the epidemiological and toxicological databases. (See also comments above on environmental monitoring).

Confounding Factors

The discussion on page 18 of the draft framework is very general and limited, with inadequate detail to enable the reader to understand how "OPP will consider whether relevant confounding factors are properly identified, described, measured and analyzed so that an unbiased estimate of the specific association under study can be made". What aspects and factors will be considered? Some comments with potential relevance to this discussion, particularly in relation to addressing confounding in the evaluation of relevant studies for risk assessment purposes, are presented below.

It is important to define and distinguish clearly between confounding, effect modification, synergy and other mediating effects of covariates. Definitions of confounding should explicitly address the requirement that the variable produces a distortion in the effect estimate either towards or away from the null value. For a variable to act as a confounder it must satisfy three criteria: (1) it must be associated with the disease of interest; (2) it must be associated with the exposure under analysis; and, (3) it does not lie on the causal pathway between exposure and disease. Further, the relationship between the confounder and the exposure or outcome of interest does not have to be statistically significant to have an impact on the risk estimate for the main effect.

The potential for confounding is often mentioned in critiques of epidemiologic studies, but rarely is an argument presented on the likely size of the impact of the bias. In practice, substantial confounding occurs only rarely. The classical case, in which smoking acts as a confounder of an

occupational exposure-lung cancer association, has been found to produce little effect (usually a change of 20% or less in the relative risk). It should be emphasized that a confounder must be a relatively strong risk factor for the disease and be strongly associated with the exposure of interest to create a substantial distortion in the risk estimate. It is not sufficient simply to raise the possibility of confounding; one should make a persuasive argument explaining why a risk factor is likely to be a confounder, what its impact might be, and how important that impact might be to the interpretation of findings.

If unmeasured confounders are thought to affect the results, researchers should conduct sensitivity analyses to estimate the range of impacts and the resulting range of adjusted effect measures.

When deciding whether to include a potential confounder in a regression model, it is important to make sure that the factor is actually a risk factor on its own and not only related to the exposure of interest. Adjusting for a factor that has an association with the disease of interest wholly or partly because of its association with the exposure of interest will attenuate an exposure-disease association, if one truly exists. Adjusting for season of conception may be such an example. Although several reasons are possible for the seasonality of adverse birth outcomes (e.g., disinfection byproducts in drinking water, nitrates, other fertilizers, air pollutants, SES factors), it is also possible that some of the association between season of conception and adverse birth outcomes may be due to, for example, exposure to a pesticide itself. If this is the case, adjusting for seasonality will attenuate the pesticide-adverse birth outcome association (i.e., bias toward the null). In an investigation included in the Case Study A, a variable for "farming exposure" consisting of the proportion of crop-planted land around a home was added to the regression models. Again, this could lead to attenuation of an exposure-disease association because, if it is associated with the outcome of interest, this is likely to be due to it also being associated with the exposure of interest and not a risk factor on its own.

Additionally, the lifestyle factors mentioned in this section as potential confounders are rather limited in scope. For example, confounders may include dietary factors other than high energy diets; physical activity may not be just inactivity; and other factors (e.g., genetics, comorbidities, medications, alternative therapies, alcohol consumption, etc.) might be potential confounders.

When collinearity exists between contaminants (e.g., trichloroethylene and perchloroethylene often occur together in drinking water, or pesticide mixtures containing several chemicals), it is difficult to disentangle the effects of the individual contaminants. If a chemical that is correlated with the chemical of interest is also a risk factor for the disease being studied, then it can create confounding. "Residual confounding" occurs when the additional exposure(s) or other confounding factors are not measured and controlled in data analyses. Studies should include a discussion of residual confounding (and possibly a sensitivity analysis) if the impact may be important enough to affect the interpretation of the findings.

Finally, it should be emphasized that errors in the measurement of a confounder (i.e., confounder misclassification) may result in the inability to provide adequate control for its confounding effects.

Effect Modification

It is important to separate issues of confounding from effect modification. Confounding is a bias that investigators seek to minimize. Effect modifiers are variables that change the magnitude of the association across strata of the study population, e.g., age, race/ethnicity, socioeconomic status, genetic polymorphisms. Investigators evaluate effect modification by the examination of interaction terms in multivariable models or by stratification of the data across levels of the modifier. Effect modifiers may or may not be confounders; e.g., smoking may be a confounder in studies of airways disease but may also be an effect modifier if the risk of exposure to an air pollutant is higher among smokers than non-smokers. Investigators search for evidence of effect modification because differences in risk across strata of the population may provide important information in evaluating the association between exposure and the effect of interest and may be important in identifying susceptible populations. Unless a study is designed specifically to evaluate effect modification (e.g., smoking-asbestos interaction for lung cancer), it may lack statistical power to evaluate such an effect if a sufficient number of individuals are not available in each stratum.

Misclassification of Exposure and Outcome

The text on page 19 of the draft framework should make a clear distinction between differential and non-differential exposure misclassification. With respect to non-differential exposure misclassification, care should be taken to differentiate the direction of the distortion of the effect estimate toward or away from the null depending on the categorization of exposure (dichotomous vs. multiple levels of exposure). A table to illustrate the effects of non-differential exposure misclassification, such as those contained in several texts, would be helpful to illustrate the changes in the odds ratio or relative risk that occur with modest, but anticipated and relatively frequently occurring degrees of misclassification in the range of 10 to 20 percent.

Methods for assessing the impacts of exposure misclassification bias, selection bias, and confounding bias exist. Inclusion of these in relevant studies should be encouraged. The "healthy worker effect" is not mentioned in the text and can create an important bias in occupational studies. The "healthy worker effect can lead to bias toward the null and below, creating the interpretation that the exposure is "protective".

Disease misclassification bias should be elaborated in the text. In the studies involving birth defects included in Case Study A, the disease misclassification resulting from the use of birth certificates to ascertain birth defects was probably non-differential, resulting in a loss of statistical power. However, it is possible that the bias could have been differential with respect to exposure status leading to bias either toward or away from the null.

Issues in Statistical Analysis

The discussion on page 19, related to statistical analysis, should be expanded to include several additional points. When an outcome under evaluation is rare or the sample size is relatively small, it is possible to introduce "statistical bias" in the analysis by including too many covariates in the model. The resulting effect estimate may be biased in either direction but is

unlikely to reflect the true estimate of effect. Statistical bias is also possible when conditional methods are used (e.g., conditional logistic regression when the design includes matching of the comparison group to the group under study). If too few discordant pairs (or discordant sets) are observed (i.e., the number of discordant pairs is small, thus analogous to the situation of a small sample size), statistical bias can also occur, producing a biased estimate of effect. Although it may be important to control for potential confounders, one must take care not to over-control or to end up with cell sizes that are so small that statistical bias is introduced as a result. In such situations, it may be more important to seek parsimonious models that adjust only for the most influential confounders so that the effective sample size is adequate to address the research question without introducing bias due to the statistical modeling.

In evaluating statistical results, it is important to consider the magnitude of the association as measured by the relative risk, odds ratio, risk ratio, regression coefficient, etc. Strong relative risks are unlikely to be due to unmeasured confounding, while weak associations may be due to residual confounding by variables that the investigators did not measure or control in the analyses.

The draft document should also include a discussion of statistical significance in the context of the clinical/biological/scientific significance of the result and the difference between biological and statistical significance. Some statistically significant associations may have little clinical or biological relevance; conversely, some associations that fail to meet the criteria for statistical significance (which are somewhat arbitrary) may be important clinically or from a public health perspective and merit further investigation, especially when the association is strong (but imprecise).

Interpretation of null studies

Exposure measurement error resulting in exposure misclassification is the most likely cause of null findings when an association truly exists, followed by lack of statistical power due to inadequate sample size to detect small effects or to include sufficient numbers of individuals with the health outcome(s) of interest (page 20). In the studies of birth defects considered in Case Study A, ascertainment was based on birth certificates, likely resulting in substantial under-ascertainment that would reduce statistical power to detect meaningful differences.

In addition to the factors described above, a study may be "null" because the exposure is below a threshold at which an effect would occur or be detected. An assessment of the study's statistical power to detect the magnitude of effect of interest is important in interpreting null results. Information on mode of action as a basis to assume site concordance is critical in interpretation of results for risk characterization. If adequately addressed, such results can be used to "bound" dose-response estimates from toxicological studies.

Publication bias related to exclusion of null studies from the literature is discussed on page 24.

External validity

The issue of generalizability in etiologic research (page 20) concerns whether exposures were similar (dose, timing, duration, etc.) and whether important effect modifiers (e.g., sensitive or vulnerable populations) were considered. It is not only an issue of whether a sample is "representative" of the larger population of which it is a sample. Therefore, if exposures are similar, the results found for agricultural workers in NC and Iowa should be fairly generalizable to other agricultural workers in other states, unless important effect modifiers are present that are distributed differently among agricultural workers in NC, Iowa and/or other states (e.g., CA), for example, racial/ethnic distribution when genetic predisposition may modify the effect of exposure. The results of the AHS may not be generalizable to migrant farm workers and their families, because these populations may have a different distribution of risk factors than the AHS population, including race/ethnicity and potentially lifestyle factors and comorbidities, that can act as effect modifiers in addition to their exposures being different (e.g., they may have more intense exposures to pesticides).

Benefits and uses of epidemiologic data in human health risk assessment

This section describes well some uses and benefits of epidemiologic data for risk assessment (i.e., hazard identification/characterization, exposure characterization, and dose-response characterization). Among the unmentioned advantages of epidemiologic studies is that they evaluate the actual conditions of exposure in human participants. On the other hand, in animal studies drinking water exposures are simulated often solely by oral administration, sometimes at unrealistic doses for humans, and occupational exposures are similarly simulated by a single exposure route, usually inhalation.

On page 20, the first paragraph under (C) contains a statement about how "high quality studies with robust exposure assessment may be used to estimate risk quantitatively". This is subsequently qualified to indicate that "most epidemiology studies suffer some limitations in size, scope, exposure assessment or data analysis which prevent their use in quantitative risk assessment" (referenced to Calderon, 2000). Although some epidemiologic studies may not be useful to quantify risks, others have been used for this purpose (e.g., occupational and drinking water studies have been used in trichloroethylene risk assessments, studies of occupational exposure to radon have been used to assess lung cancer risk). As indicated in the general comments included at the beginning of the response to this subquestion, it would be informative to provide an analysis of in what circumstances and how epidemiologic data have been used in risk assessment (for pesticides, specifically and more generically, for broader U.S. EPA program mandates). One example of data from an epidemiologic study having been used to inform the outcome of risk assessment in a quantitative manner is the NIOSH dioxin study (Fingerhut et al., 1991).

Question 1.2 Section III of the draft framework describes the major sources of *human incident data* along with their strengths and limitations. Section III also describes ways to use human incident data in risk assessment. Please comment on the soundness and completeness of these discussions. Please include comments on additional factors to consider when evaluating the quality and weighing the utility of human incident data in risk assessment/characterization.

Panel Response 1.2

Information on how OPP uses human incident data is very general. Case Study C is also fairly unique and as a result, has limited applicability. As indicated in the general comments in the response to Question 1.1, it might be helpful to include an analysis of how incident data have been used elsewhere in the Agency. In particular, some examples as to how this information has indicated the need for a new risk assessment or risk management (as specified in the report) would be helpful.

The description in this section presenting the toxicity data within these human incident reports seems mainly focused on their severity ranking. However, when data are used in risk assessment, the nature of the toxicity endpoints is important, and, while this was discussed in the draft document with respect to the diazinon case study, greater clarity in the text of Section III would be helpful. Most importantly, it should be emphasized that the utility of these data within the framework of risk assessment can be maximized when evaluated in the context of possible mode(s) of action and any related *in vitro* data.

In this context and in relation to potential other uses of the human case reports and surveillance of acute poisonings, incident data can be helpful in considering similarities of site concordance of target organs between animals and humans. This is important in mode of action/human relevance analysis. Use of these data is not restricted to hazard identification, but extend to hazard characterization. Other possible uses of the data include identifying vulnerable or sensitive population subgroups (e.g., age, gender, occupation and demographics), albeit not at molecular levels. This may be helpful as a basis for refining the risk assessment to ensure adequate protection for all subgroups or possible exposure scenarios, or for targeting safety policies or outreach for the reduction or prevention of poisoning events.

In view of their limitations, however, reliance on incident reports in quantifying risks is necessarily limited. Incident reports usually involve high doses and frequently involve illegal or accidental exposure, and, as a result, they will not be reflective of normal use exposures. The incident data are frequently of limited detail and largely the observations of non-medically trained individuals, based on short term exposure with only limited follow-up. The exposure estimates in these cases are normally semi-qualitative at best, and of limited reliability. In addition, they generally involve exposure to products rather than single chemicals, so the possible interactions of the main active ingredient with other chemicals are unknown. Also, probably little, if any, information is available in incident reports to indicate the nature of other contributing factors or confounders. In addition to those mentioned above, the uses cited for incident data (i.e., need for changes in risk management, monitoring success of mitigation

measures and targeting enforcement activities), are all reasonable, assuming their reliability has been assessed critically.

Of special concern in the interpretation of incident data is the reporting of symptoms by medically untrained individuals and without specific disease criteria. When an accidental exposure occurs at a potentially high dose, it is likely that reported signs and symptoms may reflect physiologic responses to the fright induced by the situation rather than the effects of the chemical. Caution is urged in the interpretation of symptoms that could be attributed to physiologic stress reactions if these are not consistent with the plausible toxicological effects for a chemical. The classic flu-like symptoms that are frequently cited as an acute adverse consequence of exposure to some pesticides could also be attributed to some non-chemically-induced causes.

The several sources of incident data described by OPP also vary substantially in their completeness, level of description and geographic scope. EPA has reasonably evaluated the utility and reliability of these five data sources, and the summary in Table 3 is particularly helpful. Additional factors to take into consideration when evaluating the quality of human incident data and their utility in risk assessment/characterization are:

a. Whether the reporting system is active (i.e., the repository agency for such information seeks out reports or incidents by contacting the relevant parties [e.g., officials, health care providers, workers, manufacturers and owners] on a regular basis) or passive (i.e., depends on the relevant parties to report to the agency or repository body).

b. Whether reporting is mandatory, (i.e., requires certain officials, physicians and other health care providers, applicators, manufacturers, farm owners, etc. to report any incident) or voluntary and who is required to report.

c. Greater clarity about how the data are used for estimating "trends over time"; more weight might be given to incident reporting that tracks pesticide use over time so that more or fewer pounds of use should correspond to more or fewer incident reports (assuming all other things such as protective gear, weather, characteristics of users or those exposed remain unchanged).

In addition to the need for analysis of validity and reliability for incident reports, and considering the limitations mentioned above, it is possible that a few individual sets of incident data from which the exposure level can be reliably estimated might contribute to modifying, improving, or confirming the certainty (or uncertainty) of an existing risk assessment based only on animal studies with default uncertainty factors.

Missing from Section III is an explicit statement of intention to estimate or quantify exposure and/or dose for the incident reports. For example, the Panel encourages the Agency to explore linking the California Pesticide Illness Surveillance Program (PISP) data to that in other related databases, such as information in site-specific pesticide use data, to see if the exposure scenarios can be better characterized or confirmed.

Throughout this section, recognition should be explicit that in part, the objective of OPP's work is to avoid incidents. The limited contribution of incident data to quantitative risk assessment needs to be considered in this context.

Question 1.3 Section IV of the draft framework describes a proposed WOE approach for evaluating human and experimental animal data from *in vitro* and *in vivo* studies. This proposed approach makes use of the "source to adverse outcome pathway" and the modified Bradford Hill criteria (like that in the MOA Framework) as tools for organizing, evaluating, and describing the human health consequence of a particular chemical based on the available data. Please comment on the proposed use of modified Bradford Hill criteria in the context of the source to adverse outcome pathway for integrating a variety of types of data at different levels of biological organization including human incident and epidemiologic data in risk assessment.

Panel Response 1.3

The nature and application of the Bradford Hill criteria or viewpoints[a] in the framework addressing source to adverse outcome pathway for integrating a variety of types of data at different levels of biological organization vary from their more traditional consideration in assessing the weight of evidence of epidemiologic data, alone. For the framework integrating human and animal data in the context of mode of action, the proposed criteria or viewpoints appropriately represent those related principally to weight of evidence rather than consideration of individual studies. They have also been appropriately modified to be more relevant to the intended context.

Framed in this context, it is important to keep in mind the original purposes of the "Hill" criteria. As first proposed by Bradford Hill, the principles were meant to bring some structure and rationality to the difficult art of interpreting observational data affected by confounders and other sources of uncertainty. Thus, their adaptation to the specific issue at hand is appropriate because they are rarely used exactly as Hill first proposed them.

Prior to considering the weight of evidence of human and animal data in the context of the framework, the weight of evidence for causality in epidemiologic studies is generally addressed. This takes into account the Hill criteria or viewpoints relevant to consideration of individual studies (e.g., temporality, strength; dose-response). Overall weight of evidence of causality is then considered on the basis of those considerations /viewpoints (e.g., consistency) relevant to the collective database. Some of the criteria/viewpoints are relevant to both (e.g., strength; dose-response).

Prior to conducting a WOE analysis, the investigators should assure themselves that they have accessed the complete body of relevant epidemiologic literature available from peer-reviewed sources. A plan should be developed for the literature search that incorporates second and third level searches in the published literature as well as using the standard approaches of literature searching such as PUBMED (http://www.ncbi.nlm.nih.gov/pubmed/). See further discussion on page 52.

