Journal of Law and Policy

Volume 13
Issue 2
SCIENCE FOR JUDGES IV:
Agent Orange Revisited and Human Behavior
Research

Article 4

2005

How Scientists View Causality and Assess Evidence: A Study of the Institute of Medicine's Evaluation of Health Effects in Vietnam Veterans and Agent Orange

Irva Hertz-Picciotto

HOW SCIENTISTS VIEW CAUSALITY AND ASSESS EVIDENCE: A STUDY OF THE INSTITUTE OF MEDICINE'S EVALUATION OF HEALTH EFFECTS IN VIETNAM VETERANS AND AGENT ORANGE

Irva Hertz-Picciotto, Ph.D., M.P.H.[1]

INTRODUCTION

The courts are often called upon to settle disputes in which health damages are alleged to have been caused by environmental exposures to chemical, physical, or biologic agents. Similarly, health scientists are often called upon to assess evidence regarding diseases or developmental injuries that might be regarded as resulting from specific exposures. The overarching purpose of this paper is to familiarize readers with the way in which scientists assess data and view evidence about causality, using the example of herbicide and related exposures incurred by U.S. military personnel during service in Vietnam.

One mechanism by which governmental or regulatory agencies at the international, national, or regional levels seek advice from scientists is by convening expert panels. These panels or advisory boards may be assembled as part of an ongoing program that reviews the state of the scientific literature on a topic or in response to specific needs. For example, panels may be assembled to help formulate a regulatory standard for a chemical in drinking water, ambient air, or the workplace environment. Thus, the

[1] Professor of Epidemiology, Department of Public Health Sciences, University of California, Davis; Chair, 2000 and 2002 Institute of Medicine/National Academy of Sciences Committee on the Health Effects in Vietnam Veterans of Exposure to Agent Orange and Other Herbicides.

documents produced by expert committees may become the foundation for the development of health-related policies.

The Institute of Medicine Committee to Review the Health Effects in Vietnam Veterans of Exposure to Herbicides ("IOM Committee" or "Committee") is one such panel. This Committee was formed under the mandate of Public Law 102-4 (better known as the Agent Orange Act)[1] to provide reports on a biannual basis to the Department of Veterans Affairs (VA), beginning in 1994. These reports were concerned with the potential adverse effects that might have been experienced by those who served in Vietnam because of exposures to herbicides, particularly the mixture dubbed Agent Orange, or contaminants found in this mixture, including the well-known chemical compound commonly referred to as "dioxin."

Part I of this article introduces the charge to the Committee, the process the Committee followed in order to reach conclusions about the evidence, the types of studies it considered, and the evidentiary categories it established for classifying specific health outcomes. Part II provides context for the decisions of the IOM Committee through a discussion of the principles that guided the Committee's evaluative process and a presentation of the scientific concepts that constitute the foundation for inferences about causation in biomedical research. Part III explains the approach used by scientists, specifically, epidemiologists, for conducting studies in populations, estimating causal effects, and examining hypotheses. It also focuses more concretely on the obstacles to inferences about causation, specifically, imprecision, which is the uncertainty that arises from studying small samples, and bias, which is the uncertainty that derives from imperfections in study methodology. Part IV narrows this discussion to a description of the major types of bias—confounding, information, selection, and statistical bias.

In contrast to the preceding sections, which focus on individual epidemiologic studies, Part V delineates the process by which scientists reach consensus and presents the framework commonly

---

[1] Agent Orange Act of 1991, Pub. L. No. 102-4, 105 Stat. 11 (codified as amended at 38 U.S.C. § 1116) [hereinafter Agent Orange Act].

used in weighing a body of evidence involving sometimes dozens of studies. Part VI returns to the work of the IOM Committee and provides a detailed discussion of the evidence the Committee reviewed regarding the four outcomes mentioned above, taking into consideration the concepts presented in Parts III through V.

I. VETERANS AND AGENT ORANGE: THE INSTITUTE OF MEDICINE COMMITTEE

A. Charge to the Committee

In light of growing concern about the health of Vietnam veterans, Congress enacted Public Law 102-4, the Agent Orange Act of 1991.[2] Through this Act, Congress directed the Secretary of Veterans Affairs to request from the National Academy of Sciences (NAS) a comprehensive evaluation of the potential health effects from exposure to Agent Orange, a chemical compound used as a defoliant by the U.S. military during the Vietnam War.[3] This legislation also called for reviews of newly available information on a biannual basis for a period of ten years.[4] The Institute of Medicine (IOM) of the NAS convened a Committee to carry out this work. The charge to the Committee was to determine "to the extent that available scientific data permit meaningful determinations" the answers to three questions regarding specific health outcomes and their relationships to Agent Orange exposure.[5] The first was "whether a statistical association with herbicide exposure exists, taking into account the strength of the scientific evidence and the appropriateness of the statistical and

---

[2] *Id.*

[3] *Id.* §§ 2-3.

[4] *Id.* § 3(g)(1) (requiring that a report be submitted to the Secretary of Veteran Affairs "at least once every two years").

[5] *Id.* § 3(d)(1). *See also* COMMITTEE TO REVIEW THE HEALTH EFFECTS IN VIETNAM VETERANS OF EXPOSURE TO HERBICIDES, INSTITUTE OF MEDICINE OF THE NATIONAL ACADEMIES, VETERANS AND AGENT ORANGE: HEALTH EFFECTS OF HERBICIDES USED IN VIETNAM 221 (1994) [hereinafter VAO 1994].

epidemiologic methods used to detect the association."[6] The Committee also was charged with determining "the increased risk of the disease among those exposed to herbicides during service in the Republic of Vietnam and during the Vietnam era."[7] Further, the Committee was asked to assess "whether there exists a plausible biologic mechanism or other evidence of a causal relationship between herbicide exposure and the disease" in question.[8] Finally, Congress charged the Committee with making recommendations

---

[6] Agent Orange Act § 3(d)(1)(A); *see also* VAO 1994, *supra* note 5, at 221.

[7] Agent Orange Act § 3(d)(1)(B); *see also* VAO 1994, *supra* note 5, at 221 (stating "the increased risk of each disease in question among those exposed to herbicides during Vietnam service").

[8] Agent Orange Act § 3(d)(1)(C); *see also* VAO 1994, *supra* note 5, at 221. Some authors have argued that the first charge does not mandate the Committee's examination of "cause" or "causal association," but instead requires only that the Committee look into a possible "statistical association." However, the third question indeed requests that the Committee evaluate the existence of "evidence of a causal relationship." Notably, any determination about the existence of "statistical association" that takes into account "strength" of the evidence and "appropriateness" of the methods examines the same concerns that enter into a consideration of evidence for causation. These concerns (for example, the strength of the association and the methods used) give rise to issues such as bias and confounding, which are defined in detail in Parts II through IV. Thus, although the Committee was not charged with drawing a conclusion about causation, the combination of responses to questions one and three effectively results in a lengthy consideration of virtually all of the issues that would be discussed if such a conclusion were required. As stated in a recent update issued by the 2002 Committee:

> The evaluation of evidence to reach conclusions about statistical associations goes beyond quantitative procedures at several stages: assessing the relevance and validity of individual reports; deciding on the possible influence of error, bias, confounding, or chance on the reported results; integrating the overall evidence within and between diverse fields of research; and formulating the conclusions themselves. Those aspects of the committee's review required thoughtful consideration of alternative approaches at several points and could not be accomplished by adherence to a narrowly prescribed formula.

COMM. TO REVIEW THE HEALTH EFFECTS IN VIETNAM VETERANS OF EXPOSURE TO HERBICIDES, INSTITUTE OF MEDICINE OF THE NATIONAL ACADEMIES, VETERANS AND AGENT ORANGE: UPDATE 2002 28 (2003) [hereinafter VAO UPDATE 2002].

for areas in which further study might help answer the questions of concern.[9]

Faced with the above mandates, the Committee first clarified the exposures to be evaluated. The Committee focused specifically on exposure to Agent Orange. Agent Orange and the other defoliants used in Vietnam were comprised of combinations of one or more of four herbicides: 2,4-D (dichlorophenoxyacetic acid), 2,4,5-T (trichlorophenoxyacetic acid), 4-amino-3,5,6-trichloropiclorinic acid (picloram), and dimethylarsenic acid (DMA or cacodylic acid). Mixtures containing 2,4-D or 2,4,5-T were contaminated by chemicals formed during the production process, including 2,3,7,8-trichlorodibenzodioxin (2,3,7,8-TCDD). Although other dioxins and dibenzofurans were also formed, 2,3,7,8-TCDD is considered the most toxic and, as such, was the compound reviewed most extensively by the Committee.

In the course of examining the effects of Agent Orange exposure, the Committee never evaluated the claim of any individual veteran, as this was not its charge. Indeed, the Agent Orange Act specified that such decisions would be made by the VA. Moreover, the Committee was instructed not to consider the issue of potential compensation in its deliberations.

Before beginning its work, the members of the Committee were required to disclose potential conflicts of interest or biases, or anything that might create the appearance of a conflict of interest. These included financial holdings, consulting activities, government service, areas of research, and professional affiliations as well as any public statements or intellectual positions relevant to the topic of the Committee. Committee members served without

---

[9] Agent Orange Act § 3(e) (directing the National Academy of Sciences to "make any recommendations it has for additional scientific studies to resolve areas of continuing scientific uncertainty relating to herbicide exposure"); s*ee also* VAO 1994, *supra* note 5, at 15. One of these recommendations was to commission an historical exposure reconstruction. *Id.* at 17-18. This recommendation led to the studies of Agent Orange exposure in Vietnam described by Dr. Jean Mager Stellman in this issue. *See* Jeanne Mager Stellman & Steven D. Stellman, *Characterization of Exposure to Agent Orange in Vietnam Veterans As a Basis for Epidemiological Studies,* 13 J.L. & POL'Y 505 (2005).

compensation, except for reimbursement of expenses.

Because the scope of the review is broad, the Committee includes health scientists representing expertise in a wide range of fields covering epidemiology, oncology, neurology, reproductive health, and toxicology. The IOM staff assists with the review, conducting library searches that begin with hundreds, if not a few thousand, of articles, and works with the Committee to progressively narrow them down to those articles that are pertinent to the questions at hand.

B. Types of Evidence Reviewed

For the scientists on the Committee, it was obvious that any findings of health effects from these same exposures could serve as evidence for potential effects in the Vietnam veterans, even if the results were obtained in other populations. Scientists consider biological systems in human beings to be sufficiently similar throughout the world that a high proportion of research findings, especially those that have been replicated in several studies, can be "generalized" to much larger populations beyond those that were studied. Because few studies actually were conducted on Vietnam veterans, other data sources were frequently used as the basis for the Committee's decisions regarding the first and third questions posed to it by the Act.

The three main sources of epidemiologic data used to address the first question were studies conducted in: (a) occupational groups with exposures in the workplace, such as chemical manufacturing, farming, application of herbicides, or paper and pulp manufacturing (where 2,3,7,8-TCDD is produced as a by-product of the bleaching process);[10] (b) populations with environmental exposures, which typically result from accidents that contaminate residential or recreational areas,[11] or alternatively, from residence in agricultural regions in which herbicides are

[10] 2,3,7,8-TCDD is produced as a by-product of the bleaching process

[11] Such an event occurred in Seveso, Italy, when an explosion at a chemical plant caused 2,3,7,8-TCDD to contaminate a wide residential area. *See* Pier Alberto Bertazzi et al., *Health Effects of Dioxin Exposure: A 20-year Mortality Study*, 153 AM. J. EPIDEMIOLOGY 1031 (2001).

widely used; and (c) veterans who served in Vietnam, including not only the U.S. armed forces, but also those from Australia.

Although hundreds of studies have been reviewed each time the Committee has been convened (every two years), there are some cohorts of exposed persons that have played a prominent role in the deliberations. These cohorts had high exposures and were evaluated numerous times, often for different health outcomes (such as cancer, heart disease, diabetes, neurologic disorders, immune function and allergies, reproductive events, etc.), each time contributing more information to our knowledge base. Some of the most important of these were the National Institute of Occupational Safety and Health (NIOSH) cohort of workers employed after 1942 at twelve plants that manufactured chemicals containing 2,3,7,8-TCDD;[12] a similar multinational cohort from more than half a dozen European countries, assembled by the International Agency for Research on Cancer (IARC);[13] the cohort exposed to the explosion of a chemical plant in Seveso, Italy, in 1976 that released 2,3,7,8-TCDD over an area populated by more than 200,000 persons;[14] and the Air Force Health Study (AFHS), also known as the "Ranch Hand" study, of U.S. Air Force personnel responsible for flying spraying missions to defoliate North Vietnam using Agent Orange (these missions were termed "Operation Ranch Hand").[15]

Each of these cohorts was characterized by higher than usual exposures. For instance, in the two occupational cohorts, a subset of the workers had experienced chloracne, an acute reaction of skin

---

[12] *See, e.g.*, Marilyn A. Fingerhut et al., *Cancer Mortality in Workers Exposed to 2,3,7,8-tetrachlorodibenzo-p-dioxin*, 324 NEW ENG. J. MED. 212, 212 (1991); Kyle Steenland et al., *Cancer, Heart Disease, and Diabetes in Workers Exposed to 2,3,7,8-tetrachlorodibenzo-p-dioxin.* 91 J. NAT'L CANCER INST. 779 (1999).

[13] Manolis Kogevinas et al., *Cancer Mortality in Workers Exposed to Phenoxy Herbicides, Chlorophenols, and Dioxins: An Expanded and Updated International Cohort Study*, 145 AM. J. EPIDEMIOLOGY 1061, 1061(1997).

[14] *See* Bertazzi, *supra* note 11.

[15] SCI. APPLICATIONS INT'L CORP., AIR FORCE HEALTH STUDY, AN EPIDEMIOLOGIC INVESTIGATION OF HEALTH EFFECTS IN AIR FORCE PERSONNEL FOLLOWING EXPOSURE TO HERBICIDES (1997), FOLLOW-UP EXAMINATION RESULTS (2000).

eruptions that follows high exposures to 2,3,7,8-TCDD. In addition to the examination of multiple health endpoints, numerous subsets of the cohorts also were examined more extensively. Furthermore, in each of these four cohorts exposure was, at some point, measured in blood drawn from a subset of participants.

The Ranch Hand study was the most extensive study of veterans. The Air Force initiated this cohort study, which attempted to recruit about 1,200 servicemen who were identified as Ranch Hand personnel and 1,700 Air Force personnel who were assigned to duty in Southeast Asia, but were not exposed occupationally to herbicides. In 1982, a baseline examination was conducted of both groups of men, and follow-up took place every five years thereafter.

The Committee also held public hearings and invited written submissions. These provided the Committee members with an opportunity to hear from those most familiar with the conditions and sequelae of service in Vietnam, as well as the authors of relevant papers, including some that were in press, but not yet published.

C. The Process

To provide a framework for its decisions, the initial committee, which began meeting in 1992, defined four categories of evidence.[16] The first category consists of those health outcomes for which the available data provide *sufficient evidence* of an association.[17] This category applies when multiple studies are consistent in showing an association, and bias, confounding, or random variation are not likely to explain the findings. The second category consists of those health outcomes for which the available research provides *limited or suggestive evidence* of an association.[18] This category may apply when multiple studies observe an association, but the magnitude is sufficiently small that bias, confounding, or random variation cannot be ruled out.

---

[16] *See* VAO 1994, *supra* note 5, at 246.

[17] *Id.*

[18] *Id.* at 247.

Alternatively, there may be one or more reasonably high quality studies showing an association that other studies do not confirm.

The third category is used to identify those health outcomes for which the literature provides *inadequate or insufficient evidence* from which to determine whether an association exists.[19] This category is used when there are very few studies, none of which is definitive, or where there are many studies, but the quality is inadequate (the studies might have failed to control confounding) or the findings are inconsistent. Finally, the last category is used to designate those health outcomes for which the extant research provides limited or suggestive *evidence of no association*.[20] This category is used when there are numerous studies of reasonably high quality, and they consistently show no association between the exposure and the outcome.

The above categories were applied to the first question with which Congress charged the Committee. With regard to the second question, the paucity of data on those who served in Vietnam precluded, for the most part, making a determination about the magnitude of increased risk. First, the inability to assign exposures to individual veterans, including the vast majority of those who participated in the research studies that were conducted, was seen as an enormous obstacle. When an agent induces a response, it is recognized that the magnitude of the response, or the likelihood of developing a disease, tends to increase as the exposure gets larger. This phenomenon is referred to as "dose-response." The Committee concluded that, without knowledge of the exposure level, the size of the risk could not be quantified. Even if an average exposure level were known, it would still be difficult to estimate an average risk because the existing research, whether in veterans, exposed workers, or accidentally-exposed populations, usually could not establish, that is, did not quantify, the dose-response relationship. Given the lack of information about how steeply the risk for each of the health outcomes evaluated would be expected to rise, the Committee concluded that it was unable to answer the second question regarding the "increased risk of each

---

[19] *Id.*

[20] *Id.*

disease among those exposed" with any specificity.[21]

The responses to the third question—whether a biologically plausible mechanism or other evidence supporting a causal association existed—expanded the work of the Committee beyond epidemiology and engaged the Committee in the review of a broad spectrum of studies. The premise of this question is that any determination about a causal relationship is strengthened when the mechanism of action is understood. For example, researchers might find that blood pressure in persons who have taken a certain drug is generally lower than in those who have not. If in addition researchers determine that this drug relaxes the smooth muscles in the arterial walls known to be inversely related to blood pressure (muscle contraction increases pressure, but relaxation reduces it), then the plausibility of the drug having a causal effect (in this case a protective one) is enhanced. The data used in evaluating biologic plausibility may derive from a wide range of scientific fields, including toxicokinetics, which examines how and at what rate compounds are absorbed into the body, distributed to different organs, chemically metabolized, and excreted; whole animal toxicology, which addresses the pathologic and homeostatic responses of the organism, often in rodents but also in other species, including humans; and molecular and cellular biology, which seeks to understand the biochemical alterations that result from an exposure and the subsequent consequences for cell functioning. Thus, biologic evidence from experimental studies in humans, other animals, and test systems such as cell cultures is used to determine whether a plausible mechanism exists. Such evidence is considered to provide support for inferring causation when statistical associations have been observed in human studies.