[a] While referenced as the "Bradford Hill criteria", in fact, these criteria or viewpoints first appeared a few months earlier, in nearly the same form, in the 1964 Surgeon-General's report on the health effects of cigarette smoking. (Smoking and Health: Report of the Advisory Committee to the Surgeon-General of Public Health Service. U.S. Dept. of HEW, Public Health Service, GPO, 1964.), and potentially even before. They might more appropriately be quoted as "frequently attributed" to Bradford Hill.

For example, point 1 on page 28 of the draft document refers to the literature search, but does not say how that search will be organized and conducted, nor does it refer to the necessary screening of papers to find those with some merit for the present purpose in, for example, problem formulation. Given that the literature on important public health issues is often vast, complex, largely of poor quality and not well-focused (Bailar 1989), it is recommended that the likely contribution of various sources of data be considered early in scoping as a basis to conserve resources.

Study quality is an important consideration in evaluating epidemiologic data for inclusion in WOE approaches. Criteria for assessing study quality should be explicitly described, though are necessarily somewhat empirical in nature. General guidelines are found in the meta-analysis literature where similar considerations are relevant for selecting studies for inclusion. Elements of study quality may include but are not limited to the following considerations: study design, the existence of an *a priori* hypothesis versus an exploratory analysis, sample size and statistical power adequate to detect the size of the meaningful effect under evaluation, ascertainment of the outcome in terms of sensitivity and specificity, quality of the exposure assessment and the potential for differential and non-differential misclassification, measurement of key potential confounders, assessment of other forms of potential bias, evaluation of effect modification, statistical analysis, the possibility of multiple comparisons unsupported by *a priori* hypotheses, or other supporting data or biological plausibility and others. Many of these issues are introduced on page 28 of the draft framework.

Studies demonstrating no association with a pesticide exposure are equally as informative in a WOE analysis as those that do, provided they meet the criteria for quality described above. Publication bias resulting from rejection or failure to submit 'null' studies is of concern. However, several journals such as Epidemiology have an explicit editorial policy of not rejecting 'null' studies on the basis of the findings.

It is often useful to distinguish between two types of evidence from epidemiologic studies: the quality of the information available and the size of the effects reported. The OPP draft might usefully point this out, and make the intended meaning clear each time the phrase is used.

In the context of assessment of the collective weight of evidence for causality in human studies, alone, the modified Bradford Hill criteria are highly appropriate for framing the description of the likelihood of a specified consequence of a particular chemical using data that are typically available. These considerations are well-accepted for assessing evidence of causation in the field of epidemiology and in public health. It is important to view the considerations not as a checklist, but rather as a group of characteristics that taken collectively provide a systematic way to aggregate observations and guide assessments and conclusions. They require flexible interpretation.

As noted at the beginning of this response, the application of these considerations in the context of assessing the weight of evidence for hazard based on integration of epidemiologic and toxicologic data necessarily varies from their more traditional consideration in determining the weight of evidence of causality from epidemiologic studies, alone. It may be important, then, to clarify terminology in the documentation and to distinguish explicitly potential variations in

interpretation of the considerations/viewpoints in the context of the framework vs. consideration of epidemiologic data, alone.

Biological plausibility needs special attention in this context. In an epidemiologic sense, for example, it sometimes means the integration of everything else one knows at the time of analysis, to develop a sense of the likelihood that a real effect is present. This is the statistician's concept of prior probability, and may be the simplest and most straightforward way to express the concept. However, other interpretations are possible.

In the context of weight of evidence for integration of toxicological and epidemiologic data, biological plausibility takes into account side by side comparisons not only of chemical specific toxicological data including that on mode of action, but integrates what is known from a broader understanding of biologic principles.

Apparent lack of biologic plausibility should not, in itself, be a reason for inaction and the Agency should leave open the possibility that new effects or different manifestations of effects might be revealed in human data. The definition of biologically plausibility evolves over time and may involve a great deal of uncertainty. Biological plausibility based on toxicity studies should not be used to negate contradictory evidence from epidemiologic studies. A potential way of further clarifying the distinction between use of these viewpoints in integrating epidemiologic and toxicologic data versus assessing the weight of evidence of epidemiologic data alone would be to consider differential weighting of the criteria. However, lack of available data may preclude this for specific applications. Establishing dose-response relationships is highly desirable in epidemiologic studies and should be weighted heavily. Temporal association is critically important in that exposure must precede effect and should be weighted accordingly. Strength, consistency and biological plausibility should also be weighted heavily. On the other hand, specificity of the exposure-response relationship has fallen out of favor, particularly in studies of chronic diseases following systemic exposure and should be considered the weakest criterion, if at all.

Based on increasing experience in application of the mode of action/human relevance framework and to avoid the mistaken idea that this addresses exposure in any way, it is suggested to consider revising reference to "dose-response relationships" to concordance of dose-response relationships between the key and end events". Earlier key events are expected to occur at lower doses, and, if this is not the case, the data do not support the hypothesized mode of action. The incidence of earlier key events is also expected to be greater than or equal to that for the end toxic effect; if this is not observed, the weight of evidence does not support the hypothesized mode of action. See Le Marchand (2005) for a discussion of how biological measurements at the cellular or molecular level are being used in cancer risk epidemiology.

Question 1.4 OPP has extensive experience applying the MOA Framework to experimental animal data. However, OPP has not yet completed a WOE approach that also includes epidemiology or human incident data like that proposed in Section IV of the draft framework. Please include in your comments what, if any, additional scientific considerations not discussed in the draft framework OPP should take into account when conducting such WOE analyses.

Panel Response 1.4

The draft framework for integration of *in vitro*, *in vivo* animal and human data has many advantages and the Agency should be congratulated for their efforts. Application of frameworks as a basis for increasing transparency and consistency in the evaluation of weight of evidence undoubtedly has potential for improving risk assessment.

The use of the "source to adverse outcome pathway" and the modified Bradford Hill criteria (like that in the MOA Framework) is also extremely helpful not only as a basis for organizing, evaluating, and describing the potential implications for human health of a particular chemical based on the available data but also in identifying critical data gaps. It is also important in moving the focus in toxicology and risk assessment from late adverse effects to earlier biomarkers of exposure and effect, so that more informative human data at relevant dose levels can be collected. Further, the framework is helpful in directing attention to dose-response relationships for early key events at a very early stage in the assessment of available data; it is these dose-response relationships that are critical in the subsequent risk characterization. The source to adverse effect pathway and framework as proposed in the documentation offers significant potential for the transparent and appropriate integration of human and toxicologic data.

Clear benefit is to be gained, however, in more clearly distinguishing the qualitative and quantitative aspects of a framework analysis as a basis for integration of human data in subsequent dose-response characterization. While pre-existing epidemiologic and incident reporting can be helpful in hazard characterization, unless exposure has been robustly and quantitatively addressed, preferably with inclusion of appropriate biomarkers of exposure and effect based on identification of key events in a mode of action context, the contribution to dose-response characterization will necessarily be more limited (Figure 1 in EPA Epi-Incident Framework Draft document).

Early recognition of the likely contribution of human data in this context in scoping/problem formulation will additionally increase transparency, facilitate peer engagement and necessarily conserve resources. For example, lack of adequate characterization of exposure-response relationships in epidemiologic studies may preclude the need to do an extensive weight of evidence analysis for these data, since they cannot contribute significantly to dose-response analysis and hence, risk characterization.

The value of framework analysis in coordinating assessment and research has also not been emphasized in the documentation. For example, there is repeated reference to problem formulation in the draft framework but without indication of how the broader toxicologic and

epidemiologic databases might be considered in an integrated fashion at this stage as a basis to identify uncertainties and critical data gaps to inform the assessment. This would be an appropriate way to identify limitations of available human data in the context of the overall database as a basis to focus additional research. Appropriate human data, might include, for example, *in vitro* studies in human tissues or cell lines and perhaps focused epidemiologic studies to address specific questions in potentially susceptible subgroups by evaluation of early biomarkers of effect.

Alternatively and/or additionally, scoping of the likely relative contribution of the existing human data (including but certainly not limited to epidemiological data), in the context of the overall database, can be helpful as a basis for engaging peers in the review of preliminary considerations of this information and for focusing resources on additional work to complete the assessment.

In the context of specific content of a framework analysis, based on increasing experience with evaluations of MOA/HR, potential alternatives for hypothesized MOAs would normally be considered at the outset as a basis for distinguishing relevant pathways and key events in an integrated fashion (Page 31, "Other MOAs").

There is also limited reference in the draft framework to how uncertainty will be considered. For example Page 29 refers to "Postulated MOA", but does not deal with the difficulties of determining the MOA with reasonable certainty, or how residual uncertainty should be dealt with in the analysis.

It is important to remember the historical context and purposes for developing the MOA frameworks, which focused originally on cancer outcomes and have only recently been extended to non-cancer endpoints. For decades, positive findings from animal cancer studies were assumed to be relevant for human hazard identification. In the late 1970s and early 1980s research programs were initiated to systematically examine the biologic events that appeared to correlate with and perhaps account for the induction of cancer in a number of common sites for tumor responses in rodent cancer studies. The original framework for mode of action for experimental animal tumor sites and types was established to ensure that the many hypotheses that were being proposed as a basis for considering their relevance were, in fact, based on a solid scientific foundation.

Extension of the framework to consider human relevance represented a logical progression from consideration of the nature of key events in animals to address both qualitative and quantitative aspects of kinetics and dynamics. This proved to be an illuminating exercise for those who participated, and highlighted the extent of erroneous perception that certain animal tumor types were not relevant for humans based on the expectation that humans would not be exposed at sufficient levels to develop tumors, rather than on true differences in physiology or biology.

The most important aspect of the framework for assessing the human relevance of tumors in animals is that in the absence of sufficient weight of evidence that an animal cancer MOA could not occur in humans, the default assumption is that the animal cancer finding is relevant for human health assessment. In this context, it is important that the proposed framework for

incorporation of data from *in vitro, in vivo*, human incident and epidemiologic data does not imply that inconsistent findings in any one area could lead to inaction on the part of the Agency.

Consider, for example, the relevancy of a particular endpoint in a high throughput screening assay to the toxicity pathway that is under consideration or the observation of a strong tumor response in an organ that doesn't exist in humans; e.g., the Harderian gland, Zymbal's gland or forestomach in rodents. The uncertainty associated with the relevance to humans could be considered comparable to that associated with confounding in an epidemiologic study. Confounding is often used to minimize the relevance of elevated risk estimates, when in fact confounding can attenuate or exaggerate a true association. The Agency should guard against inappropriate conservatism in the face of uncertainty when attempting to combine different types of data in reaching public health decisions. By considering all of the "relevant" data in a framework analysis, the EPA cannot raise the bar so high that nothing is recognized as a threat to public health.

While the Panel was given assurance that strong epidemiologic signals would not be ignored, having more data can lead to confusion as easily as to clarity, especially when the data are inherently variable in quality or in relationship to the endpoint of interest. Professional judgment of the strength of the data in the separate areas in the context of both hazard characterization and dose-response analysis will still be necessary before a decision can be reached on the collective cohesiveness or biologic plausibility.

The draft framework notes that consistency may not exist between human and animal model responses. In such situations, it is proposed that the most sensitive endpoint in animal models be used to ensure protection of humans. By way of example, consider that pesticide X inhibits an enzyme resulting in cardio excitation as the most sensitive endpoint in rodents. A similar effect occurs in humans, but at lower exposure concentrations, pesticide X causes headache, slight confusion and nausea. These endpoints are not detected in animal studies. In such situations, it is the relationship between enzyme inhibition and the most sensitive apical effect in humans, which is critical. If, for example, a 25% inhibition results in headaches, etc. in humans, then a 25% reduction in enzyme activity should be used as the appropriate endpoint in animal experiments.

Epidemiologic studies may also suggest other MOAs and/or key events. This needs to be taken into account in relative weighting of information.

The Agency also needs to recognize the difference between poor quality epidemiologic data and epidemiologic data in which results do not fall within the comfort zone whose boundaries were established by the Bradford Hill criteria. Both have limited use in the risk assessment process, but while the former can be ignored (or deferred for possible future use), the latter should raise the possibility that these boundaries need to be expanded, and there should be a strategy to pursue this possibility.

Aspects not addressed within the draft framework include the following: extrapolation among species (mentioned in paragraph 1 on p. 32), and extrapolation of high-dose toxicity in experimental animals and human incident cases to environmental exposure in humans that

typically occurs at much lower doses. Individual susceptibility, genetic variation and other potential effect modifiers are also not addressed.

Additional Specific Comments on the Draft Document

Page 8, Figure 1: It would be helpful to emphasize that the critical connector is mode of action. Biomonitoring data, biomarkers of exposure and effect need to be selected based on what we know from the animal studies. This enables meaningful integration of the toxicologic and epidemiologic data.

Page 9, Table 1: Provides a nice overview of relevant documentation from a range of agencies on which OPP has drawn to develop the framework

Page 10, last paragraph and Figure 2: The current description is not very informative. It would be helpful to describe the nature of the information from a framework analysis that is considered in problem formulation since this is likely to play a key role in the integration of epidemiologic and toxicologic data. This would be the appropriate step, for example, to bring together relevant MOA information as a basis to inform epidemiologic design.

Page 21, paragraph 1, lines 5-7: Some Panel members felt that the NRC (2007) vision relates to a much broader range of human data, than the type currently collected in epidemiologic studies. In fact, it seems that *in vitro* studies in human tissues or cell lines are more likely to be informative in a more predictive mode of action context. In the future, epidemiologic studies will be much less exploratory in nature but focused to address specific questions in identified subgroups defined through consideration of early biomarkers of effect.

Page 21, paragraph, 4, line 2: Animal toxicology studies can be designed to cover a broad range of exposure levels. Unfortunately, in the vast majority of cases, they are not.

Page 31, paragraph 1: The intended meaning of the second full sentence of this paragraph is not clear. "When animal and epidemiological data do not provide a consistent toxicological picture of a particular pesticide more weight would likely be given to those studies with robust study design and availability of replication or confirmatory data." Would more weight be assigned to the human epidemiologic studies?

Page 31, paragraph 1: This paragraph references multiple effects. An adverse effect is considered to be a function of one mode of action. There are key events leading to this MOA with several potential pathways for perturbation.

Page 31, paragraph 2: In the interest of better integrating epidemiologic and toxicologic data, what we should be seeking are biomarkers of exposure or early biomarkers of effect that can be measured in humans, and are based on an understanding of mode of action for critical effects. One Panel member did not fully understand (or agree with) the conclusion that by selecting a biologically plausible and sensitive endpoint from animal studies, the risk assessment is also protective of human health. Why can we assume that the dose-response curves are similar?

Agency Charge

2. Case Study A: Retrospective and Ecologic Non-Cancer Epidemiology Studies

OPP has a dual purpose for developing the Case study A on recent ecologic and retrospective epidemiology studies reporting adverse birth outcomes associated with atrazine exposure. First, the case study illustrates key methodological issues that OPP must consider when integrating ecologic and retrospective epidemiology studies in risk assessment/characterization. Second, this case study reviews several recent studies that will be considered in the re-evaluation of atrazine. Building on the feedback from the SAP at the February, 2010 meeting, these studies will be incorporated in the overall WOE analysis and risk characterization for atrazine. The atrazine WOE is scheduled for review by the FIFRA SAP in September, 2010.

Question 2.1 As discussed in Question 1.1, the draft framework provides general descriptions of the strengths and limitations of ecologic and retrospective epidemiology studies with respect to human health risk assessment. Please describe what you consider to be characteristics of robust, well-designed ecologic and retrospective epidemiology studies.

Panel Response 2.1

General Comments

Clarity of terms must be provided in the Framework document. Ecologic studies (which are based on group level data) must be separated from retrospective studies (which are based on individual level data) because their potential use in risk assessment is very different. Greater clarification is needed, particularly for the term "retrospective epidemiology studies" because many epidemiologic investigators use this term to describe case-control studies in which information about exposure is assessed retrospectively. In Case Study A, the term "retrospective" is also applied to "retrospective cohort studies" (better labeled historical cohort studies) but with a different design. In historical cohort studies the exposed and unexposed members of cohort are ascertained retrospectively, and these persons are followed up to determine incidence rates of the health outcomes.

Case-control studies and retrospective cohort studies share some of the same challenges of accuracy and completeness of retrospectively ascertained exposure information, but they may determine exposure using different methods. In case-control studies, participants are usually asked about prior exposures, so the information gathered may suffer from inaccuracy and incomplete recall. In retrospective cohort studies, the exposed and unexposed cohorts are often identified from existing records on prior exposure; e.g., in an occupational setting. Thus, retrospective cohort studies have the potential to provide more accurate and complete assessment of exposure than participant recall of exposures in case-control studies. It should be noted, however, that even though the Agricultural Health Study is a prospective cohort study, much of its initial exposure assessment that determined exposure groups was based on retrospective recall of exposures, thus making it somewhat of a hybrid retrospective and prospective cohort study.

It should further be noted that nested case-control or case-cohort designs, which were not included in Case Study A are efficient designs that may provide useful information for risk assessment. For example, a nested case control study within the Agricultural Health Study could

provide less potential for bias in ascertainment of exposure than traditional case-control or retrospective cohort designs.

Finally, the use of the term "predictors" in this section is inappropriate in ecologic (and in some cross-sectional) studies because the exposure of interest is usually assessed at the same time as the health outcome so that the temporality of the relationship cannot be determined and thus it is unknown if the exposure is truly a risk factor.

Due to the inherent differences between ecologic studies and other study designs, retrospective studies of various types will be considered separately in response to questions 2.1 and 2.2.

The Panel recommends that this OPP report acknowledge that ecologic studies are inherently weak vehicles for quantitative estimation. Data from ecologic studies may have considerable strengths in other ways: generating hypotheses, supporting smaller and inconclusive data of stronger inherent character, providing "floors" to the size of some effects to support legislation or regulation. However, the strength of ecologic studies should not be overstated in the interpretation and analysis of problems. The quality of ecologic studies spans a spectrum of strengths; not all ecologic studies are equally informative.

Ecologic studies are considered to be of most use in hypothesis generation and are rarely suitable for hypothesis testing because both the exposure and the health outcome data for such studies are collected at the group, rather than the individual level. Thus, they are not useful in quantitative risk assessment. An inherent problem in ecologic studies is an inability to control for potential confounding at the level of the individual when exposure and outcome are assessed at the level of the group. Therefore, adjustment for confounding factors at the population level may not sufficiently remove confounding effects. This in turn, may create disparate findings from studies where adjustment for confounding is performed on an individual basis to obtain summary statistical results comparing groups. This shortcoming, often termed the "ecologic fallacy" renders ecologic studies less useful in the problem formulation stage of risk assessment as well as in risk characterization.

Additional considerations for evaluating the information in Table 1 of Case Study A on page 41 of the draft framework should include whether information on potential confounding variables was obtained and used for adjustment in the statistical analyses. This is particularly important when comparing rates between geographic areas or between a state and the United States as a whole because differences in the population distributions for variables including age, racial/ethnic, socioeconomic status, lifestyle factors (e.g., smoking, diet), and family history, as well as other pesticide, chemical and air pollution exposures, could affect the rates and thus influence the results.

In this respect it is important to note that the CDC natality database that was used for the ecologic studies by Mattix et al. (2007) and Winchester et al. (2009) include data on potential confounding factors, such as maternal demographic variables and behavioral risk factors such as tobacco and alcohol use. However, some of these data are missing for certain states where recording of this information is not required (e.g., tobacco in California, Pennsylvania and Washington) or are not comparable over the years or across the states. This can result in significant amounts of missing data on potential confounding factors so that significant residual confounding cannot be ruled out, and furthermore results may not be directly comparable across

states after adjustment for confounding because the adjustment is based on inclusion of different confounding factors.

However, not all ecologic studies have similar potential for bias, and therefore the strengths and weaknesses of such studies should be considered individually as to their potential to inform risk assessment. The group level data may be more or less refined. For example, a study of disease rates by contamination levels in water on a state basis might be ecologic, but more information is available in comparing counties, still more in towns, or water districts, and still more in comparing rates with persons classified by household, all without individual data, though the last might be better described as individual-based. Some investigators refer to studies in which outcome is assessed at the level of the individual and exposure (such as concentration of a pesticide in the municipal water supply) at the group level as "semi-ecologic".

In rare instances, such as examining census tract or county rates of disease prior to and following a well-defined event such as introduction or removal of a pesticide from use, or following widespread contamination of the environment following an industrial accident, ecologic studies may be informative enough to incorporate in hazard identification. Such relatively informative examples of ecologic studies compare rates of disease in populations before and after events (with consideration for latency) or across geographic strata of well-defined exposed and unexposed populations.

Specific Comments for 2.1

The Panel provides the following comments regarding characteristics of robust, well-designed epidemiology studies:

Exposure measurement error is the predominant weakness in environmental epidemiology in general and in studies of pesticide exposure specifically because it is rare that individual data from biomonitoring are available to estimate internal dose. Exposure measurement error in turn, creates exposure misclassification when individuals are assigned to exposure categories using cutpoints from continuous data. Therefore, investigators should make maximal efforts to assure that their exposure data are as accurate and precise as study conditions permit, and that validation of estimated exposures be performed if conditions permit. As mentioned in the draft framework, exposure assessment is an important consideration in all epidemiologic studies, irrespective of the design.

As summarized in Table 1 of Case Study A on page 41, exposure assessment generally involves surrogate measures of exposure (e.g., levels of atrazine in the surface or drinking water or proximity to fields in which atrazine was applied), rather than measures of body burden or concentrations in drinking water and amount of tap water consumed to estimate exposure levels (noted on page 45). Robust, well-designed ecologic, case-control, historical or prospective cohort studies should use the best possible measures of exposure to estimate dose. For drinking water exposures, tap water concentrations are preferred over ground or surface water measurements, and information about individual amounts of tap and bottled water consumption at home and at work improve the accuracy of the exposure estimate. For studies that assess exposure through drinking water, sufficient measurements in the distribution system must be available to characterize monthly levels of contaminants. Alternatively, monitoring data and modeling can be used to estimate levels for the critical periods of gestation for specific outcomes. Many reproductive epidemiologic studies estimate exposures for specific trimesters

of gestation; e.g., studies of birth defects and first trimester exposures. Studies involving public water systems are easier to conduct than studies involving private wells due to availability of monitoring data for public systems. If the study focuses on public systems, the study area selected should have information sufficient to define adequately the coverage area of each public system and the source(s) of water for each system. If a study examines exposure from pesticide drift, modeling should be used to estimate levels near the home, and if possible validated by collection of dust samples from within the home. If chronic diseases or cancers are the outcomes of interest, the assessment must be able to characterize exposures in the distant past, which can present formidable challenges.

A potential source of exposure misclassification in reproductive epidemiology is the use of residential address of the mother at the time of the birth to assign exposure categories. Several studies have shown that approximately 25 percent of pregnant women move during pregnancy (Canfield et al., 2006), creating the potential for exposure misclassification when the critical period of gestation occurred prior to the relocation. Similarly, studies of chronic diseases often use the address at the time of diagnosis. Use of these addresses can result in misclassification of the relevant area of residence (e.g., residence at conception for studies of reproductive effects or residence years before disease onset in studies of chronic disease) and thus in the assessment of exposure levels. This potential source of misclassification is recognized in the draft framework. Robust, well-designed case-control and cohort studies should acquire residential histories and use the most relevant residential location(s) for assessing exposure. However, it should be noted that even obtaining relevant residential location may misclassify individual exposures because individuals spend substantial portions of their lives at work or otherwise away from their residences, and the exposures in these locations are often not considered. More recently, investigators have been collecting work address/location information in occupational histories which may reduce measurement error.

Accuracy, precision and reliability of exposure data are important components of high quality exposure assessments. These considerations may include the precision of measurement of exposure or estimated dose, the extent to which exposure measures or categories are well-defined, the incorporation of data on external exposure (e.g., from measurement in the individual's micro-environment) or internal dose through biomonitoring, the use of analogous data as an exposure surrogate, and predicted exposure estimates from validated modeling. Individually collected exposure data, preferably absorbed dose data, would be of greatest value. Exposure indices that have poor predictive ability should be avoided. If an exposure index is used, it should be validated with data. The exposure measurement must provide adequate discriminating power to detect an exposure-related hazard (at a minimum provide reliable gradations of relative amounts of exposure). Exposure metrics can represent dose estimates (for example, average daily dose or peak dose), duration of exposure or a combination of these in a cumulative exposure metric (e.g., area under the curve statistic).

Misclassification of the outcome can also occur in epidemiologic studies. Therefore, case-control, historical and prospective cohort studies should derive data on reproductive or cancer outcomes from registries with mandatory reporting and active surveillance with explicit and consistently used criteria and definitions of outcomes. Similarly, if exposed and unexposed cohort members are followed by routine, regular screening for outcomes, explicit and consistently used criteria and definitions of outcomes should be used. If self-reports are to be used, efforts to confirm diagnoses should be made by review of medical records or data linkage with birth certificate, cancer or other registries. Robust, well-designed studies of all types should

derive reproductive and cancer outcome data from registries with mandatory reporting and active surveillance with explicit and consistently used criteria and definitions of outcomes. In case-control and historical cohort studies, ascertainment should be performed in a similar manner for the exposed and non-exposed group for the outcome as well as for potential confounders. Differences in ascertainment can introduce selection and information bias.

A well-defined study population with inclusion of an appropriate comparison group to address the study hypothesis and objectives is an important component of study quality. The use of population-based registries for cancers and birth defects, and confirmed diagnoses if information is obtained from self-reports is recommended to reduce selection and information bias, respectively.

In the studies summarized in Table 1 of Case Study A on page 41 of the draft framework, much of the data on occurrence of birth defects, preterm delivery and small for gestational age was derived from birth defects registries, birth records and national datasets. Important considerations in using such data sources (i.e., also for cancer registry data) include:

a. whether reporting to the registry or on the birth record is mandatory as this would tend to make these sources of information more complete, and reporting from areas where it is not mandatory could be influenced by factors that might also be related to exposure (e.g., socioeconomic status may be related both to likelihood of reporting and of exposure to pesticides);

b. whether the registry actively identifies birth defects or depends on passive reporting of defects: an active identification system would tend to provide more complete ascertainment of cases;

c. whether reporting to the database depends on who reports, e.g., health care providers or parents or both because the more individuals who are reporting, the greater the likelihood of more complete ascertainment;

d. whether the criteria for and definitions of outcomes (e.g., birth defects, preterm delivery and small for gestational age) have been explicit and consistently used so that data are comparable across years and across regions; and

e. whether the length of follow-up is appropriate. For birth defects, some outcomes are not captured in birth records because they become manifest some time after birth, during the first year of life.

Robust, well-designed studies should make a maximum effort to assure that potential confounding is controlled to the extent possible. Thus, investigators should obtain complete information on as many potentially confounding variables and risk factors as possible from all individuals to reduce the possibility of residual confounding by unmeasured variables. Potential confounders should be evaluated with appropriate criteria and methods to determine whether they are related to the exposures and outcomes and whether their inclusion in multi-variable models produces a change in the effect estimate. Those potential confounders that meet the criteria should be retained in the final models.

Robust, well-designed studies should have sufficient statistical power and information to enable analyses for effect modification. Effect modification is present when the magnitude of the risk estimate varies across strata of another variable such as age, gender, race/ethnicity or socioeconomic status. Incorporating analysis of effect modification is an important component of good epidemiologic studies and is relevant to risk assessment for identification of potentially susceptible subsets of the population such as children. Change in the risk estimate across racial/ethnic groups could be due to genetic differences in the frequency of a polymorphism in a gene that controls transformation or metabolism of a xenobiotic. Effect modification can be detected by stratifying the analysis across levels of the variable or by including interaction terms in the model to assess statistical interaction. Therefore, sufficient sample size and information should be provided to enable stratification on the potential effect modifier(s) with sufficient statistical power to detect meaningful differences.

Robust, well-designed studies should incorporate careful consideration of appropriate analytic methods. Some epidemiologic analyses are based on categorization of exposure; e.g., into tertiles or quartiles depending on the nature of the exposure metric. Linear models are also used when data are available for continuously distributed exposure variables. Smoothing methods may be used to inform the categorization of exposures (instead of categorizing by percentiles, or to check the appropriateness of percentile categorization). When meaningful cutpoints exist (e.g., the MCL) these may be used in the analysis. In addition, when data are available, sensitivity analyses should be conducted for potential exposure misclassification bias, information bias, selection bias (including loss to follow-up), healthy worker/survivor biases, and other confounders to estimate their potential impact(s) and interpretation of the findings. All findings should be reported. Interpretation of findings should not be based solely on statistical significance testing. The precision of risk estimates is based on the confidence interval around the risk estimate.

High quality studies should account for temporal, spatial and individual variability as related to exposure. They should also incorporate a sufficiently long observation period with respect to the expected latency of health effects. For example, exposures to pesticides in water should include a complete residential history and history of water consumption habits; studies of workers should include a full occupational history with address/location information.

Statistical power is an important consideration in evaluating associations. Studies should provide sufficient sample size to examine the relation of exposure to reproductive, cancer or other chronic disease outcomes. Statistical power is the ability of a study to detect an association between an exposure and a health effect, if in fact, one exists. Statistical power should be considered in evaluating null studies to reduce the probability of failing to detect an association if one exists (false negatives). Statistical analysis should include examination of the confidence intervals for a risk estimate as well as the determination of statistical significance. For studies with low statistical power, failure to find statistically significant differences should be interpreted cautiously and should consider the magnitude of the observed effect and variability in the effect estimate.

Data from Winchester et al. (2009) [Table A-2 on page 43 of Attachment A of the draft framework] illustrate another issue in interpretation of some ecologic and cohort studies relevant to sample size. Specifically, sample sizes can be quite large when comparing populations of entire regions e.g., county level data, resulting in very small risk estimates being statistically significant. These data raise the issue of statistical versus biological or clinical significance.

Failure to consider heterogeneity in the categorization of the outcome when developing the case definition may lead to improper conclusions. Analyses that consider all forms of an outcome together (all cancers, all birth defects) are not informative since different etiologic mechanisms are likely to be involved in specific forms of the disorder. Teratogens may be responsible for specific birth defects but have no association with other defects or other adverse reproductive outcomes.

A potential issue in consideration of epidemiologic studies is that of multiple comparisons, particularly when these are not based on well-founded *a priori* hypotheses. Epidemiologists recognize that multiple analyses for multiple outcomes (e.g., multiple categories of birth defects) or analyses that classify exposures in many different forms using varying temporal or spatial constructs or measures of exposure may generate statistically significant differences, some of which may be due to chance alone due to the multiple testing. Such exploratory studies may be useful, but caution must be exercised in interpreting results. In such circumstances, investigators should recognize the potential impact of multiple testing unless a well-developed hypothesis indicates the likelihood of the exposure being related to more than one outcome. Differences in opinion exist with regard to the appropriateness of formal statistical adjustment in these situations, and these are discussed in further detail in response to question 2.4.

Finally, for some exposures and outcomes, published pooled and meta-analyses may be informative. Pooled analyses are particularly useful when original data from multiple investigators are available since study power will be increased. Meta-analyses which incorporate sample sizes and risk estimates from a number of individual studies into a single risk estimate with its confidence interval are also useful to provide proper perspective in reaching conclusions, particularly in a weight-of-the-evidence approach for epidemiologic studies.

Many of the points made by the Panel are also summarized in a paper by Swaen (2006). These characteristics of robust case-control, historical and prospective cohort studies should not be reduced to a checklist because variation occurs from problem to problem and between studies, so that thoughtful interpretation by epidemiologists will remain necessary. These criteria can also be applied to ecological studies, although most ecological studies will fail to meet many of these criteria.

A parallel to this discussion is found in a paper by Shore et al. (1992) which states in part "Good-quality epidemiological studies are those with sound methodology, lack of bias, long enough follow-up times to observe a (carcinogenic) *health effect* response, adequate exposure information, and dose-response information. Before a lack of (carcinogenicity) *health effect* can be inferred, it is essential that the exposures be of substantial duration and intensity, and that the number of exposed persons be reasonably large."

Question 2.2 Ecologic and retrospective epidemiology studies are particularly useful in identifying new hypotheses about the human health effects of pesticide exposure and may confirm the human relevance of findings from experimental animal studies. However, these types of studies do not typically include robust characterization of exposure and they do not address confounding factors as well as prospective studies. Although there may be exceptions, generally, ecologic and retrospective epidemiology studies are not sufficiently robust for use in quantitative risk assessment (i.e., for use in deriving a point of departure or in quantitatively informing extrapolation factors, etc). In light of the strengths and limitations of ecologic and retrospective studies, please comment on appropriate ways to use these types of epidemiology studies in risk assessment/characterization or their utility in problem formulation (e.g. defining additional analyses or research/testing).

Panel Response 2.2

General Comments

As described above, exposure measurement error and misclassification of exposure comprise the most important limitations of epidemiologic studies for incorporation in risk assessment. Epidemiologic studies of pesticides face challenges in assessing exposure accurately and to a lesser extent in identifying and measuring potential confounders and effect modifiers. This is illustrated by the difficulties experienced in trying to obtain reliable biomonitoring data that matches the characteristics of the exposure and the outcomes of personal monitoring studies in studies of workers. Even with some degree of control over the measurements on identified individuals, large variability and associated large uncertainties still occur.

External exposure is a surrogate for internal exposure (exposure of a target organ). For a single well-defined externally applied dose, internal exposure can vary by an order of magnitude between individuals. Even direct measurements can be misleading. For example, the study of Barr et al. (2007) indicates that atrazine mercapturate measurements (the frequently used marker of atrazine exposure) underestimate exposure to atrazine and its break-down products. In both occupationally and non-occupationally exposed individuals, other metabolites (diaminochlorotriazine, desethylatrazine) predominate. Despite these limitations, rigorous estimates (preferably confirmed by measurements; e.g., multiple 24-hour urine samples analyzed in a study of absorbed dose to 2,4-D) (Hayes and Aylward, 2009; Aylward et al., 2010) can provide validation of the external exposure estimates. Similarly, a series of studies by Harris and colleagues show that epidemiologic studies can be conducted using absorbed dose estimates of 2,4-D and other herbicides as well as external exposure estimates and self-reported questionnaire information (e.g., pesticide use) (Harris et al., 2002, 2005; Harris 2007; Harris and Wells, 2007).

Worker exposures typically result in the highest expected levels of exposure, so epidemiologic studies of occupational cohorts can be useful in identifying associations with adverse effects (if these exist). However, the use of worker exposure data does not provide an assessment of exposure of other members of the population. Spouses and children can be exposed through contact with contaminated clothing, contaminated equipment or recently treated areas. This became obvious with the families of workers exposed to asbestos. Although this exposure

should be minimal if good practice is followed, contamination can occur. Other potential sources of exposure (spray drift, water, food) need to be considered.