In evaluating the evidence pertaining to its congressional mandate, the Committee made decisions regarding the relationship between Agent Orange exposure and dozens of health outcomes. Four of these will be reviewed in detail: (a) non-Hodgkin's lymphoma, (b) Type II diabetes, (c) prostate cancer, and (d) the

[21] The term "risk" will be further defined later in this article but, for now, it shall suffice to say that, in the judgment of the Committee, the number of cases of each disease among Vietnam veterans due to herbicide exposure could not be estimated with any reasonable accuracy.

presumptive period for respiratory cancer. The first three are health outcomes, while the final one is an issue of timing and causation. The Committee concluded that the evidence for an association with non-Hodgkin's lymphoma was sufficient and that biologic plausibility was established. For Type II diabetes and for prostate cancer, the Committee concluded that there was limited but suggestive evidence for an association. The VA had ruled that respiratory cancer could be considered service-related only if it manifested within a period of thirty years after the end of service in Vietnam, termed the "presumptive period." Charged with determining whether this presumptive period had a scientific basis, the Committee concluded that it did not. To provide the reader with a background for understanding how the Committee reached each of these conclusions, this article now turns to an exposition of key concepts in scientific and epidemiologic research.

II. CAUSAL EFFECTS

The concept of causation is fundamental to scientific inquiry, which seeks to understand cause and effect relationships of physical, chemical, or biological phenomena. Within biomedical sciences, potential causes of disease and developmental disorders are studied using a variety of tools, including epidemiology and toxicology. In these fields, however, the concept of "cause" differs from that which courts use in settling individual or even class action cases.

To study the causal effects of an exposure, two identical "units" must be compared, one exposed and one unexposed. A unit might be, among other things, a person, a laboratory animal, a cell, or a piece of tissue. In order for the study to produce results about causal effects, it is essential that the two studied units be absolutely identical, which is to say that they differ *only* with regard to the exposure. Each unit is evaluated for some response, such as growth, chemical or electrical activity, or structural or functional change. The difference in response between the exposed unit and the unexposed unit represents the "causal effect."

Studies can be classified in many ways, but one significant distinction is between experimental and observational studies. Table 1 compares these two types of studies:

Table 1

Experiment	Observational Study
Identical units, such as a single strain of laboratory mice	Units not identical
Scientist manipulates exposure and determines which units are or are not exposed	Exposure occurs beyond the control of the scientist
Scientist determines (measures) exposure levels	Scientist measures exposure (measurements may be subject to greater error than in experiment)
Outcome is measured: reliability will vary with the nature of the outcome and the quality of the protocol for its measurement	Outcome is measured: reliability will vary with the nature of the outcome and the quality of the protocol for its measurement

Notably, it is easier to ensure the use of identical "units" in an experimental study than in an observational study. However, the more fundamental difference is that, in an experimental study, the exposure is controlled by the investigator; that is, the investigator decides which unit will receive the exposure and which will not. Typically, this decision is made randomly and either of the units could be the chance recipient of the exposure. The investigator also determines the level of exposure each unit receives and may assign the units to different amounts or intensities of exposure. In an observational study, by contrast, exposure is not assigned, but rather, occurs for reasons that have nothing to do with any

researcher's actions. Historical, social, political, and physical forces, as well as individual choices that may also be shaped by any of the above factors, will determine when and where exposure occurs. As a result, the exposed unit is rarely identical to the unexposed unit. For this reason, the concept of causation in observational studies has been more elusive than in experimental studies.[22]

Recently however, a conceptual paradigm has been developed that aids in the understanding of the conditions under which causal inferences can be made from observational investigations.[23] The underlying concept is the "counterfactual" that contrasts two scenarios. Under the first scenario, the individual unit (usually a person, but also possibly a non-laboratory animal) is exposed and its response is measured.[24] Under scenario two, we suppose that the individual is not exposed and, therefore, we can measure the response that would have occurred had the individual, counter to fact, not been exposed. We call this the counterfactual response. Thus, the individual causal effect in an observational setting is the difference between the actual and the counterfactual response.

Unfortunately, the individual causal effect can never be known since researchers can never observe both the factual and counterfactual experience. Epidemiologists, however, strive to measure the group-level causal effect, which represents a type of average of the individual-level effects, under the assumption that the two groups (exposed and unexposed) each represent the counterfactual experience, on average, of the other. In order to do so, epidemiologists must first define the following terms: risk (R),

---

[22] It should be noted that even in double-blind, placebo-controlled trials, inferring general causation can be problematic, due largely to the fact that persons who participate in these studies and who are compliant with the treatment regimes are often a select group. Additionally, despite randomization, the exposed and unexposed may differ in unmeasured ways, by chance, particularly in small trials. Randomization reduces the likelihood of confounding but does not eliminate it. *See infra* Part III.

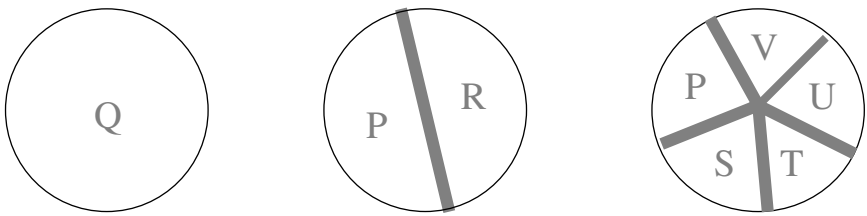
[23] *See* Donald B. Rubin & Roderick J. Little, *Causal Effects in Clinical and Epidemiological Studies Via Potential Outcomes: Concepts and Analytical Approaches*, 21 ANN. REV. PUBLIC HEALTH 121 (2000).

[24] A response might be a continuous measurement such as blood pressure, or a binary outcome, such as the presence or absence of disease.

risk factor (RF), risk difference (RD), and risk ratio (RR). Risk is defined as the probability of disease, while risk factor refers to an exposure or characteristic that increases risk or serves as a surrogate for a factor that increases risk. Risk difference is calculated by subtracting the "risk if exposed" value from the "risk if unexposed" value. Finally, the risk ratio, also known as relative risk, is defined as "risk if exposed" divided by "risk if unexposed."

The importance of the counterfactual assumption cannot be overemphasized. In any study, it is possible to make measurements. In many studies, the risk difference or risk ratio can be measured. However, defining these terms or measuring them does not in itself make them meaningful in terms of causation. Epidemiologists and other scientists often say that, "association does not necessarily imply causation." One can gain additional insight into the source of various conditions and diseases through the use of the sufficient causes model. Figure 1 provides an example of the application of this paradigm:

Figure 1



A: Single cause is sufficient (rare situation)

B: Two causes are required and sufficient

C: Multiple causes are required (common situation)

Each circle, or pie, represents a set of sufficient causes.[25] In circle A, a single cause will result in the disease. This cause might be, for example, the measles virus. The virus alone causes the clinical entity we call measles. In circle B, a second factor is needed; this example might apply if not all individuals exposed to the virus actually exhibited the clinical symptoms of the disease, that is, if some individuals lacked susceptibility to the virus. Thus, in circle B, P is the virus, but R is also necessary because neither P alone nor R alone results in the disease. In circle C, five factors are necessary to cause disease. This type of scenario corresponds to most chronic, or non-infectious, diseases, for which multiple factors are likely to play a role in any individual case. By way of example, it has been suggested that there may be more than ten genes for autism.[26] However, each child with autism probably does not require all such genes to develop this disorder, and there may be several environmental factors also involved. Note also that for any given disease there may be several different sets of sufficient causes; some individuals will require one set and others will require a different, though possibly overlapping, set. For example, among workers who smoke and are exposed to arsenic at their workplace, some might develop respiratory cancer from the cigarette smoke alone, while others might develop cancer from the arsenic alone, and still others might develop respiratory cancer only because they received both exposures.[27]

The sufficient causes model is also instructive in terms of inferring individual causes. Knowledge about the presence or absence of other known risk factors changes the probability that a suspect risk factor was causal for an individual case. If an individual who has never smoked, whose parents never smoked, and who is not married to a smoker develops lung cancer, the probability that this cancer was caused by some other known lung

---

[25] MODERN EPIDEMIOLOGY 8-12 (Kenneth J. Rothman & Sander Greenland eds., 1998).

[26] Sarah J. Spence, *The Genetics of Autism*, 11 SEMINAR PEDIATRIC NEUROLOGY 196, 198 (2004).