As with experimental toxicology, epidemiologic studies tend to concentrate on one compound in isolation. It is difficult to identify all of the important compounds in the "chemical soup" that might interact with the action at the target or that might produce similar outcomes. Humans are rarely exposed to individual chemicals, particularly in the context of agricultural chemicals. The assessment of hazard associated with chemical mixtures poses a challenge to toxicologists due to complexities associated with the selection of chemicals in the mixture and exposure concentrations of the individual chemicals. These problems are not confined to pesticides since humans are exposed to many other compounds (e.g., industrial chemicals, pharmaceuticals, components of household products, personal care products {some with seasonal use e.g. sun screen products}) on a regular basis. Well-designed retrospective epidemiologic studies may provide a foundation for assessing hazards associated with exposure to chemical mixtures. Occupationally exposed cohorts represent one opportunity for such studies. Epidemiologic studies may provide insight into exposures to mixtures of relevance to humans. Although challenging, associations with components of the mixture or with the entire mixture as measured could be evaluated using the WOE framework and, when appropriate, used to provide guidance in the design of animal toxicity studies.

Identifying potential confounding factors is an important part of the design of epidemiologic studies. It is not possible to know whether all relevant factors have been identified, and it is important to assess whether known potential confounders have been considered. Given these reservations, well-designed epidemiologic studies have the potential to contribute to the risk assessment/characterization process in a number of areas. An important area is the identification of potential health problems (previously not considered) that may be associated with exposure to pesticides. This can identify compounds that should be regarded as of concern, and provide guidance in the prioritization of research. They can provide some guidance to environmental levels of exposure that may impact adversely on health, and could inform research to determine internal exposures corresponding to observed external exposures. Well-designed epidemiologic studies may help in the identification of sets of lesions that could be investigated using toxicodynamic and molecular methods to determine the mode of action of the target compound.

A formal framework for validating epidemiologic methods, particularly for exposure, is needed. The development of such a framework has transformed the field of analytical chemistry over the last twenty years. Validation protocols have led to greatly increased reliability of analytical chemical data. Epidemiologic and toxicological studies could provide better information (for instance, improved definition of the uncertainties associated with the estimation of exposure, and identification of target sites relevant to human exposure) if investigators from varying disciplines collaborated with each other. For example, banking of bio-specimens and environmental samples could be used to increase the utility and reliability of future cohort studies and permit validation of exposure estimates.

It would be beneficial to have a framework for assessing the scientific validity of the outcomes being investigated. This would ensure that studies using categories containing multiple endpoints resulting possibly from different modes of action are identified, and evaluated

appropriately. Although the statistical methods used by the various authors were described in the case study, their appropriateness was not evaluated. A framework in which to consider the strengths and weaknesses of any analysis would be of benefit, and would focus attention on possible alternative ways of analyzing the data.

Specific Comments for Types of Epidemiologic Studies

Ecologic Studies

As described in the response to question 2.1, ecologic studies are considered to be of most use in hypothesis generation and are rarely suitable for hypothesis testing since data are collected at the group, rather than the individual level. Thus, they are not typically useful in quantitative risk assessment unless the target of interest is the larger group as is sometimes the case in descriptive epidemiology. The use of surrogate measures of exposure may create sizeable misclassification of exposure. Thus they are largely useful (arguably more so than incident data), for suggesting hypotheses to be addressed in future well-designed studies of individuals and in examining consistency of findings.

Due to the problems inherent in the "ecologic fallacy" and the inability to control for potential confounding at the level of the individual, ecologic studies are also less likely to be useful in the problem formulation stage of risk assessment. In rare instances, such as when examining census tract or county rates of disease prior to and following a well-defined event, such as introduction or removal of a pesticide from use, or following widespread contamination of the environment following an industrial accident, ecologic studies may be informative enough to incorporate in problem formulation. An example of the latter occurred at Seveso, Italy in 1976 when a substantial proportion of the residential population was exposed to dioxin - 2,3,7,8-tetrachlorodibenzo-p-dioxin (TCDD). The event was followed by short-term studies of morbidity and by a long-term prospective cohort study to evaluate mortality and cancer incidence based on zones of exposure to the contaminant (Bertazzi et al., 2001). In this example, exposure was assigned based on soil measurements of TCDD and a limited number of human samples (Bertazzi et al, 1998). Such relatively informative examples of ecologic studies compare rates of disease in populations before and after events (with consideration for latency) or across geographic strata of exposed and unexposed populations.

Ecologic studies are often exploratory. However, observations made in these studies can be used to direct future hypothesis setting and analysis. They can also provide insights into future needed analyses and research, i.e., identifying gaps in knowledge and informing problem formulation and driving the research agenda. Exposure characterization is an inherent weakness of ecologic studies since it is not focused on the individual. Accordingly, these studies are of limited value in this aspect of the risk assessment process. This makes it difficult to use such studies as the basis of hazard characterization. However, the results can be corroborative of the hazard characterization obtained from toxicological studies using animal surrogates, and can identify hazards that are unique to humans, and not seen in animal models.

One important potential use of ecologic epidemiologic studies is in the monitoring of the effects of mitigation measures implemented as a result of the risk assessment process. Data at the

population level may be informative in a temporal analysis in which disease rates are compared for the same population before and after implementation in a regulation or changes in a standard (i.e., as a check on the correct identification of mitigation requirements and the effectiveness of risk management measures aimed to protect human health). Further, on the basis of published studies it may be possible to identify opportunities for validation of findings where a particular pesticide is removed from use, or restricted in use in a region (e.g., atrazine in many European countries) or where exposure is changed (e.g., due to improved drinking water treatment). However, the latter would simultaneously reduce levels of other contaminants found at low levels in treated drinking water. Additional analyses and research can be appropriately proposed based on preliminary findings from ecologic or historical cohort studies.

Retrospective Epidemiologic Study Designs

Several study designs comprise the group designated as "retrospective", including case-control studies, historical cohort studies, nested case-control studies within cohort studies and cross-sectional studies. These vary in their design characteristics and in their strengths and weaknesses. In turn, their usefulness in informing risk assessment will vary depending on the design, the quality of the exposure assessment, study power, the ability to adjust for potential confounders and other considerations described in response to question 2.1.

Case-Control Studies

The case-control design is frequently used in epidemiologic research because it is suitable for the study of rare outcomes and is relatively efficient. Well-designed and executed case-control samples can provide valid results comparable with what can be gleaned from a cohort study but at considerably reduced time and cost (Rothman, 2002). Various forms of bias can be introduced into case-control studies unless rigorous attention to appropriate design is provided. Information bias due to differential recall or reporting and selection bias due to differential participation or inappropriate selection of control groups are potential problems. However, these potential flaws are not necessarily present in and are not restricted to case-control studies, and reviews of their usefulness should examine the potential for introduction of bias and if present, the likely extent and direction of such bias. The retrospective nature of the case-control study is due to the sequence of events in carrying out the study. Typically, cases and controls are identified from suitable populations and information about exposure is collected subsequently. For studies of reproductive outcome, the length of time between identification of cases and controls and ascertainment of exposure may be relatively short, thereby reducing the extent of misclassification of exposure due to faulty recall of distant events.

A relevant example of a case-control study is one in which men with reduced semen quality (cases) were compared with a group of men with normal semen quality (controls) for biomarkers of pesticide exposure (Swann et al., 2003). The hypothesis was developed when the investigators initially found that men who lived in an agricultural area of Missouri had reduced sperm concentration and motility compared with men who lived in an urban area. They then conducted a case-control study by enrolling men with abnormal semen parameters (low concentration, lower percentage motile sperm and higher percent of abnormal sperm morphology) along with men with semen parameters within normal limits from the same areas in

the Midwest. They then analyzed urine samples from all men provided at the time of semen collection for concentrations of metabolites of eight pesticides. They found that pesticide metabolite levels in cases living in one of the states (central Missouri) were higher than those in controls for alachlor, atrazine and the insecticide diazinon (Swann et al., 2003). This study has some characteristics of a cross-sectional study because information about the disease status and exposure was collected simultaneously. Nonetheless, it illustrates the potential for obtaining quantitative data about pesticide exposure from biomonitoring using a case-control approach that may be incorporated into quantitative risk assessment.

Historical Cohort Studies

In the historical (or retrospective) cohort study, the cohort is identified from historical exposure records or data, such as biologic samples that may be available from samples collected in the past. Follow-up is conducted for cohort members, and disease status is ascertained at the time of the study, permitting comparison of rates among persons with varying levels of historical exposure. Thus, this design is more efficient and less costly than the prospective form of the cohort study, provides a better temporal sequence of exposure to outcome than ecologic or cross-sectional studies (making it more useful in risk assessment), and is often used to assess the relationship between exposures and health outcomes among occupationally exposed persons. When historical data are available for exposures to pesticides that occurred in the past, this form of retrospective study may be useful for risk assessment. Two examples of historical cohort studies are included in Case Study A (Villaneuva et al., 2005; Ochoa-Acuña et al., 2009) and are discussed further in response to question 2.3. These retrospective cohort studies are based on measured concentrations of atrazine in municipal water treatment plants and ascertainment of outcome for women who lived in the communities supplied by their respective water utility. The quality of the historical data and the degree to which exposure can be assigned to each individual will determine the extent to which the data will be valid for quantitative risk assessment.

Nested Case-Control Studies

In some instances, case-control studies may be incorporated within cohort studies and are referred to as "nested case-control studies". This design is useful when the costs of analysis for pesticide exposure are too high to study the entire cohort. A control group is selected from among all eligible controls and compared to cases that developed during the study period. For example, Krieger et al. (1994) conducted a nested case-control study of 150 women with breast cancer and 150 disease-free women, selected from a cohort of 57,040 women who had been enrolled during a multiphasic health examination in the late 1960s and had a serum sample collected and frozen at the time of examination. The women were followed up through 1990 for the development of breast cancer. Each case was matched by race to a cancer-free control and the concentrations of DDE [1,1-dichloro-2,2-bis(p-chlorophenyl)ethylene], the main metabolite of the pesticide DDT [2,2-bis(p-chlorophenyl)-1,1,1-trichloroethane]), and polychlorinated biphenyls (PCBs) were compared among the groups of white, African-American and Asian women. Matched analyses found no significant differences in the concentrations of these organochlorine chemicals between cases and controls in any of the three exposure measures. In this nested case-control study, the quantitative information about (absence of) risk associated

with development of breast cancer at specific serum concentrations of DDE and PCBs could be incorporated into a quantitative risk assessment for these chemicals.

Cross-Sectional Studies

Cross-sectional studies often assess exposure and outcome at the same point in time. Thus, their major limitation is an inability to establish the temporal sequence of events unless additional information is available. However, many cross-sectional studies obtain historical exposure information adequate to determine temporality. An additional limitation is vulnerability to healthy worker effect/survivor biases because the exposure may affect disease incidence, disease duration, or both. Their value is that they can evaluate early indicators of a health effect (e.g., effect biomarkers), and if done serially, can provide longitudinal data on changes in effect biomarkers over time. Cross-sectional studies may be informative in evaluating pesticide exposures and potential health effects. For example, Farr et al. (2004) conducted a cross-sectional study to assess the association between pesticide use and menstrual function among 3,103 women living on farms in Iowa and North Carolina who participated in the Agricultural Health Study. Premenopausal women completed two self-administered questionnaires on pesticide use and reproductive health at enrollment. They reported exposures about lifetime use of pesticides and hormonally active pesticides. The characteristics of their menstrual cycles, including cycle length, missed periods, and intermenstrual bleeding were ascertained at the same time. After controlling for age, body mass index, current smoking status and occupational physical activity, the authors reported associations between pesticide use and longer menstrual cycles and increased odds of missed periods compared with women who never used pesticides. Women who used probable hormonally active pesticides had a 60–100% increased odds of experiencing long cycles, missed periods, and intermenstrual bleeding compared with women who had never used pesticides (Farr et al., 2004).

In summary, case-control, historical cohort and cross-sectional studies have significant advantages over ecologic studies in evaluation of individual level data on exposures, health outcomes and confounding variables and are thus much more useful in risk assessment. The extent to which data from these epidemiologic designs can be applied in a quantitative context will depend on the methods used for exposure assessment and the ability of the investigators to make relatively accurate and precise measurements of dose as well as the outcome. Well-designed case-control and historical cohort studies may have quantitative value in the risk assessment process, depending on their ability to establish dose-response relationships.

Question 2.3 The atrazine case study (Case study A) provides specific examples of ecologic and retrospective epidemiology studies. Please comment on OPP's reviews of the studies discussed in Case study A. In your comments, please provide specific feedback on the OPP's descriptions of each study design, exposure assessment, use of appropriate statistical methods, and ability to address bias and confounding in addition to other factors that may be important in the interpretation of these studies.

Panel Response 2-3

General Comments

In general, OPP has performed an accurate and thorough analysis of the five published studies included in Case Study A, and has captured most of the limitations of these studies.
A weakness of Case Study A is related to limiting the criteria for inclusion to studies that examined a reproductive outcome and were published since the 2003 IRED decision. This limitation had several undesirable effects. First, all studies in Case Study A used an ecologic or retrospective cohort design. Thus the Panel did not have the opportunity to explore how other epidemiologic designs that used cross-sectional, case-control, or other approaches might be incorporated in the risk assessment process. Second, the overall quality of these studies was relatively poor, thus limiting their applicability to the upcoming review of atrazine or the more general issue of incorporating epidemiology in risk assessment. Third, two of the five published studies used an ecologic design (Mattix et al., 2007; Winchester et al., 2009). As pointed out on page 61 of the draft framework and elsewhere in this report, ecologic studies are not useful for hypothesis testing and can rarely be used to establish exposure-disease relationships. At best, these studies might contribute to hazard identification, but better studies of atrazine and reproductive outcomes are available to meet the goal of assessing exposure-disease relationships.

Overall, the approach of the Agency to evaluating the epidemiologic studies provides a useful framework, and covers the important factors that need to be considered. However, some additional general factors should be addressed in the evaluation process. First, the background material needed for a full evaluation of these studies was lacking. For instance, background information on the target health effects would have been useful. It would be useful to compare the methods used to handle analyses below the limit of detection in the various papers, because this can affect summary statistics, associated confidence intervals and any statistical testing and modeling. It would also be useful to assess whether the endpoints are specific and whether their definitions are precise enough to distinguish a number of specific endpoints resulting from different modes of action. It would be useful to see whether any feasible modes of action have been identified as underlying the observed lesions. An analysis of whether the observed endpoints were compatible with the known reproductive effects of atrazine observed in animal studies was not included. Some background material on temporal and spatial aspects of the reproductive health effects would be helpful in an assessment of studies such as those evaluated in Case Study A. Background information on the reproductive outcomes evaluated; i.e., low birth weight, SGA, preterm birth and birth defects in the general population would have been useful in evaluating the case study. This information could include total incidence rates, incidence rates by maternal age and race, recognized risk factors for each outcome (tobacco and alcohol use, body mass index, nutritional status, prescription and over the counter medications,

recreational drug use etc), accuracy and completeness of reporting from birth certificates, examples of well-recognized associations with environmental exposures from epidemiologic studies, data from experimental studies in laboratory animals and other considerations.

An evaluation of background material on the seasonality in live births would be helpful because such seasonality is strongly influenced by socio-demographic factors (Bobak and Gjonca, 2001; Darrow et al., 2009). It would be worthwhile to compare the observed periodicity in exposed populations with that in other areas with limited exposure to toxicants of interest. The Agency should consider how this sort of background information could be used in assessing the quality and utility of epidemiological studies.

The Agency and the authors of the five studies in Table A-1 of the draft document missed potentially useful analyses related to the papers in this group. Some review papers going back to the mid 1960's considered the medical significance of date of birth (Bailar and Gurian, 1964; Kesselman and Bailar, 1964; Bailar and Gurian, 1965). Three forms of congenital malformations (congenital hip dislocation, spina bifida and other neural tube defects, and patent ductus arteriosus) showed a seasonal pattern. These were also unusual in being the only three that occurred more frequently in females. This work may be important because atrazine was not in use at that time in the sixties, and evidence for or against a seasonal pattern at that time might tend to help in the interpretation of more recent data.

The five studies presented in this draft report share several problems in common. All are reporting effects that are small in relation to the background "noise", so that unrecognized or uncontrolled forms of bias may be quite important. Each of these studies had numerous sources of potential bias. Some of them also have a problem of multiple comparisons, though it appears that none of the studies addressed the issue. Multiple comparisons may create difficulty in interpreting findings, because some associations may be due to chance. The interpretations, particularly of the individual-level studies (Ochoa-Acuña et al., 2009; Ochoa and Carbajo, 2009; Villanueva et al., 2005) had flaws, according to one reviewer. The sixth study (Mohanty and Zhang, 2009) is an ecologic study based on a slide presentation. This study has not been published in the peer-reviewed literature. The Panel recommends that abstracts from scientific meetings, presentations and other data sources that have not been subject to the peer review process should not be incorporated into risk assessments by the Agency.

The retrospective cohort studies by Villaneuva et al. (2005) and Ochoa-Acuña et al. (2009) used municipal drinking water concentrations of atrazine from monitoring data to assign exposure levels for the analysis. This method of exposure assessment may lack precision at the level of the individual for several reasons described below. However, retrospective cohort studies that measure concentrations of a chemical in the municipal water supply for exposure assessment have been used frequently in epidemiologic studies of drinking water contaminants and may be used in risk assessment when study quality is high. For example, several retrospective cohort studies were conducted to examine associations between exposure to the disinfection by-products (DBPs) that are produced during chlorination and reproductive outcomes (Dodds et al., 1999). Their usefulness was limited by the quality of the exposure assessment and specifically for reproductive outcomes, the ability of the investigators to estimate exposure accurately during the critical period of gestation for each member of the cohort. In the research on DBPs,

mandatory quarterly testing requirements at the plant and in the distribution system of municipal systems serving 10,000 persons or more in the United States and Canada provided a data base for estimation of exposure. Assigning concentrations to specific trimesters of gestation period was performed by regression methods in one study (Dodds et al., 1999). With less frequent sampling, the quality of the exposure assessment will be negatively affected if exposure varies over time.

The degree to which daily, weekly or seasonal fluctuations in concentration occur, potential variability in concentration throughout the distribution system and the stability of the pesticide in water must be considered to reduce exposure misclassification (Nieuwenhuijsen et al., 2000). Important factors in estimating exposures to cohort members include the potential use of bottled water for consumption, the amount of tap water consumed daily, the use of tap or home filtration systems, whether the system is flushed prior to obtaining drinking water and other behaviors such as consumption of water outside the home. The use of tap water for making soup, cold drinks, coffee and tea and juices needs to be considered in this pathway. Exposure to volatile chemicals such as the trihalomethanes will be reduced during heating of water for preparation of soups and heated beverages. Conversely, exposures to volatile organic chemicals, such as the trihalomethanes, through showering and bathing comprise an important exposure pathway (Gordon et al., 2006). Residential mobility during pregnancy has been shown to be relatively common, affecting up to 25 percent of women in some studies. Therefore, reliance on the address found on the birth certificate may not provide an accurate assessment of exposure if the participant moved across boundaries during her pregnancy to locales where the concentration of the contaminant would be expected to vary widely.

One Panel member was concerned that all non-statistically significant findings, including findings that showed a monotonic exposure-response relationship, were ignored in the Agency review. Insufficient attention was paid to exposure-response relationships, how the studies evaluated them, and whether the evaluations were adequate (e.g., were the categorizations of exposure simply based on percentiles, were meaningful categories used [e.g., an MCL], was the use of a continuous variable in a logistic regression or binomial regression model warranted (i.e., to what extent did the model selected adequately characterize the likely shape of the dose-response curve?). Another Panel member felt that an over-emphasis was placed on confounding bias in the draft framework and that confounding rarely has a big impact in epidemiologic studies. This Panelist also felt that an under-emphasis was evident on the bias most likely to have a big effect on the estimation of the magnitude of effect and the exposure-response relationship, i.e., exposure misclassification. Further, inadequate attention was paid to the statistical power of the studies and to the impact of disease misclassification due to factors such as under-ascertainment of birth defects from information on birth certificates, and the grouping of heterogeneous birth defects that may have differing etiologies.

In the evaluations of the papers it might be beneficial to place some emphasis on the structure of the hypothesis, and on the validity of the exposure measurement. These issues were raised in comments by a number of Panel members. In evaluating the atrazine epidemiologic studies, the Agency needs to be consistent in its assessment of the possibility of concomitant exposure to other triazines and their breakdown products which have the same mode of action according to U.S. EPA's cumulative triazine risk assessment. (U.S. EPA, 2006)

One Panel Member commented that overall, the evidence presented in the case study on recent epidemiologic findings on the association between atrazine and birth outcomes was weak and inconsistent and seemed quite compatible with no effect of atrazine on birth outcomes. This individual felt that the collection of data could serve as a good example of a "negative" dataset for use in developing guidelines for the interpretation of human epidemiologic and incident data in health risk assessment. Others felt that the quality of the studies included in the Case Study was not adequate to reach any conclusion regarding potential associations between atrazine exposure and adverse reproductive outcomes. A more complete record of published papers, and or a more complete search of the literature, may or may not reverse this conclusion.

Conversely, another Panel member offered the opinion that the ecologic studies cited by the Agency may not suffer from the severe ecologic fallacy that the Agency identified. Information presented during the meeting, including analyses presented by public commenters, weight the evidence toward the environmental contribution to birth defect incidence. It was clear to this reviewer from the presentations by the Agency and the public comment (Syngenta Panel slide #44) that birth defect incidence in the US is strongly affected by seasonality both in states with and without heavy atrazine usage. This seasonal pattern coincides with the use of atrazine in those states during late spring/early summer, gradually diminishing in the fall and winter seasons that could be explained by the half-life of atrazine in the environment. This reviewer commented that the national birth defect incidence data compiled by CDC sheds light on the association of birth defect and atrazine concentrations in water (both surface and finished drinking water) in the states where atrazine has been heavily used.

Ultimately, the Agency needs to pose the question: to what degree does the epidemiologic studies decrease uncertainty associated with extrapolation from animal studies for the protection of human health? The Agency should use their answer to the question to approach uncertainty factors and the inherent limitations of observational research accordingly.

Specific Comments Regarding Case Study A

Mattix et al., 2007

The data in the paper by Mattix, Winchester, Scherer (2007) contain a gap from 1990 to 1995-2002. Two sources of data were available; no comparison was made between the CDC data for Indiana and that reported by the state. If serious discrepancies are apparent between the two sources, then one or the other (or both) must be wrong, and it would be important to find out which, and why. In the Agency's assessment of bias, confounding and other factors, the figures presented in the bottom, left-hand column on p. 948 suggest that, of the abdominal wall defects (AWD) occurring in the Riley Hospital (279 over 1990-2002) fewer than half (133, 48%) were simultaneously identified by the State Registry. This low rate of capture of AWD incidence by the State Registry raises a question about state-to-state differences in the ascertainment of AWD incidence and whether the higher rate noted in Indiana might not be due to a higher than national average rate of capture of AWD incidence. The state registry information is based on birth certificates which often fail to capture many birth defects; AWD is one that is often omitted. The critical issue with using data reported on birth certificates is not whether they contain errors, but

whether the errors are differential between states. If reporting across states is relatively constant over a specified period of time or location, then analysis of those times or locations may provide relatively useful data.

One reviewer commented that the strength of findings presented in this paper should be reduced to account for the "multiple comparisons problem". The authors noted that the elevated Indiana rate was statistically significant only in 1996, 1998, and 2001, but a critical question is whether the reported rates for all years are statistically comparable. In short, was statistical power great enough to say that an effect was present (or greater) in some years than in others, or are we just looking at the effects of having small numbers of AWDs in each year? Were any features of atrazine use unique during the higher-incidence years? It appears that the data in Figure A-2 were not adjusted for nitrates.

Winchester et al., 2009

One reviewer commented that the relationship shown in figure A-1 is far from striking, especially when one views this in the context of the overall scale of cases (per 100,000), a maximum variation of perhaps 6% is observed. This is even more concerning due to numerous possible season-related confounders. The peak incidence in terms of last menstrual period (LMP), roughly, date of conception, was May-June. The data would be more convincing if the authors had found a lack of such a pattern in mothers who had been drinking ground water. Also, the text provided no evidence that the authors adjusted for other seasonally changing chemical exposures, nor did they look at concurrent data from other states with lower atrazine exposures to see whether the reported patterns were unique to atrazine exposure. It was appropriately noted in the Agency's critique that chemical concentrations were measured in surface water and not drinking water, and that these were population-level data based on rates.

According to one reviewer, as shown in Tables A-2 and A-3, all but one of the birth defect types occurred more frequently in April-July than in other months, and the exception ("Nervous", not further specified) barely fell below a ratio of unity. About half of the differences were statistically significant. However, chemical teratogens tend to be more specific with effects targeted to the organ that was in a critical stage of development at the time of exposure. The lack of specificity and broad pattern of the evidence suggests a pervasive bias related to some other seasonally changing factor.

Ochoa-Acuña and Carbajo, 2009

The retrospective cohort study of limb defects by Ochoa-Acuña and Carbajo (2009) identified birth defects among 48,216 singleton births between 2000 and 2004 in rural Indiana. Although not stated explicitly, it appears that the investigators compared cases of birth defects with unaffected infants to calculate odds ratios for exposure to cornfields or soybean fields within 500m of the residence. Unfortunately, the exposure analysis relied on proximity to these fields and as an exposure surrogate and was not validated by other methods.

For some birth defects, e.g., neural tube defects (NTD) and abdominal cavity, the statistical power to detect meaningful differences was low. The category "heart defect" combines defects

with very heterogeneous etiologies and should be analyzed, if sufficient numbers of cases exist, by subgrouping (e.g., conotruncal heart defects). The study controlled for the variable "farm exposure", i.e., the percent of cropland around a home, which likely led to a bias toward the null. This factor is part of the exposure of interest and not a confounding variable. If distinguishing farming exposure effects from "drift" to bystander populations is a concern, then the analysis could be stratified by this variable. The authors noted that when the models did not include this variable, respiratory defects "appeared increased".

In the evaluation of this study, the statistical methods were described but not evaluated. Limb defects showed an odds ratio with a 95% CI above 1.0 for <3.4 ha vs. >3.4 ha. Other defects also had high odds ratios, but the lower limit of 95% CI fell below 1.0. Elevated odds ratios were observed for several birth defects and soybeans, which were not discussed because they were not statistically significant. The strongest relationship in Table 3 was the finding for NTD (odds per unit increase in exposure = 1.72) and soy area, but was not discussed. It would have been useful to have comments on whether additional or different statistical methods could have been used to analyze the data. For instance, it may have been better to divide the crop areas into three or four categories rather than using an exposed group and not-exposed group comparison. The methods used may have been too conservative. Multiple comparisons were performed; this is of concern because the effect estimate is modest and the confidence bounds on the adjusted ratio for fields of corn (OR = 1.22, 95% CI: 1.01, 1.47) barely exclude unity. The odds ratio for soybean fields does not suggest an effect (OR = 1.04, 95% CI: 0.85, 1.28). The use of a continuous variable in an exponential model (Table 3) is problematic because it assumes an exponential increase in risk for each increment of exposure (Hosmer and Lemeshow, 2000).

Villaneuva et al., 2005

This is a retrospective cohort study with exposure assigned to individuals and covariates measured at the individual level (not an ecologic study as designated by the Agency). As shown in Table A-6 of Attachment A, a non-significant association was present for preterm delivery and exposure to atrazine in finished water which was not discussed by the authors or the Agency. An exposure-response relationship was observed (ORs 1.0, 1.22, 1.93 across the tertiles), and the finding in the higher exposure group (i.e., OR = 1.93) was stronger than any other finding in the study. A monotonic exposure-response relationship was apparent; therefore, this finding should have been thoroughly examined. In table A-7 the ORs for first trimester exposure and preterm delivery (1.36, 95% CI 0.95, 1.95) and third trimester exposure and SGA (1.37, 95% CI 1.04-1.61) are nearly identical. The finding for SGA is statistically significant due to larger numbers of cases. This example demonstrates the limitations in relying solely on p values in interpreting epidemiologic data.

The Agency's assessment of the exposure data included the issues of using drinking water concentrations averaged across several years, while the birth effects were for a single year, the relatively low concentrations of atrazine, and the fact that the concentrations of atrazine in the three exposure groups were not adequately heterogeneous to discriminate potential associations. The latter point was also discussed by the authors. Breakdown products of atrazine were measured in the study, which constitutes a strength in the exposure assessment for that

investigation. Exposure was transformed into geometric mean values without any explanation. If it was because of skewness in the distribution, then it was inappropriate to use this parameter because it is the high-exposure points that are of concern, and it is counter-productive to reduce their impact on the analysis. Only one year was examined, and so any possible year-to-year patterns cannot be studied. No analyses were included of possible covariates correlated with distribution units, such as ground vs. surface water, or local contamination by known sources of toxic chemicals. The authors were thorough in their identification of potential bias and confounding factors, which included selection and recall bias, maternal smoking and alcohol consumption, exposure to agents with known effects on abnormal birth parameters (disinfection byproducts, air pollution, PCB's and lead) that may have seasonal patterns that coincided with that of atrazine, and atrazine exposure other than by drinking water.

Table A-9 contains a statement that the study by Villaneuva et al. (2005) is limited because it was conducted in France and the results may not be generalizable to the U.S. The rationale for OPP's statement is unclear.

Ochoa-Acuña et al., 2009

In this study of atrazine in drinking water systems in Indiana, study data were adjusted for season. This adjustment may have led to a bias toward the null if the seasonal effect on SGA or preterm birth was partly due to atrazine exposure. The study used a continuous variable in an exponential model, which makes the strong assumption that the risk increases exponentially with each increment (log atrazine level) of exposure. Moreover, use of a log transformed exposure variable may not be appropriate for characterizing the exposure-response curve. The categorization uses percentiles which may or may not be appropriate to characterize the exposure curve. The SGA effects were statistically significant although small. They fall in the range of 1.06-1.2.

In the study by Ochoa-Acuña et al. (2009), the availability of atrazine sampling data at 7 to 14 day intervals constitutes a strength of the exposure assessment. Some issues involved with interpolation of atrazine concentrations and the quality and comparability of the four drinking water monitoring programs were addressed. However, a weakness in the exposure assessment was that estimates were based on sparse data, especially for the winter months. The latter is critical since it is this period that comprises the unexposed months of gestation.

Roughly 70% of the birth records available to Ochoa-Acuña et al. (2009) came from one mid-sized community, which raises questions about selective effects on reporting. It is not clear why Fort Wayne predominated in the data, or whether unmeasured confounders occurred more frequently in Fort Wayne.

Weak evidence of an association between atrazine exposure and SGA was detected for exposures in the third trimester and "entire pregnancy". No association with preterm delivery was evident. LBW was not reported in this part of the analysis. The range of the confidence bounds was smaller for SGA than for preterm delivery with larger sample sizes available for the third trimester and entire pregnancy than for the first or last month of gestation in the preterm delivery analyses.

One Panelist introduced several additional factors to be considered in the WOE analysis, particularly with respect to the ecologic studies in Case Study A:

a. Window of susceptibility – In the atrazine case study investigators looked at either the association of atrazine concentrations in surface water and the months of LMP in relation to the national birth defect data (Winchester et al., 2009), or the association of atrazine concentrations in surface water during the 3rd trimester and SGA (Ochoa-Acuña et al, 2009). This Panelist felt that the data presented provide a convincing link between birth defects and atrazine concentrations in surface or drinking water at the critical points of time during development. If the window of susceptibility for a birth defect is not taken into account (as presented in a public comment) the correlation between month of LMP and atrazine concentrations in water disappeared.

b. Longitudinal or temporal variations of atrazine exposure and the correspondence with birth defect outcomes – In the case of atrazine exposure and birth defects, temporal variation of atrazine concentrations in either surface water or drinking water is critical for assessing birth defect risks. If atrazine concentrations in water were to remain constant throughout the year while birth defect incidence rates varied, the WOE of this association would be non-existent. The converse is also true.

c. Evidence for protecting public health. – One of the missions of the Agency is to safeguard the public from unnecessary pesticide exposures. In the opinion of this Panelist the evidence for an association between atrazine and birth defects is adequate to consider it as "Some Evidence For", and thus it would be prudent not to dismiss it. As this framework is developed and evolves, and better quality epidemiologic and/or incident data become available, the Agency will be able to change the designation toward or away from "Some Evidence For" with greater confidence without endangering public health.

Question 2.4 In light of scientific issues discussed in Questions 2.1-2.3, OPP requests input from the SAP on factors to consider when integrating these studies in the atrazine WOE analysis currently under development.

Panel Response 2.4

Of primary concern to the Agency in integrating the results of epidemiologic studies in the atrazine WOE analysis is an assessment of study quality. The criteria for assessing the quality of such studies are discussed in detail in response to question 2.1. Study quality must be a key component in selection of studies to incorporate into the analysis. An important issue is how the Agency decides whether to use particular sets of data. It is not uncommon for the Agency to be criticized by some experts for *excluding relevant* data from their risk assessments and criticized by others for *including poor quality data* in the same risk assessment. The Agency should establish a set of criteria for determining the acceptability of epidemiologic studies. These criteria may be based on quantitative criteria, scientific judgment, or some combination of these. Inevitably, it will be necessary to exercise some degree of scientific judgment in this assessment. The Panel recommends that epidemiologists participate actively in the process. Observational research is subject to potential error due to the nature of the science. However, the presence of uncertainty in epidemiologic research does not necessarily imply that the study cannot be used. In practicing their "art" epidemiologists make judgments about the extent of potential biases, such as an inability to measure exposure precisely or to arrive at valid estimates of dose. They also make judgments about the probable direction of these uncertainties, i.e., whether the misclassification of exposure is likely to have biased the risk estimate toward or away from the null. In particular, epidemiologists consider the possibility that exposure misclassification may have biased a dichotomous categorization of exposure toward the null and distorted an exposure-response relationship or that differential misclassification, e.g., recall bias biased the risk estimate away from the null.

The interpretation of new information about the effects of pesticide exposure must be considered carefully. If a novel observation is made in a study, but the decision was made not to use the study based on other criteria, the data should be archived "pending further investigation". This will add transparency regarding the use of data and preserves the possibility that novel data will be resurrected and used if corroborated by additional studies. Epidemiologists generally believe that no single observational study should be considered "definitive" and that the findings from well-conducted studies still require confirmation in other populations (the consistency criterion of Bradford Hill). Further, novel findings should have a biologically plausible framework if they are to be considered in the WOE. The possibility that individual statistically significant associations may be due to chance should not be ignored by the Agency or other reviewers.