[27] Irva Hertz-Picciotto et al., *Synergism Between Occupational Arsenic Exposure and Smoking in Lung Cancer Induction*, 3 EPIDEMIOLOGY 23, 28 (1992).

carcinogen is increased. If, for instance, that individual is known to have high exposure to radon, the likelihood that the radon caused the cancer is higher than it would be for another individual with the same high exposure to radon who smoked or was exposed passively to tobacco smoke.

III. STUDIES IN GROUPS: ESTIMATION AND PRECISION

Researchers prefer to enroll groups for their studies rather than rely on individuals, primarily because individuals almost never provide definitive evidence about causal effects. In recognition of this limitation, measurements are made on an enrolled group with the idea that the results can be extrapolated to the population from which the group arose, and hence, to other individuals who were not participants in the study. The group that is studied is termed a "sample," and any measure on the sample is considered an "estimate" of the parameter (risk ratio, for instance) for the complete population.

For example, researchers concerned that adolescents with symptoms of depression may engage in binge drinking of alcohol might sample a group of high school students. In the sample, researchers may determine what proportion of tenth and eleventh graders attending one high school selected at random from all high schools in a metropolitan school district exhibit depressive symptoms (perhaps by use of a questionnaire). This result provides an *estimate* of the true proportion of high school students with depressive symptoms in that school district and perhaps in that metropolitan area, that state, the country, or all similar countries.

If researchers also found out how many of those high school students engaged in binge drinking, they could estimate the risk ratio for binge drinking by comparing those with depressive symptoms to those without such symptoms. The resulting risk ratio would be an estimate of the risk ratio in the population. If the risk ratio were 1.5, it would mean that high school students in the study who had depressive symptoms were one and one-half times more likely to engage in binge drinking than those who did not. If the risk ratio were 1.0, it would mean that each group of high school students had the same risk of engaging in binge drinking.

However, epidemiologists recognize that the sample studied might be different from or unrepresentative of the complete population and thus they also construct a range around this estimate. This range is known as a confidence interval and represents a range of values that, on average and under certain conditions, is expected to include the true population value. The width of this interval is roughly a function of the study size, a good indication of a more technical quantity known as statistical power. In a study in which less than half of the population is exposed and disease is not common, this power is mostly determined by the number of exposed persons with disease. As the number of exposed persons who develop disease increases, the confidence in the estimate increases and the interval becomes tighter around the estimate. For example, a small study with an RR of 1.5 might have a 95% confidence interval of 0.5 to 4.5, in which case we would say that the precision is low. A much larger study that also had an RR of 1.5 might have a 95% confidence interval of 1.3 to 1.7, indicating very high precision.

The above exposition emphasizes the estimation of effects. The precision of these estimates (reflected in confidence intervals) is related to another concept used by scientists and invoked in recent court decisions regarding admissibility of scientific evidence, namely "statistical hypothesis testing." A common practice in many scientific fields is to construct a "null hypothesis," which states that there is no association between the exposure and the outcome. Once the study has been conducted, the result is compared with the null hypothesis. If the study result is extremely different from what is predicted by the "null hypothesis," then, assuming the data are reliable, one may conclude that the null hypothesis is not supported because if it were true, then large deviations from the null would be improbable. To quantify the improbability of the result, one calculates its probability of occurring under the dual assumptions of no association and complete absence of any other information. The resulting probability is called a "p-value." It is sometimes referred to as an "error" rate.

The merits and misuses of p-values have been the subject of

considerable debate within the scientific community.[28] One criticism of the p-value relates to the convention of using a cutpoint of 0.05 to determine whether a finding is "significant" (the designation when the p-value is less than 0.05, that is, when the probability of the result is less than one in twenty if the null hypothesis were true), and declaring all results with p-values above 0.05 as "nonsignificant."[29] The result of a statistical hypothesis test is a decision of whether to "reject" the null hypothesis. In practice, it is difficult to argue that results with a p-value of 0.051 are qualitatively different from those with a p-value of 0.049. Another problem is that the p-value combines two different aspects of the study result: the magnitude of the association and its precision. For instance, one study may have a p-value of 0.04 when the RR is 8.0; this will be a less precise estimate (the confidence interval will be wider) than another study with a p-value of 0.04 and an RR of 2.5. To address this concern, many epidemiologists have preferred to express their results with "estimates" and "confidence limits," thereby keeping these two aspects of the study findings clear and separate.

It has also been noted that although a p-value provides information about the consistency between the "null hypothesis" and the data collected, it provides no information at all about any other hypothesis. If one wanted to hypothesize that a risk is doubled for individuals who are exposed, one would not calculate a p-value. Similarly, if previous studies have already suggested that the null hypothesis may not be true, then it may be of greater interest to evaluate whether the new data are consistent with the previously published findings rather than whether they are consistent with a null effect. In fact, p-values do not provide the means for placing findings in context,[30] or for considering the possibility of biases.[31] Instead, they are calculated by either

---

[28] One website lists "326 Articles/Books Questioning the Indiscriminate Use of Statistical Hypothesis Tests in Observational Studies." *See* http://www.cnr.colostate.edu/~anderson/thompson1.html.

[29] Note that if the 95% confidence interval includes the null value (0 for a risk difference, 1 for a risk ratio), then the p-value will be greater than 0.05.

[30] *See infra* Part V.

[31] *See infra* Part IV.

assuming no other information or by explicitly ignoring it.

Overall, it is unwise to make decisions on the basis of a single set of data, a practice that is encouraged by the use of p-values. Science does not actually proceed in the manner implied by statistical hypothesis testing and, rather than relying on decisions at the end of each study, scientists gather and review the body of evidence as a whole. It has been suggested that the practice of hypothesis testing detracts from scientific thinking; indeed, one journal in the field of epidemiology strongly discourages the use of p-values to summarize results[32] and frequently asks authors to remove them as a condition of accepting a paper.

Further critiques point out that the common use of $p<0.05$ as a criterion for deciding to reject the null hypothesis is based on the implicit assumption that there is a high cost to mistakenly rejecting the null hypothesis and thereby "finding an association." In other words, this convention presumes that such a conclusion should be made only very cautiously (society cannot afford to make this mistake more than 5% of the time). In some circumstances, this implicit assumption may prove problematic. If the harm from an exposure is severe, a regulatory body, for example, may wish to err on the side of protecting public health. This, however, would require the use of different criteria. In the courtroom, a "more likely than not" standard is used in some circumstances as the bar against which to evaluate evidence. In a single study (absent any other research), a $p<0.50$ means the probability is less than 50% that the data (or more extreme data) arose from a population in which exposure and disease are not associated. Hence, a $p<0.50$ would be much closer to the criterion of "more likely than not" for evaluating whether the data arose from a population in which exposure does not cause disease.[33]

While epidemiologists strive to conduct studies that produce precise estimates, there is always the possibility that the estimate could be wrong, not because of chance "sampling error" that occurs with small or even moderate-sized samples, but because of

---

[32] *See* Janet M. Lang et al., *That Confounded P-Value*, 9 EPIDEMIOLOGY 7, 8 (1998); The Editors, *The Value of P*, 12 EPIDEMIOLOGY 286 (2001).

[33] I will not presume to guess what probability would correspond to the criterion of "beyond a reasonable doubt."

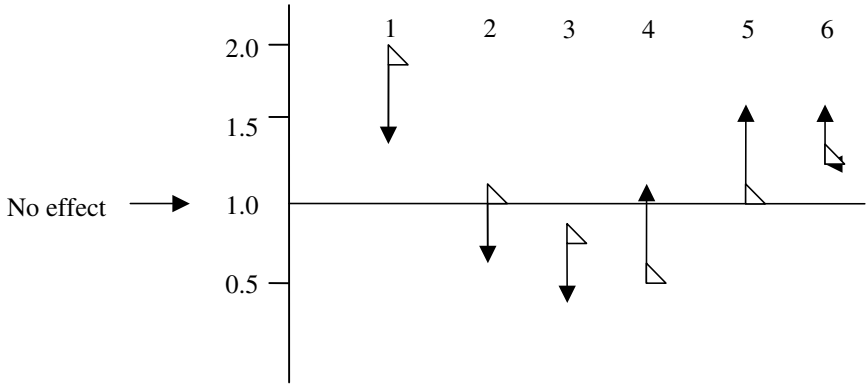
a more systematic problem known as bias. When bias is present, not only is the estimated association incorrect, but the p-value does not represent the purported "error rate." As we shall see, bias is usually a greater concern than errors due to random fluctuations that produce these error rates in observational studies.