Studies demonstrating no association with a pesticide exposure are equally informative in a WOE analysis as those that do so provided they meet the criteria for quality described above. Publication bias resulting from rejection or failure to submit 'null' studies is of some concern. However, several epidemiology journals, such as Epidemiology, have an explicit editorial policy of not rejecting 'null' studies on the basis of null findings alone when the study is otherwise well-conducted and address an important health concern. Publication bias is addressed in detail in the literature for meta-analytic epidemiologic analysis and methods, such as the use of funnel

plots, for detection of potential publication bias are routinely used in meta-analyses (Greenland, 1998).

The process for selection of studies in the WOE analysis for atrazine or more generally for incorporation in risk assessment begins with a comprehensive literature search to identify the full array of available studies. In conducting a WOE analysis the investigators should be assured that they have accessed the complete body of relevant epidemiologic literature available from peer-reviewed sources. A plan for the literature search should be developed that incorporates second and third level searches in the published literature as well as using the standard approaches of literature searching such as PUBMED (http://www.ncbi.nlm.nih.gov/pubmed/).

Careful evaluation of the findings should be performed by trained epidemiologists. For example, if drinking water studies are under evaluation, then epidemiologists experienced in conducting these kinds of studies, and scientists with expertise in water modeling and drinking water exposure assessment should be asked to review them. Similarly, review of occupational retrospective cohort studies should be conducted by researchers familiar with the issues of occupational exposure assessment and the statistical methods used for analysis of this form of cohort study. The field of epidemiology is diverse; therefore, those who conduct reviews and make expert judgment regarding inclusion of studies and eventual use of the data should have the training and experience required to do so.

Epidemiologic studies of reproductive outcomes have substantial strengths as summarized previously (Savitz and Harlow, 1991). As widely recognized, the fetus represents a susceptible subset of the population that may be exquisitely sensitive to the effects of environmental contaminants. The events that encompass conception and gestation and the exposures that may affect the processes of implantation, development and growth of the fetus occur in a relatively short time frame of one year or less. Thus, in the evaluation of potential adverse effects of *in utero* exposures to atrazine and other pesticides, epidemiologic studies focus attention on the critical temporal windows of exposure for each outcome; for example a single birth defect, spontaneous abortion or growth retardation. The truncated time frame of interest provides opportunity for more precise exposure assessment and reduces the probability of recall error when questionnaires are used to obtain information from parents. In studies of reproductive effects of disinfection byproducts, investigators were able to focus the analysis on exposures during specific months and weeks of late gestation in assessing associations with low birth weight, intrauterine growth retardation and pre-term birth (Hinckley et al., 2005).

When animal data show that a pesticide affects pathways essential to human reproduction and thereby establish biological plausibility for an effect, the Agency should examine the full suite of endpoints that may be perturbed (Moses, 1994). Epidemiologists have studied human fecundity by using time to pregnancy as a marker of success (Baird et al., 1986). This marker has been applied to pesticide exposures (Thonneau et al., 1999) and was described in early studies of agricultural workers exposed to dibromochloropropane on banana plantations in Costa Rica (Whorton et al., 1979).

Spontaneous abortion is an endpoint frequently examined in human studies of chemical exposures. For example, Arbuckle et al. (2001) studied the effects of pesticide exposures on the

risk of spontaneous abortion in a Canadian farm population. Studies of spontaneous abortion can be limited by the introduction of selection and reporting bias (Wilcox et al., 1984) and by the fact that in 20 to 25% of reproductive failures fetal loss is not manifested clinically (Wilcox et al., 1988). Late fetal loss, after 20 weeks of gestation can be assessed using data on stillbirth and neonatal mortality. Perturbations in fetal growth incorporate studies of birth weight (continuous variable), low birth weight (< 2500 grams), very low birth weight (< 1500 grams) and preterm delivery (37 weeks of gestation). Intrauterine growth retardation or small-for-gestational-age is assessed by comparing the infant's birth at a specific week of gestation to the norms for that racial/ethnic group using the 5^{th} or 10^{th} percentile as the basis for designating each birth. Examination of fetal growth parameters is frequently performed by analysis of birth certificate data that can be accessed readily and linked to environmental exposure data for the cohort or case and control samples.

The hormonal control of processes such as the onset of menarche, the patterns of menstrual cycle activity and the menopause provide the biologic framework for epidemiologic studies of these endpoints vis-a-vis pesticide exposure. Examples of studies which assessed associations with atrazine are available in the scientific literature (Farr et al., 2004; Farr et al., 2006). These studies should be incorporated into the WOE analysis.

Finally, it is widely recognized that approximately 50 percent of impaired fertility in humans is attributable to the male. Therefore, studies that assessed exposure to atrazine and semen quality (sperm concentration, percent motility, percent abnormal sperm) are important components for the WOE analysis. In particular, the study of Swan et al. (2003) should be informative. Some of the studies of reproductive outcomes that were not included in Case Study A should be relevant to the Agency's review of atrazine later this year.

Several issues in analysis and interpretation of findings from epidemiologic studies are often the source of discussion among epidemiologists, biostatisticians and others. First is the consideration of the interpretation of statistical significance and the sole reliance on the use of the p value for decision making (e.g. < 0.05). A series of papers exists in the literature in which epidemiologists and others have made a strong case for interpretation of the precision of risk estimates using confidence intervals in lieu of a strict interpretation of the p value (Savitz, 2003; Rothman and Greenland and Lash, 2008). It is important to evaluate all findings that show an elevated or reduced risk estimate or an exposure-response relationship regardless of statistical significance. Reviewers should consider the likelihood that the study lacked adequate statistical power and if needed, conduct the appropriate power calculations to assess the magnitude of risk that could have reasonably been expected to be detected if a true association existed.

A second issue that has been widely discussed by epidemiologists is that of multiple comparisons and the possibility that some findings in epidemiologic research may be due to chance as a result of multiple statistical tests. Epidemiologists warn against making adjustments to p values or confidence intervals that are inappropriate, overly conservative and wasteful of information (Rothman, 1990; Savitz and Olshan, 1995; Savitz, 2003; Rothman, Greenland and Lash, 2008). In considering this issue, reviewers should distinguish between analyses that incorporate multiple exposures and or outcomes in searching for any and all associations from those that explore *a priori* hypotheses in databases that permit multiple analyses to be conducted. The results of a

study should not be discounted simply because it efficiently and comprehensively evaluates multiple outcomes and multiple exposures (or multiple exposure indices). Several approaches to the multiple comparisons issue are discussed in detail in the literature (Steenland et al., 2000; Rothman, Greenland and Last, 2008). Multiple inference procedures involving hierarchical models are useful if the research interest concerns a joint hypothesis (i.e., a "family" of similar or "exchangeable" exposure-disease associations) or the purpose is simply exploratory (e.g., to answer the question, "Which, if any, of a "family" of exposure-disease associations should be followed-up in future investigations?"). However, in most instances the research question concerns a separate, single exposure-disease comparison or hypothesis. In the latter case, each comparison should be evaluated as if it were the only comparison in the study (this approach is also appropriate when there is doubt about whether the research interest is in a joint comparison or single comparisons – see Rothman, Greenland & Lash 2008, page 237).

Epidemiologists recognize the problems inherent in such analyses and the possibility of chance findings. Epidemiologists are cognizant of the inherent problems in conducting multiple statistical analyses and are trained to interpret these findings carefully and to employ many of the guidelines suggested in the Bradford Hill criteria discussed elsewhere in the report. The strength of an association, presence of a dose-response relationship, consistency of the finding across studies, coherence with available biologic information and other criteria are routinely employed in interpretation of data. Reviewers should also consider the reproducibility of observations among studies in terms of the direction of the effects observed, the magnitude of the effect and the concentrations at which these effects occur. The latter consideration is often ignored. In addition, the criterion of specificity of effect in the original Bradford Hill criteria has largely fallen into disfavor due to the systemic effects of many environmental exposures. Furthermore, the lack of evidence regarding biologic plausibility is not sufficient reason to discount or ignore that the remaining criteria may constitute sufficient weight of evidence in assessing an exposure-outcome relationship.

Epidemiologists often categorize exposures categorically, using tertiles or quartiles to evaluate potential dose-response relationships. Cutpoints for such analyses are typically developed from exposure data from the unaffected members of the cohort, or in case-control studies from the distributions among controls. Incorporating these studies in the WOE analysis will require considerable thought since the model departs from the traditional examination of linear dose-relationships. The shape of the dose-response curve in humans may not be linear if a threshold exists below which the chemical has no effect. The Agency should examine dose-response models and human data for other chemicals (non-carcinogens) to determine the optimal methods for integration of categorical exposure data from human studies into the WOE analysis.

Agency Charge

3. Case Study C: Human Incident Data-- Retrospective Case Study Using Diazinon

EPA is undertaking an effort to more systematically and transparently review and use human incident data in risk assessment/characterization or in problem formulation than has been done previously. As part of this effort, a case study using human incident data on diazinon is included.

Question 3.1 Case study C describes various analyses and evaluations that can be conducted when evaluating human incident data. Please comment on ability to use incident data for the following types of analyses: trend of incidents over time, frequency of reported symptoms, common product clusters, frequency of repeated exposure scenarios, and assessment of children vs. adult symptom profiles), in the diazinon case study and suggest alternative and/or additional analyses, if appropriate.

Panel Response 3.1

The majority of the Panel members agreed that little weight should be placed on self-reported incident data in routine risk assessment. Although human incident data can sometimes be useful in providing information on trends or differences in the frequency and severity of symptoms and whether human effects are consistent with those observed in toxicologic experiments or epidemiologic studies, the limitations of using human incident data for risk characterization and risk assessment outweigh the advantages. The major limitations include: 1) likely under-reporting of cases due to the lack of mandatory reporting other than for registrants; 2) uncertainty regarding the exact exposure conditions; 3) capture of largely only acute events and not events with long latent periods or events associated with long-term exposures; and, 4) the applicability of self-reported human incident data only to pesticides with notable acute toxicity.

The diazinon case study, as presented by the Agency, is unique because of the distinct symptoms resulting from cholinesterase inhibition and because of the risk mitigation measure of removing diazinon from residential use and the consequent reduction in incidents. Other pesticide groups, such as the triazine herbicide family, that do not produce symptoms of acute toxicity would probably not generate usable incident data for the following analyses.

Trend of Incidents over Time

Incident data have value in assessing effects resulting from changes in use patterns and implementation of use restrictions, and thus can serve as a good measure of the success of risk management procedures for minimizing acute toxicities. For diazinon, the reduction of reported incident cases appears to reflect its restricted access to the general public. It is unclear, however, whether exposures at lower levels would trigger incident self-reporting and whether these incidents would also have been reduced. For pesticides that do not pose marked acute toxicity potential, incident data are presumed to be sparse and inconclusive and of limited use in risk characterization/assessment.

It is worth noting that the different incident reporting systems seem to be generally reliable, particularly for evaluating frequency of incidents over time, such that all five systems showed a relatively similar decrease in incidents over time with the removal of diazinon. The data are also collected in a relatively uniform manner (e.g., product information, severity rankings and symptoms) among the different data sources. However, increased communication and coordination among the different reporting systems to make the data collection instruments more uniform could improve the collective data generated.

Frequency of Reported Symptoms and Frequency of Repeated Exposure Scenarios

The reporting of similar symptoms following exposure to a particular pesticide product by different individuals within a defined time period should raise concern for the use pattern of this particular product. Reporting of common symptoms by exposed persons may also indicate a pattern of acute toxicity previously undetected in experimental studies or in reports by individual registrants. Under circumstances in which incident data reveal a health outcome that was not previously observed in toxicology or epidemiologic studies, such human incident data could be valuable in terms of exploring biologic plausibility associated with specific pesticide exposures. However, it is very likely that although these incident reports can be effectively collected, the available data may not adequately discriminate between high and low level exposures.

Common Product Clusters

Common product clusters resulting from incident reports occur, but are more important as an immediate public health concern, rather than serving a risk assessment purpose. For instance, methyl parathion poisoning cases in the southeast in the 1990s resulted from misapplication of the product. Such incident data would be inappropriate for consideration in risk analysis. This is also true for cases of abuse or suicide in which the data would not be relevant for risk management or risk assessment because the exposures for those abuse/suicide cases would be expected to exceed label recommendations or be by ingestion. If clusters of incidents point to a risk management failure, then certainly incident data should be used to protect those individuals who might be at unanticipated elevated risk.

One Panel member thought that a cluster is likely to indicate a more severe problem than an isolated case, but not more severe in proportion to the number of persons reported. This is because reporting by one individual is likely to stimulate reporting by others, so that a cluster is artificially created in excess of what would happen if reporting were independent. However, the reporting of clusters has contributed to the risk assessment process, for example in the aldicarb contaminated watermelon episodes in California (Goldman et al., 1990). In this instance, the reported cluster played an important role in the risk assessment for this particular pesticide.

Assessment of Children vs. Adult Symptom Profiles

Comparisons of the distributions of symptoms in children and adults can provide supportive evidence of similarity of effect, but lack of similarity does not necessarily mean that the mechanisms are different because they could reflect: 1) different levels of sensitivity of reporting (e.g., effects in children may be more likely to be reported than similar effects in adults); 2) different routes of exposure; and 3) different sizes of populations exposed (e.g., small numbers

exposed persons might result in less certainty in the distribution of symptoms). It is also likely that the magnitude of exposure (or dose) to certain pesticides that would trigger the reporting of the incident would be very different between adults and children.

Other limitations of using human incident data for risk characterization and assessment further reduce its utility. For instance, follow-up on these incident reports is typically minimal, limiting the information on possible long-term consequences. Since more subtle, long-term effects are typically more difficult to detect in animal studies, better follow-up of reported incidents may be beneficial in that regard. It is also very likely that self-reported incident data may consist of anecdotal or emotional observations that have limited factual evidence of connection to a specific exposure. Because the quality of the self-report incident data is extremely difficult to determine, it will no doubt introduce bias and uncertainty in future analyses.

Overlap of self-reported cases among the five different incident databases is also a concern. It is unknown whether the overlap is concentrated on severe poisoning cases or in certain geographic areas. It appears that these sources of data are studied independently until the later stages of analysis, when results are compared across databases to identify signals of a problem. It is, therefore, necessary to consider how these five sources might be used in combination at earlier stages, not necessarily by matching cases, but at least by organizing the data in ways that draw on the strengths of the various sources. This might be accomplished by focused study in one or two areas where three, four, or even all five reporting systems operate.

One Panel member recommended that the Agency should clarify the presentation of human incident data contained in Attachment C with respect to how the data were compiled for the tables in Appendix B. In addition, several other Panel members suggested that before further considering the utility of human incident data, potential confounders or other exposures that may have been responsible for the symptoms reported should be identified and controlled in the analysis.

Question 3.2 OPP plans to conduct analyses of human incident data like that described in Case study C for other pesticides undergoing registration review. In light of scientific issues discussed in Questions 3.1, OPP requests input from the Panel on factors to consider when evaluating the reliability of human incident data and determining the relative weight that should be placed on such data in risk assessment/characterization or in problem formulation.

Panel Response 3.2

In general, very little weight should be placed on incident report data in routine risk assessments because of their diverse nature with regard to estimated dose levels, product characteristics, and the ability of the observer to assess symptoms accurately. If the numbers of incident reports are large, the exposures are well-estimated and the symptoms are highly consistent, then perhaps incident data would be useful. If incident clusters point to a risk management failure, then certainly incident data should be used to protect those individuals who might be at unanticipated elevated risk. In cases of abuse or suicide, the data would not be very helpful for overall risk management because these exposure levels would be well beyond label recommendations or by another pathway such as ingestion. Reports that contain vague or subjective information, including flu-like symptoms or those that could arise from physiological stress should be interpreted with caution. These reports could represent general symptoms from a variety of illnesses or conditions including infectious diseases or stress. It may be impossible to distinguish pesticide effects from other conditions that could mimic those due to exposure to the product. Fear of poisoning could lead to neurobehavioral symptoms, with the typical reactions associated with stimulation of the sympathetic nervous system. The diazinon case study was uniquely suited for such an analysis because of the well-defined set of acute symptoms due to its anticholinesterase activity. The availability of data before and after introduction of risk mitigation in removing diazinon from residential uses and the consequent reduction in incidents were also relatively unique. Most other pesticides would probably not be adaptable to such a clear presentation.

The incident reporting systems described in Attachment C seem to be generally reliable. This is particularly clear when evaluating frequency of diazinon incidents over time, such that all five systems showed a relatively similar drop in incidents over roughly the same time period. Considering the lack of specific training for persons recording information and the non-specific nature of some symptoms, relatively good reliability was observed among the reporting systems for different classes of signs/symptoms associated with reported diazinon exposure. The recognized strength of this type of data, in contrast to information from animal toxicity studies, is that responses in humans are detected under real-life situations, with conditions of differential individual sensitivity, modifying factors and other influences possible in the human population. A major weakness of this type of information, however, is the uncertainty regarding the exact exposure conditions, concentration, or amount of the chemical to which the individual was exposed. Even the specific chemical may not be known with certainty, and it is likely that no information exists on inert ingredients or co-exposures that may have been involved. Follow-up for these incidents is typically minimal, limiting the information on possible long-term consequences. As more subtle, long-term effects are typically more difficult to detect in animal studies, better follow-up of reported incidents may be beneficial in that regard. As noted in the draft framework, the various incident reporting systems may report different symptoms and signs

(or the same signs called something else), or differential severity rankings for symptoms of toxicity. Considering the weaknesses of this type of information, the weight given to use of incident data in risk assessment process should be low, with a more qualitative than quantitative influence on the process. Surveillance for unanticipated effects in incident reports could be useful in suggesting alternative mechanisms of action or toxicities not previously described for a pesticide.