IV. STUDIES IN GROUPS: BIAS

Bias is present when, on average, the estimates tend to be either too high or too low relative to the true population parameter. Figure 2 displays how bias can distort a relative risk by creating either artificial effects or masking true effects:

Figure 2: Upward and Downward Bias

The base of the wedge at the start of each arrow is the true value of the risk ratio, and the arrowhead is at the biased (observed) value. In case 1, an increased risk due to a harmful exposure will be presumed to be smaller than it truly is. In case 2, an exposure with no effect will be presumed to reduce risk. In case 3, an exposure that is beneficial (reduces risk) will appear to be more beneficial than it really is, whereas in case 4, a beneficial exposure will appear as though it has a small harmful effect. In case 5, an exposure that has no effect will be presumed to be harmful and, in case 6, a slightly harmful exposure will be presumed to be more harmful than it is.



Cases 1-3 show "downward" bias, and cases 4-6 show "upward" bias.

The main types of epidemiologic bias are selection bias, information bias, confounding bias, and statistical bias. Selection bias occurs when the subjects in the study sample do not represent the targeted population with regard to the exposure and the disease. Consider Figure 3, the epidemiologic two-by-two table in which each individual falls into one of four cells: exposed with disease, exposed without disease, unexposed with disease, or unexposed without disease. A completely representative sample will take approximately the same proportion from the population out of each of the four cells. (This means that if 90% of the population is in the cell for unexposed without disease, then 90% of the sample also would be from that cell.) It is also possible to intentionally sample at a different rate from one column or one row and still obtain an unbiased estimate, but only if the investigator ensures that both cells in that row or column are sampled in the same proportion.

Figure 3: The Two-by-Two Table

	Exposed:	Unexposed:	
Diseased:	Exposed with disease	Unexposed with disease	Total diseased
Not diseased:	Exposed without disease	Unexposed without disease	Total without disease
	Total exposed	Total unexposed	Grand Total

More concretely, suppose that a study is conducted to examine the hypothesis that the use of hot tubs by pregnant women increases the risk of spontaneous abortion. Suppose further that women who use hot tubs are more likely to participate in the study because they have more leisure time and that women who have spontaneous abortions are also more likely to participate because they are concerned about why they lost their pregnancies. In essence, a larger percentage of the population in the upper left cell of the epidemiologic two-by-two table participated in the study than the population percentage in the other cells. In other words, proportionately fewer women who did not use a hot tub or who did not spontaneously abort would participate in the study.[34] This would lead to an upward bias in the estimated RR. Thus, if, for the sake of argument, the true risk ratio for spontaneous abortion from hot tub use were 1.2, in this study we might see an estimated risk ratio of 1.5.[35] If the true risk ratio were 1.0, we might, for example,

[34] This point was made in a commentary by Irva Hertz-Picciotto & Penelope P. Howards, *Invited Commentary: Hot Tubs and Miscarriage: Methodological & Substantive Reasons Why the Case Is Weak*, 158 AM. J. EPIDEMIOLOGY 938 (2003) (critiquing Li De-Kun et al., *Hot Tub Use during Pregnancy and the Risk of Miscarriage*, 158 AM. J. EPIDEMIOLOGY 931 (2003)).

[35] *See* case 6 in Figure 2.

observe a risk ratio of 1.2 or greater.[36]

In other examples, the bias might occur in the opposite direction. For instance, researchers who studied high fat diets and diabetes might find that persons eating high fat diets and diabetics would be less likely to participate. In this case, the upper left cell in Figure 3 would be underrepresented as compared to the population at large. Therefore, if the true RR were 1.8, then one might observe an RR of, say, 1.2;[37] alternately, if the true RR were 1.0, one might observe a lower RR of, for instance, 0.7.[38] In the former case, researchers might incorrectly conclude that there is only a small detrimental effect of the high fat diet when it is quite harmful, and, in the latter case, the study incorrectly suggests a protective effect, that is, a lower risk of diabetes among those who eat high fat diets. In short, selection bias can lead one to draw the wrong conclusion.

Information bias, by comparison, occurs when information about the disease diagnosis differs between those who are exposed and those who are unexposed. For example, bias might result where individuals of low socioeconomic status who do participate are more likely to be exposed than those at higher socioeconomic levels, but less likely to be diagnosed because they lack health insurance and rarely see a physician. Thus, persons with the disease may be misclassified as healthy because they are not yet diagnosed. Information bias also might occur when data on exposure differs with regard to those who have the disease and those who do not. For example, in a study of the possible connection between pesticide use around the home and incidence of childhood leukemia, parents may be asked to recall what pesticide products they used and when. The parents of affected children might be more likely to recall every insecticide or fungicide used in or around their house than the parents of healthy children. This would result in a specific type of information bias termed reporting bias or recall bias, which usually results in upward bias.

---

[36] *See* case 5 in Figure 2.

[37] *See* case 1 in Figure 2.

[38] *See* case 2 in Figure 2.

Another type of bias, confounding bias, occurs when an alternative risk factor for the disease (one that is not the exposure of interest for the study) happens to occur more or less frequently in the exposed as compared with the unexposed. Consider, for example, a study to examine the hypothesis that exposure to polychlorinated biphenyls (PCBs) during infancy adversely affects the cognitive development of children. Suppose that a major source of PCBs to infants is breast milk. Then suppose that mothers who breastfeed their infants are more educated and are more likely to read to their children or offer other intellectual stimulation. Notice that an experiment to test the hypothesis that PCBs adversely affect cognitive development would randomly assign some mothers to breastfeed and others to give formula. However, in the real world, women who choose to breastfeed are not the same as those who do not elect to breastfeed and hence cannot serve as the "counterfactual" experience for those who do not breastfeed. The result is confounding bias: children with a higher exposure to PCBs were given greater intellectual stimulation. In this example, the RR would be biased downward, but it is possible that in other examples the RR could be biased upward.

A fourth type of bias is statistical bias. Statistical bias occurs as a result of errors in statistical analysis or limitations in data. Sometimes the methods used for analysis do not match the conditions in which the data were collected or the variables as defined by the investigator; hence, bias results. In other instances, the adjustment for confounders is done incorrectly, and bias is introduced inadvertently. Thus, to avoid statistical bias, epidemiologists, in addition to having an intimate understanding of the subject they are studying, must be knowledgeable about both statistical methods and proper confounder selection strategies.

It is important to keep in mind that all of these types of bias could be in either the positive (upward) or negative (downward) direction. However, one cannot dismiss the results of a study simply because there is a possibility of bias or confounding. Frequently, one can glean information that bears on the direction of bias. For example, if the factors tending toward downward bias are stronger than those that would magnify the association between exposure and disease, one would expect the true relationship to be

stronger than the one observed in the study.

This discussion about bias and precision can now be used to answer the problem of how to measure causal effects in groups. The key is that when certain conditions or requirements are met, the association between exposure and disease may be interpreted as a causal one or at least one can conclude that such an inference probably does not stray far from the truth. These conditions and requirements may be satisfied when (a) study subjects have been properly sampled and recruited;[39] (b) exposures and disease have been measured or diagnosed accurately;[40] (c) confounder data are complete and adequately measured; and (d) the appropriate multivariable statistical techniques have been used to analyze the data. Under these conditions, once all confounders have been accounted for, the unexposed group provides a good representation of the counterfactual experience of the exposed group and the analysis properly compares the group responses.

In other words, as long as the quality of data is reliable and the analysis is statistically correct and appropriately takes account of confounders, then the two groups (exposed and unexposed) can be validly compared. In this scenario, one can infer that the study RD or RR will be a measure of the *causal* effect of exposure. Of course, this measured causal effect may or may not be a precise estimate, as that will depend on whether the study has an adequate-sized sample. The quality of an individual study, therefore, depends on there being (1) minimum bias, which is achieved through careful design, sound methods of data collection and measurement of exposures and disease, and appropriate statistical treatment of the data; and (2) an adequate-sized study sample.

V. REACHING CONSENSUS

In practice, epidemiologic studies are never perfect, and even the best studies only approximately meet the necessary conditions for risk ratios or risk differences to be interpreted as causal effects. For this reason, it is nearly always true that causation cannot be

---

[39] This ensures a low probability of selection bias.

[40] This acts to reduce information bias.

inferred from a single study, but rather, must be examined in a multitude of studies. The problem is less acute when experimental (randomized) studies are possible, such as for the evaluation of drugs that are believed to impart a benefit to those taking them. In this situation, the evaluation of evidence is more straightforward than it would be for exposures for which it would be unethical to conduct such research (such as cigarette smoking or asbestos exposure). It is these latter, allegedly harmful exposures that have generated discussion about how to infer causation. This discussion has focused on how epidemiologists should evaluate a body of evidence from multiple studies, including human epidemiologic investigations, experimental data from whole animals, and mechanistic research in which cells or tissues are manipulated to understand physiologic or biochemical processes believed to be related to pathogenesis in the human body.

As the body of evidence grows and new hypotheses are proposed, the research community begins the process of reaching consensus regarding which studies and ideas it finds convincing. Arriving at a consensus can take months, years, or decades. For example, consensus regarding the role of the human immunodeficiency virus (HIV) in AIDS took a relatively short time, whereas the environmental contribution to breast cancer still remains contentious.

Consensus does not require and is not synonymous with unanimity. Even today it is possible to find some who are unconvinced about the relationship between HIV and AIDS or between smoking and lung cancer. That being said, the reaching of consensus often follows a typical pattern, in which evidence accrues and scientific opinion shifts. For example, consider a study that finds a previously unstudied association in which exposure E is related to an increased risk of disease D. To receive attention, the study often would have observed a strong association. Frequently, these first findings are based on a small sample size. Some scientists may reject these findings because they object to the study's methodology. Other researchers will then attempt to replicate the finding using improved methodology and maybe larger study samples, but it is possible that only some of the studies will confirm the original result. Over time, the weight of the

evidence will tend to fall on one side or the other. At some point, a meta-analysis or "quantitative review," which is a combined analysis of multiple studies, will be conducted. For this, it is preferable to use high quality studies as this type of analysis is more effective in addressing the precision of results than the biases. Meanwhile, toxicologic or other basic science studies may or may not demonstrate a plausible mechanism. Thus, the consensus will build either in support or in contradiction of a causal effect.

Although ideally scientists will evaluate evidence in a value-free context, it is increasingly recognized that it is impossible for scientists to be totally "objective" because individuals are unavoidably influenced by their particular cultures and personal experiences. Studies have documented how these experiences influence the way in which individuals assess scientific studies and place greater weight on certain studies or lines of evidence as compared with others. It should be noted that the IOM, in assembling its committees, consciously seeks to achieve not only diversity of fields of expertise, but also "balance" among possible biases on its committees.

Although criteria have existed for inferences about microbial causes of infections for more than a century, the discussion about causal inference for chronic diseases is more recent. In the 1960s, the debate as to whether cigarette smoking causes lung cancer provided the impetus for the development of a specific set of guidelines for inferring causality. These were summarized by Sir Bradford Hill[41] and include the following primary considerations:

1. *Temporality*: A cause must precede an effect.

2. *Strength of Association*: A high RR or RD provides greater weight than a low one.

3. *Coherence*: Evidence from other fields should support, not contradict, the causal hypothesis.

4. *Biologic Plausibility*: Known biologic facts should support, not contradict, the proposed causal effect.

---

[41] Austin Bradford Hill, *The Environment and Disease: Association or Causation?,* 58 PROC. ROYAL SOC'Y MED. 295 (1965).

5. *Consistency*: Multiple studies using different designs and/or different populations should confirm the finding.

6. *Dose-response*: The greater the exposure, the greater should be the likelihood of a response.

7. *Specificity*: The outcome should be less frequent in the absence of exposure or after removal of the exposure.

These considerations are not formal criteria and Hill himself cautioned against using them as such, although such misuse is often found in the scientific literature.[42] Moreover, it can be shown that failure to observe several of these facets of an association does not necessarily detract from the conclusion of causality. In fact, it has been argued that only temporality is truly required.[43]

VI. THE IOM COMMITTEE AND ITS EVALUATION OF SPECIFIC OUTCOMES FOR VIETNAM VETERANS

Applying the epidemiological principles discussed earlier in this article, this section reviews the evidence and reasoning behind the decisions reached by the Committee with regard to non-Hodgkin's lymphoma, Type II diabetes, prostate cancer, and the presumptive period for respiratory cancer.

In the case of non-Hodgkin's lymphoma, the first Committee, which completed its review in 1994, concluded that the evidence was sufficient regarding an association with herbicides or their contaminants.[44] A sizable number of studies in occupational cohorts had been conducted, and although many showed either no association (RR=1.0) or very slight associations, quite a few

---

[42] *See, e.g.*, Carl V. Phillips & Karen J. Goodman, *The Missed Lessons of Sir Austin Bradford Hill*, 1 EPIDEMIOLOGIC PERSPECTIVES & INNOVATIONS 3 (2004), *available at* http://www.epi-perspectives.com/content/pdf/1742-5573-1-3.pdf. These considerations have now spilled over into the courts as well. *See* Joe G. Hollingsworth & Eric G. Lasker, *The Case Against Differential Diagnosis:* Daubert*, Medical Causation Testimony, and the Scientific Method*, 37 J. HEALTH L. 85 (2004).

[43] Mervyn Susser, *Falsification, Verification, and Causal Inference in Epidemiology: Reconsiderations in the Light of Sir Karl Popper's Philosophy*, *in* CAUSAL INFERENCE (Kenneth J. Rothman ed., 1988).

[44] *See* VAO 1994, *supra* note 5.

studies showed an elevated risk of non-Hodgkin's lymphoma. These included studies of Swedish workers who were exposed to phenoxy herbicides;[45] forest conservationists who worked for the U.S. Department of Agriculture (RR=2.5, 95% CI=1.0 to 6.3);[46] farmers in Kansas who had used herbicides for more than twenty days per year (RR=6.0, 95% CI=1.9 to 19.5);[47] Canadian farmers who applied pesticides to more than 250 acres (RR=2.2, 95% CI=1.0 to 4.6);[48] Washington State forestry herbicide appliers (RR=4.8, 95% CI=1.2 to 19.4);[49] and Italian farmers licensed to use pesticides (RR=1.8, 95% CI=1.2 to 2.5).[50] In addition to these studies of occupational exposures, non-Hodgkin's lymphoma was increased among male residents of Italian provinces in contaminated areas (RR=2.2, 95% CI=1.4 to 3.5),[51] and in a Finnish community in which the water supply was contaminated with chlorophenols (RR=2.8, with a 95% CI=1.4 to 5.6).[52] Also, unlike many of the other health outcomes examined, non-Hodgkin's lymphoma was observed at a higher rate in Vietnam

---

[45] *See* Bodil Persson et al., *Malignant Lymphomas and Occupational Exposures*, 46 BR. J. IND. MED. 516 (1989); Lennart Hardell, *Malignant Lymphoma and Exposure to Chemical Substances, in Particular Organic Solvents, Chlorphenol and Phenoxyacetates*, 77 LAKARTIDNINGEN 208 (1980).

[46] *See* Michael C. Alavanja et al., *Mortality Among Forest and Soil Conservationists*, 44 ARCHIVES ENVTL. HEALTH 94 (1989).

[47] *See* Shelia Hoar et al., *Agricultural Herbicide Use and Risk of Lymphoma and Soft-Tissue Sarcoma,* 256 JAMA 1141 (1986), *erratum*, 256 JAMA 3351 (1986).

[48] *See* Donald T. Wigle et al., *Mortality Study of Canadian Male Farm Operators: Non-Hodgkin's Lymphoma Mortality and Agricultural Practices in Saskatchewan*, 82 J. NAT'L CANCER INST. 575, 579 Tbl.7 (1990).

[49] *See* James S. Woods & L. Polissar, *Non-Hodgkin's Lymphoma among Phenoxy Herbicide-Exposed Farmworkers in Western Washington State*, 18 CHEMOSPHERE 401 (1987).

[50] *See* G. Corrao et al., *Cancer Risk in a Cohort of Licensed Pesticide Users,* 15 SCANDINAVIAN J. WORK, ENV'T & HEALTH 203 (1989).

[51] Paolo Vineis et al., *Incidence Rates of Lymphomas and Soft-Tissue Sarcomas and Environmental Measurements of Phenoxy Herbicides*, 83 J. NAT'L CANCER INST. 362 (1991).

[52] P. Lampi et al., *Cancer Incidence Following Chlorophenol Exposure in a Community in Southern Finland*, 47 ARCHIVES ENVTL. HEALTH 167, 171 Tbl.5 (1992).

veterans than in the general population. As of the review conducted by the first IOM Committee, an excess of non-Hodgkin's lymphoma cases had been observed in several studies of U.S. Navy personnel (RR =2.2, 95% CI=1.2 to 3.9),[53] or Marine personnel (RR=2.1, 95% CI=1.2 to 3.8[54] and RR=3.2 95% CI =1.4 to 7.4[55]).

In total, more than two dozen studies showed some indication of excess mortality or incidence from non-Hodgkin's lymphoma. Not all of these studies were of the highest quality and there were some studies that showed no excess risk, that is, no significant departures from the expected level of risk. Although many of the studies cited above adjusted for potential confounders, such variables could have created the appearance of an association (increased the estimated RR) or could have obscured an association (reduced the estimated RR). In some of the studies, the definition of exposure was extremely broad and probably included a high proportion of individuals who were not exposed to any of the herbicides that were used in Vietnam, resulting in "information bias." In such circumstances, it would be easy to underestimate the effect of an exposure. In light of what might be an expected "downward" bias, the replication across quite a number of investigations that had an adequate sample size was impressive.

Neither the Seveso cohort[56] nor the chemical production workers[57] experienced increased risks for non-Hodgkin's

---

[53] The Selected Cancers Cooperative Study Group, The Association of Selected Cancers with Service in the U.S. Military in Vietnam III, Centers for Disease Control, *Hodgkin's Disease, Nasal Cancer, Nasopharyngeal Cancer, and Primary Liver Cancer*, 150 ARCHIVES INTERNAL MED. 2495 (1990).