The limitations of the incident data for diazinon out-weigh the possible benefits of the use of such data for risk assessment/characterization. One possible enhancement to the self-reported incident data would be to implement the collection of appropriate specimens or samples, where feasible, from individuals who call in to report symptoms in the future. Laboratory analyses of such specimens and sample would serve to validate the reported human incident data and also provide critical information about the levels of exposure (dose) that are responsible for symptoms among exposed individuals. Such data would also be useful in differentiating symptom profiles and exposure levels in children versus adults. Although logistic issues, costs and feasibility of implementing specimen collection may be currently beyond the Agency's capability, the idea could be discussed further among the agencies collecting human incident data. Perhaps a limited pilot study may be feasible.

Although reliable data are generally lacking for most case studies, the Agency is encouraged not to overlook the rare cases with sufficient documentation, or clinical case reports published in the open literature with extensive follow up after poisoning. Some of these reports may uncover new toxicity endpoints of concern and should be added to risk assessment. For example, chronic neuropsychological sequelae were manifested among those who appeared to recover from cholinergic signs and symptoms after acute organophosphate pesticide poisoning which involved a different mode of action (MOA) than cholinesterase inhibition.

The Panel concluded that incident reporting data such those considered in the diazinon Case Study (Attachment C) have some value for problem formulation and hazard identification in the risk assessment process, but their application in risk characterization is very limited unless follow-up information and or laboratory data from individual incident cases become available.

Agency Charge

4. Case Study B: The Agricultural Health Study Comparison of Exposure Assessment Approaches

The Agricultural Health Study (AHS) is a large long-term prospective epidemiological study that is collecting data on the health and work practices of licensed pesticide applicators in Iowa and North Carolina. The AHS is focusing particularly on the exposure of applicators to 50 chemicals, including many of the most widely used pesticides. The study also collects information on other possible agricultural exposures, and many lifestyle factors. Investigators with the AHS have published over 100 publications on a variety of topics including characteristics of the cohort and cancer and non-cancer health outcomes that have been observed in the cohort (http://aghealth.nci.nih.gov/).

Question 4.1: The Agency believes prospective epidemiology studies with robust exposure assessment, like the AHS, have the greatest potential for use in risk assessment especially for enhancing problem formulation and risk characterization. Please comment on appropriate ways to use of these types of epidemiology studies in risk assessment.

Panel Response 4.1

General Considerations

The Agency is urged to review other situations where epidemiological data have been used in risk characterization, e.g., arsenic, as these may prove useful in developing the framework. Considerations and review of data in risk assessment of chlorpyrifos (SAP, 2008) may provide a case study of how epidemiological data can be used in risk characterization. In this evaluation, a weight-of-evidence approach was used in the final determination. In the case of developmental neurotoxicity from chlorpyrifos exposure(s), prospective epidemiologic studies, with individual measures of chemical exposure, suggested that the dose-response relationship may be much different in humans than in animals. These prospective studies suggested neurodevelopmental effects may occur in humans with early exposures to chlorpyrifos, but with possibly different types of neurodevelopmental outcomes, and at potentially much lower levels of exposure, than in animal studies. The Panel concluded overall that data from both epidemiologic and animal studies suggested a connection between chlorpyrifos (and possibly other chemicals with anticholinesterase activity) and neurodevelopmental outcomes, but that dose-response relationships, and even mode of action, may not agree between these different ways of "looking" at end effects. One caveat to the conclusions was that the several anticholinesterase agents would have been acting on the same target enzyme, so sorting out the impact of any single compound would have been extremely difficult. This same concern also is present for the AHS in that any one pesticide may be present concurrently with one or more others acting on the same target system, so conclusions need to take the mixture into consideration.

The eventual resolution of large discrepancies between epidemiologic and animal studies in apparent dose-response relationships, or substantial differences in types of responses between animal and epidemiologic studies, is unclear from the draft framework. How does the weight-of-

evidence concept rationally weigh-in on decision making in this type of situation? On page 31 of the draft framework, it is stated that "when animal and epidemiological data do not provide a consistent toxicological picture…more weight would likely be given to those studies with robust study design and availability of replication or confirmatory data". Further, it asserts that "in most situations, the epidemiological study may not be sufficiently robust for deriving quantitative risk assessment values". If epidemiologic data are used to derive quantitative values in risk assessment, determining a process for decision-making in cases in which wide differences are observed in dose-response relationships between animal and epidemiologic studies could clarify the framework and its implementation. At present, it can be argued that in most cases animal studies can more "robustly" describe dose-response relationships, using the least amount of time/resources, etc., and therefore may currently provide a more reasonable approach for characterizing dose-response relationships, for evaluating mode of action, and for quantifying points of departure. That does not rule out the possibility that in specific situations, either epidemiological data or possibly even incident data (e.g., aldicarb intoxications from watermelon consumption) could play an important role in either defining or directly contributing to estimates of departure points.

Clarification of Study Designs

In an attempt to simplify the task of assessing the utility of epidemiologic studies for risk assessment, the EPA has grouped these studies into ecologic, retrospective and prospective designs. The SAP discussed design features of different types of epidemiologic studies, methods to qualitatively evaluate them, their potential limitations, and how to make efficient use of the information that is obtained from these studies in the risk assessment and/or risk management process. Although prospective cohort studies can offer many advantages, it is typically not possible to generalize about what study design is best or most appropriate.

A clear description of different study designs and their strengths and limitations for testing hypotheses or evaluating the weight of evidence for a particular cause-effect association is needed. A presentation of study designs from weakest to strongest could include a description of: case reports (i.e., acute poisoning incidents, physician case reports), case series, ecologic studies over time and/or place, clusters, case-control studies, retrospective or historical cohort studies, prospective cohort studies, and mixed designs (e.g., nested case-control, case-cohort, case-crossover). The most appropriate study design will depend on the question being asked and the data requirements (e.g., need for cross-sectional biomonitoring data, historical data on changes in exposures over time, risk estimates for known carcinogens, hypothesis generating studies in populations where cancer incidence has increased, investigation of clusters of potential occupational or environmental origin).

The EPA draft framework should provide additional clarification on the different types of retrospective studies, with a distinction made between case-control studies and retrospective (or historical) cohort designs. Historical cohort studies offer many of the same advantages of prospective cohort studies, with the added advantage of providing much quicker answers to research questions. Further, nested case-control studies offer many advantages and because of the smaller sample needed, they are much more cost-efficient than cohort studies in studying rare outcomes. Recognizing this, it is typically assumed that the prospective cohort is the strongest

observational study design. This is in part because they provide the "opportunity" to collect the most valid and reliable exposure and/or absorbed dose data and because the characterization of exposure occurs prior to the development of disease or other outcome, thus clearly establishing the temporal sequence.

In research on the health effects of exposure to pesticides, regardless of the health outcomes, the methods of ascertaining cases, classifying diseases, selecting controls etc., exposure assessment is and will remain the most challenging aspect. The Agricultural Health Study (AHS) is an example of a prospective cohort study that has set the standard for future investigations, has developed and evaluated innovative methods of exposure assessment, and will be producing data for many years that are extremely relevant for the assessment of health risks associated with pesticide exposures.

Prospective Epidemiologic Studies and Use in Risk Assessment

The main advantage of prospective cohort studies is that individuals are followed forward in time and exposure is determined prior to the development of disease. This presents advantages and opportunities for data collection relevant to the risk assessment process. These include:

a. Single and multiple/mixtures exposures that represent environmentally relevant concentrations (and associated absorbed doses) can be measured prior to the development of the outcome.

b. Changes in exposures and/or dose can be measured over time. Cumulative exposures can be estimated as well as peak exposures and variation within and between individuals over time.

c. Biologic markers of exposure can be evaluated in relation to measured and/or predicted exposures using alternative methods, models, records or questionnaire data.

d. Dose validation studies are possible.

e. Biological markers of susceptibility (gene-environment interactions) can be measured that may modify relationships between pesticide exposures and health risks. This information can help to characterize risk.

f. Early biomarkers of effect, which may be precursors to clinical disease, can be measured; this information will be relevant for evaluation of the proposed mode of action/mechanisms in humans.

g. Quantitative exposure data and biomarkers that are intermediate on the pathway from exposure to disease can be collected.

h. Information on lifestyle, other behavioral factors, and other occupational or environmental exposures that may modify exposure response relationships or act as confounders can be obtained prior to the outcome.

Although prospective study designs have several clear advantages, they may have limited power to look at rare outcomes and can take many years to obtain results. Further, even though biomonitoring is considered the preferred approach (gold standard) to obtain valid and reliable dose estimates in these studies, these types of measures are often collected at only one point in time, and may be available only in a subsample of a cohort. Most often a spot sample (as compared to 24 hour samples) of urine or a single sample of serum is collected. Depending on the characteristics of the pesticide, the concentration may reflect only the most recent exposures.

The AHS Study: Considerations for Risk Assessment

Several important considerations were explored with Dr. Michael Alavanja, principal investigator of the AHS, after his presentation and later during the Panel's deliberations. These topics are relevant to exposure assessment and exploration of potential effects of pesticides on cancer and reproductive outcomes.

Exposure Assessment

Exposure assessment for the AHS is substantially improved over many previous epidemiologic studies of pesticide exposure. As part of the joint efforts of the EPA and AHS investigators, the PHED (Pesticide Handlers Exposure Database) data will be used to estimate exposures for cohort members (occupational exposures of licensed applicators) and compared with exposure assessments (i.e., intensity scores) for the AHS. As the Agricultural Handlers Exposure Task Force accumulates data over the next few years, these should be employed. These newer data are designed to replace the PHED to reflect changes in the equipment and newer protections of present day agricultural practices. However, in terms of estimating lifetime exposures, the more historical residues would have been accumulated with the conditions prevalent during the PHED data accumulation. Therefore, there should be consideration of the appropriate database for long-term exposure estimates. Caution is urged in the use of self-reports for historical exposure assessment when the length of recall is very long since memory may not be accurate.

Additional refinement of AHS exposure assessment may be feasible depending on the availability of banked samples and resources to analyze these samples that may be collected in the future. Biologic and environmental sampling and laboratory analyses can be conducted on population subsets (i.e., occupational groups or bystanders including spouses, farm workers and children) and to ensure cost-effective and representative data (Bakke et al., 2009).

A second issue explored during the discussion involved collection of water samples from farm residences for assessment of pesticide contamination. Although it is not clear to what extent such samples are currently available or could be collected in the future, they could be used to ascertain exposure through the drinking water pathway which does not appear to be taken into account in the current exposure assessment. Family farms typically rely on private wells for domestic consumption. Because many agrichemicals are known to contaminate superficial aquifers in regions of heavy pesticide application, this pathway should be evaluated as a potentially significant contributor to exposure.

Refinement of exposure assessment to include domestic exposure to water and house dust, along with occupational exposures may improve exposure assessment and reduce misclassification. The domestic water and/or dust pathway may be particularly important for the female members of the cohort in considering potential adverse reproductive outcomes. For women who work outside the home or do not participate in farming activities these pathways could be significant sources of exposure to pesticides.

Reproductive Outcomes

Prospective studies of the nature of the AHS may take a long period of time to compile results (cohort studies of cancer may run for 20-30 years or longer). This can substantially delay the usefulness of these data for current risk assessments. However, there are short term opportunities for evaluation of reproductive outcomes.

Although a substantial number of women is included in the cohort (approximately 31,000 enrollees), these women are now an average of 56 years of age and therefore beyond the age of childbearing. However, for a portion of the female cohort members who delivered an infant during the study, data linkage with the birth certificate files in the states of Iowa and North Carolina may provide an opportunity to explore hypotheses relating pesticide exposures to outcomes such as birth weight, low birth weight, intrauterine growth retardation and premature birth. Data linkage may also facilitate comparison between self reported reproductive outcome data and self reported data for women who gave birth during the study.

Many adverse reproductive outcomes are prevalent among live births, affecting 5 to 20 percent of women and their children. Therefore, although the numbers of women who gave birth during the follow-up period will be substantially smaller than the total number of enrollees, sufficient statistical power may exist to evaluate these outcomes for specific pesticide exposures. Further, because the relevant period of exposure for reproductive events is likely to be no more than one year, reducing the likelihood of information bias, further analyses of reproductive endpoints is recommended.

Conclusion

Overall, the use of data from a well-designed and carefully executed prospective cohort study such as the AHS should provide the best opportunity for reaching the ultimate goal of incorporating epidemiologic data into risk assessment. These data may eventually be useful for the risk characterization stage of the process. Further, they also may be useful in comparing dose-response data between humans and laboratory animals for some outcomes. Other epidemiologic designs such as mortality analyses and case-control studies, have already identified farmers and agricultural workers as high risk groups for specific cancers and other disorders. These early studies can be used to inform the problem formulation stage of risk assessment. The AHS has the potential to extend the use of epidemiology to risk characterization for agricultural chemicals. Such advances in risk assessment methodology could enhance the usefulness of the risk assessment paradigm for the eventual protection of public health.

The importance of involving relevant stakeholders and experts from all of the appropriate areas in both the planning and assessment stages was stressed. This would provide greater confidence in the proposed work plans, and in the validity and utility of conclusions. The use of interdisciplinary collaborations in this work would add value to the deliberations and would provide an opportunity to establish a sound way of operating when assessing the reliability and relevance of risk assessments in which both epidemiologic and classical toxicologic information bases are combined. The joint assessment of data sets by experts from different areas would help to ensure objectivity of the interpretations, and the quality of the data used.

Question 4.2: The Agency uses a predictive, scenario-based approach to calculate risks associated with the registered use patterns of pesticides. Estimates of risk based on varying levels of protective equipment, application methods, and use conditions are presented. The results of these assessments are used to specify label conditions that are required to support the new registration or continued registration of pesticides. In contrast, the goal of epidemiologic exposure assessment within the AHS is to develop a relative exposure ranking of individuals who are actual pesticide users within a cohort. It is not feasible to directly measure actual exposure in observational analyses such as the AHS. The AHS exposure information is ascertained from questionnaires completed by individual cohort members. Because the AHS and the Agency have different purposes for evaluating pesticide applicator exposure, there are inherent differences in the occupational handler exposure methodologies between the AHS and Agency. How to reconcile these differences in order to make optimal use of the AHS in developing regulatory policy is under investigation by a collaborative effort between EPA's OPP and investigators involved with the AHS. Case study B details a three step analysis plan for accomplishing this goal. Please comment on the proposed plan for comparing the exposure assessment approaches between the Agency and the AHS. Please include in your comments the scientific value of this comparison along with additional and/or alternative analyses which could be conducted.

Panel Response 4.2

The Panel recognized the merit of finding commonalities between the Agency's exposure assessment methodology and the AHS exposure metrics. The Agency's method results in estimates of workers' exposure as an input into risk assessment that form the basis for setting exposure limits for workers. Finding commonalities to the AHS exposure metrics would provide a way to extend the usefulness of this large and growing data base for the protection of pesticide users.

The Agency presented a single illustration for a step 1 comparison between the ground boom and air blast application. The comparison between the Agency and AHS methods is reasonable as they share a similar goal of characterizing the external exposure, and both can assess the relative exposure between work tasks. Some common grounds for the two approaches were recognized, e.g., use of PHED data, reduction of exposure by PPE. The Panel agreed with the Agency that extensive scenario-by-scenario comparisons are needed to characterize more fully the similarities and differences between the two methods before bridging methodologies for the two methods can be achieved.

The Panel suggested that the document be revised and streamlined. The AHS exposure metrics and scores, and methods of calculating these ordinal values are transparent and available in the published literature (Dosemeci et al., 2002). However, the Agency's method is less clear. The calculation of exposure (dermal, as expressed in mg a.i./day) using the PHED database could be clarified. Further discussion of the unit exposure parameters can also be added.

The Panel recommended that discussions of the variability and uncertainty associated with the foundational databases be included for each method, e.g., PHED, AHS's self-reporting and input parameter values. As far as possible, distributional analyses, not merely the point estimates

could be added to the comparison and exposure models. It is difficult to compare the two methods in that both have a series of steps. Thus, it is anticipated that the elements of the two metrics have to be "picked apart" to identify which of the elements measure the same or similar facets of exposure.

The second step involves using biomonitoring datasets from the AHS to compare the exposure metrics of both methods. This step can further elucidate factors that contribute to the differences between the two exposure estimation methods. In general, evaluation based on biomonitoring data would be less complex for shorter-term than long-term exposures. Attempts to compare exposure predictions from the AHS and the PHED generated models will be most fruitful if a common measure is used to connect the two. For example, a biologic monitoring validation study could be conducted to obtain by questionnaire PPE information and application activities (to feed into the AHS algorithm), pesticide application information (i.e., volume used and active ingredient) and other required variables for the PHED model. Exposures can be calculated using both methods and compared with the "gold standard" urinary metabolite values. The uncertainties associated with the biomonitoring data are likely to be large, and it is important that they are well-defined because this dataset is to be used as the link between the two exposure models. Measurement error, and/or bias can be assessed. When assessing the biological significance of the biological data it would be more helpful to think in terms of the coefficient of determination (R^2) that is interpreted as the proportion of variation explained, rather than the correlation coefficient (r) and its associated level of significance. Sensitivity analysis can be performed at each step. However, this will be more readily achieved for the AHS model than the PHED model. However, it would be worthwhile, because it would allow the identification of major discrepancies and align the focus for further investigations.