[54] Patricia Breslin et al., *Proportionate Mortality Study of U.S. Army and U.S. Marine Corps Veterans of the Vietnam War*, 30 J. OCCUPATIONAL MED. 412, 416 Tbl.6 (1988).

[55] PATRICIA BRESLIN ET AL., VETERAN'S ADMINISTRATION, NON-HODGKIN'S LYMPHOMA AMONG VIETNAM VETERANS (1987).

[56] *See* Pier Alberto Bertazzi et al., *Ten-Year Mortality Study of the Population Involved in the Seveso Incident in 1976*, 129 AM. J. EPIDEMIOLOGY 1187 (1989); Angela C. Pesatori et al., *Cancer Morbidity in the Seveso Area, 1976-1986*, 25 CHEMOSPHERE 209 (1992).

[57] Fingerhut et al., *supra* note 12, at 216; Andreas Zober et al., *Thirty-Four-Year Mortality Follow-Up of BASF Employees Exposed to 2,3,7,8-TCDD after*

lymphoma. As these groups were most heavily exposed to 2,3,7,8-TCDD, with little or no exposure to the herbicides in Agent Orange, the epidemiologic data tended to suggest that the associations were more likely due to 2,4-D and 2,4,5-T. However, the Committee did not attempt to make the case that these compounds were the causal agents.

Biologic plausibility that Agent Orange was capable of producing this type of cancer was supported by a study that produced lymphoma in female mice after the administration of 2,3,7,8-TCDD.[58] However, the Committee noted that the herbicides contained in Agent Orange, including 2,4-D, 2,4,5-T, picloram, and cacodylic acid, had been inadequately tested in animals.

The conclusion of sufficient evidence drew on a set of studies that showed fair consistency. While not all studies could definitively exclude bias or confounding, it was unlikely that all of the studies were biased in the same direction. Moreover, in several investigations, the groups with the best-documented or highest probability of exposure showed the greatest increase in risk. Later studies confirmed the findings of excess risk for non-Hodgkin's lymphoma in yet other populations.[59]

Type II diabetes and prostate cancer are both characterized as

---

the 1953 Accident, 62 INST. ARCH. OCCUPATIONAL ENVT'L HEALTH 139 (1990); Alfred Manz et al., *Cancer Mortality among Workers in Chemical Plant Contaminated with Dioxin*, 338 LANCET 959 (1991).

[58] *See* James Huff et al., *Long-Term Carcinogenesis Studies on 2,3,7,8-tetrachlorodibenzo-p-dioxin and Hexachlorodibenzo-p-dioxins,* 7 CELL BIOLOGY AND TOXICOLOGY 67 (1991).

[59] *See, e.g.*, COMMITTEE TO REVIEW THE HEALTH EFFECTS IN VIETNAM VETERANS OF EXPOSURE TO HERBICIDES, INSTITUTE OF MEDICINE OF THE NATIONAL ACADEMIES, VETERANS AND AGENT ORANGE: UPDATE 1996 (1996) [hereinafter VAO UPDATE 1996]; COMMITTEE TO REVIEW THE HEALTH EFFECTS IN VIETNAM VETERANS OF EXPOSURE TO HERBICIDES, INSTITUTE OF MEDICINE OF THE NATIONAL ACADEMIES, VETERANS AND AGENT ORANGE: UPDATE 1998 (1999) [hereinafter VAO UPDATE 1998]; COMMITTEE TO REVIEW THE HEALTH EFFECTS IN VIETNAM VETERANS OF EXPOSURE TO HERBICIDES, INSTITUTE OF MEDICINE OF THE NATIONAL ACADEMIES, VETERANS AND AGENT ORANGE: UPDATE 2000 (2001) [hereinafter VAO UPDATE 2000]; VAO UPDATE 2002, *supra* note 8.

having "limited/suggestive" evidence of an association. For Type II diabetes, data were considered inadequate at the time the first three Committees evaluated the evidence. (The first two Committees considered the broader grouping of metabolic disorders as a whole, largely because little research had been published relating diabetes to the herbicides used in Vietnam or their contaminants.) Nevertheless, the third Committee, which published its findings in Update 1998, noted that a number of reports, including one on the Ranch Hand personnel, showed altered glucose metabolism. The Update reported, "Further analyses and full publication of existing studies may justify a reevaluation of this conclusion."[60] A flurry of papers appeared between 1996 and 2000 suggesting some association with 2,3,7,8-TCDD ("dioxin").[61] As a result, the fourth Committee, which published its result in 2000, determined that the evidence was limited, but suggestive of an association with exposures incurred in Vietnam. Among residents exposed to dioxin because of the industrial accident in Seveso, deaths from diabetes occurred at a higher rate than in the reference population that was not exposed, particularly among females.[62] Excess mortality from diabetes was

---

[60] VAO UPDATE 1998, *supra* note 59, at 11.

[61] *See, e.g.*, Geoffrey M. Calvert et al., *Evaluation of Diabetes Mellitus, Serum Glucose, and Thyroid Function among United States Workers Exposed to 2,3,7,8-tetrachlorodibenzo-p-dioxin*, 56 OCCUPATIONAL & ENVTL. MED. 270 (1999); Gary L. Henriksen et al., *Serum Dioxin and Diabetes Mellitus in Veterans of Operation Ranch Hand*, 8 EPIDEMIOLOGY 252 (1997); Angela C. Pesatori et al., *Dioxin Exposure and Non-Malignant Health Effects: A Mortality Study*, 55 OCCUPATIONAL & ENVTL. MED. 126 (1998); John Vena et al., *Exposure to Dioxin and Nonneoplastic Mortality in the Expanded IARC International Cohort Study of Phenoxy Herbicide and Chlorophenol Production Workers and Sprayers*, 106 ENVTL. HEALTH PERSPECTIVE 645 (Supp. 2 1998), *available at* http://ehp.niehs.nih.gov/members/1998/Suppl-2/645-653vena/vena.html; COMMONWEALTH DEP'T OF VETERANS' AFFAIRS, MORBIDITY OF VIETNAM VETERANS: A STUDY OF THE HEALTH OF AUSTRALIA'S VIETNAM VETERAN COMMUNITY, VOLUME 1: MALE VIETNAM VETERANS SURVEY AND COMMUNITY COMPARISON OUTCOMES (1998) [hereinafter COMMONWEALTH STUDY]; Morris F. Cranmer et al., *Exposure to 2,3,7,8-tetrachlorodibenzo-p-dioxin (TCDD) is Associated with Hyperinsulinemia and Insulin Resistance*, 56 TOXICOLOGICAL SCIENCES 431, 433 (2000).

[62] *See* Pier A. Bertazzi et al., *The Seveso Studies on Early and Long-Term*

observed in a multinational European cohort of chemical production workers,[63] although the excess was not statistically significant. No excess was observed by Steenland et al., who studied the U.S. cohort of chemical workers assembled by NIOSH.[64]

Typically, Type II diabetes is not fatal and is often not listed on a death certificate, even if one of its complications is the cause of death. For this reason, studies of mortality from diabetes would be limited in their ability to detect associations with exposures. By comparison, diagnoses among the living might provide a more complete ascertainment of cases, and hence, studies on morbidity would be considered more definitive. In one such study, self-reports of diabetes were substantially higher than expected in Australian veterans who served in Vietnam.[65] Among Air Force personnel who participated in the "Ranch Hand" study, glucose abnormalities and use of oral medications for diabetes were elevated.[66] Additionally, higher blood serum concentrations of 2,3,7,8-TCDD were associated with an elevated incidence of Type II diabetes.[67] Table 2 shows the risk ratios for men whose blood serum TCDD was in the three upper quartiles as compared with those whose blood serum TCDD was in the lowest quartile. The data do not show a perfect trend of increasing risk, but the upper two quartiles seem to be at higher risk than the lower two.

Effects of Dioxin Exposure: A Review, 106 ENVTL. HEALTH PERSPECTIVE 625 (Supp. 2 1998), *available at* http://ehp.niehs.nih.gov/members/1998/Suppl-2/625-633bertazzi/bertazzi.html.

[63] *See* Vena et al., *supra* note 60.

[64] *See* Kyle Steenland et al., *Cancer, Heart Disease, and Diabetes in Workers Exposed to 2,3,7,8-tetrachlorodibenzo-p-dioxin*, 91 J. NAT'L CANCER INST. 779, 785 (1999).

[65] *See* COMMONWEALTH STUDY, *supra* note 60.

[66] *See* Henriksen et al., *supra* note 60.

[67] *See* Matthew P. Longnecker et al., *Serum Dioxin Level in Relation to Diabetes Mellitus among Air Force Veterans with Background Levels of Exposure*, 11 EPIDEMIOLOGY 44 (2000).

Table 2: Incidence of Type II diabetes among Air Force Ranch Hand personnel according to blood serum concentration of dioxin, in quartiles.