In comparing the two sets of metrics, it is not necessary to classify each of the thousands of subjects. A proper random sample (not necessarily a simple random sample; stratified random or other variations might work) could be used to get at the general structure of the relationships, including individual correlations. Even when concern is focused on a small group with some specified outcome, a random sample should be adequate if it is supplemented by 100% sampling of the affected subgroup. It was recognized that classifying the subgroup at a later time might raise difficulties of consistency, but the savings in time and effort could be great. Truly random sampling of some kind will be critical to the use of any sampling approach.

The proposed third step involves a larger comparison of the exposure metrics as applied to atrazine and alachlor users in the AHS database. Additional complexity can be anticipated in this step. Thus, the Panel is supportive for the Agency's feasibility analysis before proceeding to this step.

Overall, the Agency sets out to achieve two goals. One is the evaluation of exposure and risk for specific chemical(s). The other is to strengthen the Agency's current method to estimate the exposure using PHED data. It may be necessary for the Agency to establish the priority for these two goals should the task for achieving both become unattainable within the Agency's operational timeline. The Agency indicated that a plan is underway to update PHED. Accordingly, the latter goal of strengthening the Agency's exposure assessment approach may need to be modified or re-defined in the future.

Question 4.3: The Agency has a long-term goal to understand the extent to which findings from the AHS are generalizable to other populations, such as pesticide applicators in states other than North Carolina and Iowa or those who may be exposed to pesticides through other pathways and under different use conditions. Please provide suggestions for analyses which could be conducted to make best use of the results of AHS in a broader regulatory context.

Panel Response 4-3

Direction from EPA staff was provided to include a discussion on the generalizability for pesticide applicators and handlers from other states and post application workers. Further, a discussion of the generalizability of results for bystanders such as farm family members and children and the general population was requested.

Although the states of Iowa and North Carolina were deliberately selected for the AHS to provide some diversity of pesticide use patterns across the United States, other agricultural environments may not be adequately represented in the exposures found in the AHS. Therefore, great caution needs to be exercised in generalizing the health risk findings of this study to other agricultural populations, or even more so to the population at large.

That said, certain standard epidemiologic analyses can be performed to ascertain potential applicability of findings to other populations or population segments. These include analyses of the same pesticide between Iowa and North Carolina as well as comparison of findings across racial and ethnic groups when the data are available to make such comparisons. Analyses by gender, age group and by type of applicator (commercial, private) also may be informative and are likely to be conducted by the AHS investigators during their analyses.

The range of human biologic responses to a chemical agent in affecting a specified demographic subgroup in State A is likely to be virtually identical to the range, at the same doses or exposures, in State B. Therefore, extrapolation from state to state must focus mainly on exposure. Further, state boundaries are, for this purpose, artificial and almost irrelevant to exposure, though they may be important for such things as data sources, local customs (e.g., use of PPE) and possible state or local regulations (e.g., aircraft spraying), types of workers (e.g., education, supervision) but not types of work.

Results should be generalizable to "pesticide applicators" in other states if populations (racial/ethnic composition) and socio-demographic factors are similar. Factors that may modify the relationship between exposure and outcome are important (e.g., genetic factors that influence susceptibility, carcinogen activation/metabolism and other factors that may act synergistically such as smoking and other behavioral characteristics). It was recommended that a clear understanding of these and other geographic factors (such as local weather conditions and pesticide use patterns) is needed and that generalization of results to other states, based on geopolitical boundaries only, may not be appropriate.

The Panel considered exposure via other pathways and under different use conditions. Because of the broad nature of this question, which is beyond the applicator exposure scenarios, it would be clearer if the Agency presented a set of schemes and metrics for the exposure scenarios for

which the Agency is planning to apply the knowledge learned from the AHS. The AHS Intensity Level metric may have the flexibility to handle multiple exposure pathway and use conditions. It appears that if novel situations arise, it would be feasible to modify or add to the determinants of the Intensity Level calculation, but this is relevant primarily for occupational exposures.

It may be difficult to generalize the occupational exposures and associated health risks to the environmental/residential setting where exposures may be primarily via drinking water, dust, or bystander exposures. It would certainly be necessary to invest in expensive biologic assessment programs to check on the reliability of exposure models modified to use with "bystander" populations. The sample sizes need not be too large providing that the samples are representative of the strata within the general population.

Analyses to Make Use of the AHS Data in a Broader Regulatory Context

Quite early in the document a clear, succinct statement should be made about how analysis for regulatory purposes differs from analysis for scientific purposes. This matter comes up repeatedly (especially in the addendum to Case Study B), but it is never explained in a way that is clear for a non-scientist.

The Panel considered uncertainty associated with extrapolation of animal data to humans in comparison to the uncertainty associated with extrapolation from one human population to the next. National Toxicology Program (NTP) and International Agency for Research on Cancer (IARC) classifications do rely on human as well as animal data to determine categories of carcinogens, whereas EPA will on occasion use only animal data for quantitative risk assessment. EPA scientists indicated that the default, of using animal data only is not preferable to using human data. These comments prompted a discussion on the relative uncertainty of dose-response assessment based on toxicologic vs. epidemiologic studies. It is important to distinguish between qualitative and quantitative aspects in discussing the utility of epidemiologic data for risk assessment. While epidemiologic data are frequently weighted in classification systems, such as those of International Agency for Research on Cancer (IARC) which are based principally on hazard identification, their contribution to risk characterization is determined by the extent to which they inform dose-response relationships. As a result, the extent to which epidemiologic studies contribute to risk characterization varies as a function of this aspect. Uncertainty associated with characterization of dose-response relationships in human populations will depend on the differences in exposures between the populations studied and other characteristics of those populations.

One potential opportunity for analysis was described by Dr. Alavanja. HPEE, or high pesticide exposure events, have been reported by a number of applicators, and approximately 20% had symptoms (chronic neurologic, respiratory (wheeze) and detached retina), but less than 5% were reported to health care providers. It would be of interest to see how many of these events were captured in the incident databases described in question 3.

Overall, to achieve the goal of generalization, it will be necessary to use interdisciplinary teams of experts and end users at both the planning and evaluation stages. Further, the success of the process will depend on the establishment of transparent frameworks for evaluating the quality of

all data (including animal toxicological, human incident, and epidemiologic data). Poor quality information can add to the noise, and reduce the ability to discern real effects and to make accurate predictions.

REFERENCES

Arbuckle TE, Zhiqiu L, Mery LS. An exploratory analysis of the effect of pesticide exposure on the risk of spontaneous abortion in an Ontario farm population. Environ Health Perspect 2001; 109:851–857.

Aylward LL, Morgan MK, Arbuckle TE, Barr DB, Burns CJ, Alexander BH, Hays SM. Biomonitoring data for 2,4-dichlorophenoxyacetic acid in the United States and Canada: Interpretation in a public health risk assessment context using biomonitoring equivalents. Environ Health Perspect 2010; 118: 177-181.

Bailar JC III. Inhalation Hazards: The Interpretation of Epidemiologic Evidence. In: Assessment of Inhalation Hazards, U. Mohr et al., eds. Springer-Verlag, pp 39-48, 1989.

Bailar JC III, Gurian JM. Month of birth and cancer mortality. J Nat Cancer Inst 1964; 33: 237-242.

Bailar JC III, Gurian J. Congenital malformation and season of birth - A brief review; Eugenics Quarterly 1964; 12: 146-153. Reprinted in Social Biology. 1982; 29: 200-207.

Baird DD, Wilcox AJ, Weinberg CR. Use of time to pregnancy to study environmental exposures. Am J Epidemiol 1986; 124: 470-80.

Bakke B, De Roos AJ, Barr DB, Stewart PA, Blair A, Freeman LB, Lynch CF, Allen RH, Alavanja MC, Vermeulen R. Exposure to atrazine and selected non-persistent pesticides among corn farmers during a growing season. J Expo Sci Environ Epidemiol 2009; 19(6):544-54.

Barr DB, Panuwet P, Nguyen JV, Udunka S, Needham LL. Assessing exposure to atrazine and its metabolites using biomonitoring. Environ Health Perspect 2007; 115(10):1474-8.

Bertazzi PA, Bernucci I, Brambilla G, Consonni D, Pesatori AC. The Seveso studies on early and long-term effects of dioxin exposure: a review. Environ Health Perspect 1998; 106 (Suppl 2):625–633.

Bertazzi P, Consonni D, Bachetti S, Rubagotti M, Baccarelli A, Zocchetti C, Pesatori A. Health Effects of Dioxin Exposure: A 20-Year Mortality Study. Amer J Epidemiol 2001; 153 (11): 1031–1044.

Bobak M, Gjonca A. The seasonality of birth is strongly influenced by socio-demographic factors. Human Reproduction 2001; 16:1512-1517.

Canfield MA, Ramadhanti TA, Langlois PH, Waller DK. Residential mobility patterns and exposure misclassification in epidemiologic studies of birth defects. J Expos Sci Environ Epidemiol 2006; 16:538–543.

Darrow LA, Strickland MJ, Klein M, Waller LA, Flanders WD, Correa A, Marcus M, Tolbert PE. Seasonality of birth and implications for temporal studies of preterm birth. Epidemiology 2009; 20:699-706.

Dodds L, King W, Woolcott C, Pole J. Trihalomethanes in public water supplies and adverse birth outcomes. Epidemiology 1999; 10:233-237.

Fingerhut MA, Halperin WE, Marlow DA, Piacitelly LA, Honchar PA, Sweeney MH, Greife AL, Dill PA, Steenland K, Surada AJ. Cancer mortality in workers exposed to 2,3,7,8-tetrachlorodibenzo-p-dioxin. New Engl. J Med 1991; 324:212-218.

Gordis L. Epidemiology, Fourth edition. Saunders, Philadelphia, 2009.

Goldman LR, Smith DF, Neutra RR, Saunders LD, Pond EM, Stratton J, Waller K, Jackson RJ, Kizer KW. Pesticide food poisoning from contaminated watermelons in California, 1985. Arch Environ Health. 1990; 45(4):229-36.

Gordon SM, Brinkman MC, Ashley DL, Blount BC, Lyu C, Masters J, Singer PC. Changes in breath trihalomethane levels resulting from household water-use activities. Environ Health Perspect. 2006; 114(4):514-21.

Greenland S. Meta-analysis. In: Rothman KJ, Greenland S, Lash TL. Modern Epidemiology. Third Edition. Lippincott-Raven, Philadelphia, 2008, pp 652-682.

Greenland S, Rothman KJ. Fundamentals of Epidemiologic Data analysis. In: Rothman KJ, Greenland S, Lash TL. Modern Epidemiology. Third Edition. Lippincott-Raven, Philadelphia, 2008, pp 213-237.

Harris SA. Assessment of Pesticide Exposures for Epidemiologic Research: Measurement Error and Bias. In: Assessing Exposures and Reducing Risks to People from the Use of Pesticides. eds: R.I. Krieger, N. Ragsdale and J.N. Seiber. ASC Symposium Series 951, American Chemical Society, Washington, DC. pp. 173-186. 2007.

Harris SA, Sass-Kortsak AM, Corey PN, Purdham JT. Development of Models to Predict Dose of Pesticides in Professional Turf Applicators. Journal of Exposure Analysis and Environmental Epidemiology 2002; 12(2):130-144.

Harris SA, Sass-Kortsak AM, Corey PN, Purdham JT. Pesticide Exposures in Professional Turf Applicators, Job Titles, and Tasks Performed: Implications for Epidemiologic Study Design and Interpretation of Results. Amer J Indust Med 2005; 48(3):205-216.

Harris SA, Wells KM. Dose Prediction Modeling For Epidemiologic Assessment of Pesticide Exposure Risks in Occupational Cohorts. In: Assessing Exposures and Reducing Risks to People from the Use of Pesticides. eds: R.I. Krieger, N. Ragsdale and J.N. Seiber. ASC Symposium Series 951, American Chemical Society, Washington, DC. pp. 187-200. 2007.

Hays SM, Aylward LL. Using Biomonitoring Equivalents to interpret human biomonitoring data in a public health risk context. J Appl Toxicol. 2009; 29(4):275-88.

Hertzberg VS, Lemasters GK, Hansen K, Zenick HM. Statistical issues in risk assessment of reproductive outcomes with chemical mixtures. Environ Health Perspect 1991; 90:171-5.

Hinckley AF, Bachand AM, Reif JS. Late pregnancy exposures to disinfection by-products and growth-related birth outcomes. Environ Health Perspect 2005; 113(12):1808-13.

Hosmer DW, Lemeshow S. Applied Logistic Regression. John Wiley and Sons, New York, 2000.

Kesselman AB and Bailar JC III. Disease Incidence and Season of Birth: A Note on Methodology, Eugenics Quarterly 1964; 11: 112-115.

Krieger N, Wolff MS, Hiatt RA, Rivera M, Vogelman J, Orentreich N. Breast cancer and serum organochlorines: a prospective study among white black and Asian women. J Natl Cancer Inst 1994; 86(8):589-99.

Le Marchand, L. Epidemiological Approaches to Studying Cancer II: Molecular Epidemiology. In: Cancer Risk Assessment, Shields, P. G. (ed), Taylor & Francis Group, LLC, Boca Raton, FL, 2005, pp 29-59.

Mattix KD, Winchester PD, Scherer LR. Incidence of abdominal wall defects is related to surface water atrazine and nitrate levels. J Pediatr Surg 2007; 42:947-949.

Moses M. Pesticides. In: Paul M, ed. Occupational and environmental reproductive hazards: a guide for clinicians. Baltimore, MD: Williams & Wilkins, 1993, 296-309.

Munger R, Isacson P, Hu S, Burns T, Hanson J, Lynch CF, Cherryholmes K, Van Dorpe P, Hausler WJ. Intrauterine growth retardation in Iowa communities with herbicide-contaminated drinking water supplies. Environ Health Perspect 1997; 105:308–314.

Nieuwenhuijsen MJ, Toledano MB, Elliott P. Uptake of chlorination disinfection by-products; a review and a discussion of its implications for exposure assessment in epidemiological studies. J Expo Sci Environ Epidemiol 2000; 10:586–599.

Ochoa-Acuña H, Frankenberger J, Hahn L, Carbajo C. Drinking water herbicide exposure in Indiana and prevalence of small-for-gestational-age and preterm delivery. Environ Health Perspect 2009; 117:1619-1624.

Ochoa-Acuña, Carbajo C. Risk of limb birth defects and mother's home proximity to cornfields. Sci Tot Environ 2009; 407: 4447-44451.

Rothman KJ, Greenland S, Lash TL. Modern Epidemiology. Third Edition. Lippincott-Raven, Philadelphia, 2008.

Rothman KJ. No adjustments are needed for multiple comparisons. Epidemiology 1990; 1: 43-46.

Rothman KJ. Epidemiology: An Introduction. Oxford University Press, Oxford, New York, 2002.

Savitz DA. Interpreting Epidemiologic Evidence: Strategies for Study Design and Analysis. Oxford, London, New York, 2003.

Savitz DA, Harlow SD. Selection of reproductive health end points for environmental risk assessment. Environ Health Perspect 1991; 90: 159–164.

Savitz DA, Olshan AM. Multiple comparisons and related issues in the interpretation of epidemiologic data. Am J Epidemiol 1995; 142: 904-908.

Shore RE, Iyer V, Altshuler B, Pasternack BS. Use of human data in quantitative risk assessment of carcinogens: impact on epidemiologic practice and the regulatory process. Regulatory Toxicol Pharmacol 1992; 15:180-221

Steenland, K, Bray S, Greenland S, Bofetta P. Empirical Bayes adjustments for multiple results in hypothesis-generating studies. Cancer Epidemiology Biomarkers Prevention 2000; 9:895-903.

Swan SH, Kruse RL, Liu F, Barr DB, Drobnis EZ, Redmon JB. Semen quality in relation to biomarkers of pesticide exposure. Environ Health Perspect 2003; 111: 1478–84.

Swaen GMH, 2006, A framework for using epidemiological data for risk assessment, Human and Experimental Toxicology 2006: 25:147-155.

Szklo M, Nieto FJ. Epidemiology Beyond the Basics, 2nd edition. Jones and Bartlett, Boston, 2007.

Thonneau P, Abell A, Larsen SB, Bonde JP, Joffe M, Clavert A, Ducot B, Multigner L, Danscher G. Effects of pesticide exposure on time to pregnancy: results of a multicenter study in France and Denmark. ASCLEPIOS Study Group. Am J Epidemiol 1999;150(2):157-63.

U.S. EPA 2006. Triazine Cumulative Risk Assessment, March 28, 2006. USEPA, OPP, HED.

Villaneuva CM, Durand G, Coutte MB, Chevrier C, Cordier S. Atrazine in municipal drinking water and risk of low birth weight, preterm delivery, and small-for-gestational-age status. Occup Environ Med 2005; 62:400-405.

Ward E. Epidemiological Approaches to Studying Cancer I: Study Design, Confounding Variables, Misclassification, and Cancer Clusters. In: Cancer Risk Assessment, Shields, P. G. (ed), Taylor & Francis Group, LLC, Boca Raton, FL, 2005, pp 17-38.

Winchester PD, Huskins J, Ying J. Agrichemicals in surface water and birth defects in the United States. Acta Pediatrica 2009; 98:664-669.

Whorton D, Milby TH, Krauss RM, et al. Testicular function in DBCP-exposed pesticide workers. J Occup Med 1979; 21: 161-69.

Wilcox AJ, Horney LF. Accuracy of spontaneous abortion recall. Am J Epidemiol 1984; 120(5): 727-33.

Wilcox AJ, Weinberg CR, O'Connor JF, Baird DD, Schlatterer JP, Canfield RE, Armstrong EG, Nisula BC. Incidence of early loss of pregnancy. N Engl J Med 1988; 319(4):189-94.