Serum dioxin concentration:	1st quartile (lowest)	2nd quartile	3rd quartile	4th quartile (highest)
Risk ratio*	1	0.9	1.9	1.7
95% confidence interval	--	(0.5, 1.6)	(1.1, 3.2)	(1.0, 2.9)

*adjusted for family history, age, race, and military occupation

The confidence intervals (CIs) indicate that the data are consistent with anywhere between a rather small increased risk (RR just slightly above 1.0) and a fairly substantial one (a nearly three-fold higher risk). This study is notable in that the designation of diabetes was based on a clinical examination, not self-reporting. Additionally, a study conducted among residents near a hazardous waste site with dioxin contamination showed elevated risks for "high" fasting insulin if their serum TCDD concentration was elevated.[68] In general, the conclusion that the data showed limited/suggestive evidence of association was based on both the mortality and morbidity studies, with emphasis on the latter. The fact that some of these associations occurred in Vietnam veterans also weighed into the Committee's deliberations. Nevertheless, because many of the studies relied on self-reported illness, therefore raising the possibility of bias, the evidence fell far short of being sufficient.

A large number of studies have addressed the risk for prostate

[68] *See* Cranmer et al., *supra* note 60, at 431-33.

cancer.[69] Evaluation of this health outcome is difficult for several reasons. First, it is very common among elderly men, and second, most of the risk ratios are small (approximately 1.2). This is likely to occur when an outcome has multiple causes because no single cause is responsible for a high proportion of cases. Another factor to consider is the question of incidence versus mortality. Mortality is influenced by the aggressiveness of a tumor, but also by several other factors, including the quality of care, the treatment, and the stage at which the disease was diagnosed. In turn, these factors are affected by such variables as access to care and a patient's socioeconomic status. Thus, even if an exposure increases the incidence of prostate cancer, it may not show an association with mortality from prostate cancer because so much can intervene to alter survival after the occurrence of disease. Some of the early evidence used in the Committee's decision came from a well-conducted investigation of farmers or herbicide applicators, where greater exposures conferred higher risk,[70] and a number of occupational cohort studies in which risk was increased, but not significantly so. Additionally, the exposed population in Seveso showed an increased risk of prostate cancer.[71] In subsequent reviews of the evidence, the trend continued as many studies produced slightly elevated risk ratios while a few studies suggested a stronger association.

The Committee has, during updates of the reports, changed the classification of some of the health outcomes. For example, as mentioned above, diabetes was first classified as having inadequate evidence and then categorized as having limited or suggestive evidence of an association at the 2000 Update and by a separate committee convened to address this question on its own.[72] Although it has not happened yet, it is possible that the Committee could find the evidence regarding some outcome to be inadequate

---

[69] *See, e.g.*, VAO UPDATE 2002, *supra* note 8.

[70] *See* Howard Morrison et al., *Farming and Prostate Cancer Mortality*, 137 AM. J. EPIDEMIOLOGY 270 (1993).

[71] Bertazzi et al., *supra* note 56.

[72] *See* VAO UPDATE 2000, *supra* note 59; INSTITUTE OF MEDICINE, VETERANS AND AGENT ORANGE: HERBICIDE/DIOXIN EXPOSURE AND TYPE 2 DIABETES (2000).

after that disease was in the limited or suggestive category if newer studies were conducted that tended to show no association and were of higher quality than the earlier ones.

The fourth and final example of how the Committee has reviewed evidence concerns respiratory cancer and the "presumptive period." The VA had ruled that respiratory cancer could be considered service related only if it manifested within thirty years following one's service in Vietnam.[73] This thirty-year period was referred to as the "presumptive period." The Committee was asked to determine whether there was a scientific basis for this presumptive period. However, based on all of the empirical evidence from Vietnam veterans and other exposed populations, the question simply could not be answered. The analysis of time since the beginning of employment in exposed jobs suggested that the elevated risk for respiratory cancer might continue for at least the third decade. But this analysis begs the question, how long after an exposure *ends* will risk continue to be increased? Most occupational studies had not analyzed the mortality among cohorts of workers to determine whether excess risk of respiratory cancer changed with time since exposure ended. For the Seveso cohort, an insufficient period of time has elapsed to evaluate the thirty-year presumptive period (the accident occurred in 1976, fewer than thirty years prior to this writing). Thus, given the lack of pertinent epidemiologic data, the Committee relied on toxicokinetic data about how the chemicals of interest are stored in the body and on current understanding of the biology of human cancer. Dioxin is known to have a relatively long half-life in human tissues.[74] This TCDD half-life is estimated at between seven and nine years, but this period depends on the amount of fat in the studied individual[75]

---

[73] *See* Disease Associated with Exposure to Certain Herbicide Agents (Multiple Myeloma and Respiratory Cancers), 59 Fed. Reg. 29723-01 (June 9, 1994) (to be codified at 38 C.F.R. pt. 3).

[74] The half-life is the time it takes for the concentration to decrease to half of what it was.

[75] *See* Dieter Flesch-Janys et al., *Elimination of Polychlorinated Dibenzo-p-dioxins and Dibenzofurans in Occupationally Exposed Persons*, 47 J. TOXICOLOGY ENVTL. HEALTH 363, 377 (1996); Joel E. Michalek & Ram C. Tripathi, *Pharmacokinetics of TCDD in Veterans of Operation Ranch Hand: 15-*

and may differ between men and women. Hence, after external exposure ends, the compound remains in fatty tissue, circulates in the blood, and deposits itself in various organs. At any time during this period, disease induction can occur even though external exposure has ceased. In addition, disease detection may occur long after induction.[76]

Cancer progresses through multiple stages, beginning with initiation, the time at which a cell's DNA is damaged. The damaged cell then escapes the surveillance of the body's repair system and the immune system, which usually hunts out damaged cells. Other changes, known as promotion, may occur until the cell begins to divide unchecked, resulting in proliferation. Further stages enable the tumor to develop its own blood supply.

The point at which diagnosis occurs is determined by biologic, social, and individual psychologic factors. Biologic determinants will include the aggressiveness of the tumor, age of the person, and presence of other medical conditions that might influence immunologic competence. The social factors will include access to care, the quality of any screening program, and the skill and vigilance of the health provider. Individual characteristics that influence how early in the disease process a diagnosis is made include the propensity to seek medical care, which is highly variable in the population and is related to the degree of trust placed in the medical profession, and the fear of a diagnosis of cancer.

Given the above considerations, the Committee concluded that there was no scientific justification for a presumptive period of thirty years for respiratory cancer. The possibility that circulating TCDD might result in the initiation of cancer decades after a veteran's service in Vietnam had ended could not be excluded. A further consideration was the uncertain length of the latency period between the initiation of the disease process and the diagnosis.

Year Follow-Up, 57 J. TOXICOLOGY ENVTL. HEALTH 369, 376 (1999).

[76] The period between the start of a disease process and the time it is diagnosed is termed the "latent period."

SUMMARY

The IOM Committees were charged with determining whether there were associations between health outcomes and herbicides used in Vietnam or their contaminants. The IOM Committees addressed three questions: whether there was a statistical association between the exposures and any health outcomes, what magnitude of increased risk Vietnam veterans would be expected to experience for each of the health outcomes due to herbicide exposures incurred while in Vietnam, and whether evidence supported the biologic plausibility of a causal association. To answer the first question, the Committees classified the outcomes into four categories of evidence (sufficient, limited or suggestive, inadequate, or limited evidence of no association) and adopted an approach that weighed the body of evidence and took into consideration the methodologic rigor of the studies. With regard to the second question, that of quantifying the risk to Vietnam veterans, the Committee concluded that the increased risk could not be identified due to the lack of adequate data quantifying the exposures of those who served in Vietnam. To address the third question, that is, whether a plausible biologic mechanism exists through which the herbicides and their contaminants could cause specific health outcomes, the Committee evaluated a wide range of data types, including toxicologic studies in humans and experimental animals, and research on mechanisms that use tissues and cell cultures.

In reviewing the Committees' findings, it is important to remember that most non-infectious diseases are caused by multiple factors and that to determine the effects of exposure, causality is defined in an individual, but can only be measured in groups. Epidemiologists therefore study groups and, for ethical reasons, frequently rely on observational rather than experimental methods. The quality of observational studies depends on minimizing the four types of bias and maximizing precision by using large sample sizes (particularly with regard to the number of exposed cases of disease). Statistical significance is a small part of evidence, and the use of p-values for causal inference can result in faulty conclusions. Single studies can add to or detract from evidence for

causality, but ultimately an inference of causality depends on replication across studies that provide precise estimates of effects and that are relatively free of bias. Accrual of epidemiologic evidence over time, along with experimental studies in animals and cell or tissue cultures that establish mechanisms, generally leads towards a consensus as to whether an exposure causes a health outcome, although this process often takes years or longer. The evidence about health effects of herbicides used in Vietnam and their contaminants was slow to accumulate, partially because a concerted effort to study the veterans longitudinally, beginning from the time of their return to the United States, was not undertaken, and partially because it was technologically difficult to study dioxin, as it is present in such small quantities.