

The Value of Epidemiological Studies

Joel E. Michalek, PhD

Epidemiology Division, USAF School of Aerospace Medicine, Brooks Air Force Base, Texas

The ongoing Air Force Health Study, the U.S. Air Force investigation of health effects in Ranch Hand veterans exposed to Agent Orange and its contaminant, 2,3,7,8-tetrachlorodibenzo-p-dioxin (TCDD), is presented as a model epidemiologic study of occupational exposure to a toxic chemical. Three points are discussed: 1) The interpretation of the many statistical associations that can arise in an epidemiologic study requires careful consideration of the multiple testing artifact and established causal criteria. 2) Recently published work indicates that epidemiologic studies designed to demonstrate safety are, in practice, not feasible, and an observed relative risk of 1.0 in a study designed to detect hazard is not a valid basis for assurances of safety. 3) Work history indices of exposure effect are subject to error when the exposure is weak or the period of exposure is short; this error can lead to a strong bias toward finding no effect.

Introduction

The role of epidemiology in the resolution of health complaints arising from occupational exposure to advanced composites in manufacturing may be viewed as one step in a scientific process to assess whether adverse health effects exist and, if so, whether they can be attributed to the exposure. Preliminary to an epidemiologic effort, toxicologists and biologists will have studied specific effects in controlled animal experiments and will have hypothesized mechanisms and metabolic pathways for the toxin. Such prior knowledge is indispensable for the planning and conduct of epidemiological studies.

Given that an epidemiologic effort is being contemplated, three cautions must be kept in mind by policy makers and study planners. They are:

1. Large epidemiologic studies are statistical investigations, the results of which must be scrutinized with respect to established causality criteria.
2. Epidemiologic studies of occupational exposures are generally never large enough to establish safety.
3. Exposure misclassification can severely bias a study toward finding no effect when in fact a substantial health effect exists.

These well-known concepts are illustrated here with the Air Force Health Study, the U.S. Air Force investigation of health effects in Ranch Hand veterans occupationally exposed to "Agent Orange" in Vietnam.

Background: The Air Force Health Study

The Air Force Health Study (AFHS)⁽¹⁻³⁾ is designed to determine whether members of Operation Ranch Hand, the unit tasked with

herbicide spray operations during the Vietnam conflict, have experienced adverse health effects and whether those effects, if they exist, can be attributed to their occupational exposure to herbicides or their contaminant, 2,3,7,8-tetrachlorodibenzo-p-dioxin (TCDD). The AFHS was initiated by the Air Force in 1978 in response to a request by Congress that the Department of Defense conduct a long-term epidemiologic study of health effects in personnel exposed to herbicides. In 1980, the White House formally directed the Department of Defense to initiate a study of Ranch Hand veterans. This decision has subsequently been reaffirmed by succeeding administrations.

The AFHS is a 20-year prospective study of 1261 Ranch Hands and an equal number of matched Comparison Air Force veterans, who are matched on date of birth, race, rank, and occupation. The Comparisons were selected from the population of Air Force personnel who flew and maintained C-130 cargo aircraft in Southeast Asia during the same period, 1961 through 1972, that the Ranch Hand unit was active in Vietnam. These men were physically examined in the baseline year, 1982, and in 1985 and 1987. The next examinations will occur in 1992, 1997, and in the concluding year of the study, 2002. The study has three arms: morbidity, mortality, and reproductive effects. The morbidity arm consists of the physical examinations and associated interviewing and laboratory assays on the study participants. The mortality arm consists of annual mortality contrasts of the Ranch Hand cohort and the entire Comparison population of 19,101 individuals. The reproductive arm is an investigation of birth defects in all 7000 children fathered by the Ranch Hands and Comparisons seen in the physical examinations. The second follow-up examination data are currently being analyzed. A report will be released in early 1990. At the same time, Air Force investigators will analyze and report the reproductive effects study to be released in mid-1990.

A new dimension to the investigation has been added by the Centers for Disease Control (CDC). Early in 1987, CDC chemists developed a laboratory assay for TCDD in human serum which they validated against the well-established, but invasive, adipose tissue assay, showing that the two methods produce nearly identical results.⁽⁴⁾ Very soon thereafter, the Air Force collaborated with the CDC to assay 200 AFHS participants, 150 Ranch Hands and 50 Comparisons, to validate Ranch Hand exposures and, with frozen serum from the 1982 examination, estimate the half-life of TCDD in humans. The results⁽⁵⁾ show that the Ranch Hands still possess high body burdens of TCDD approximately 17 years after exposure and that the half-life of TCDD in Ranch Hands is approximately 7.1 years.⁽⁶⁾ The relatively long half-life means that most Ranch Hands are within two to three half-lives of their

Vietnam exposure. The CDC is currently assaying all Ranch Hands and Comparisons who complied with the blood draw during the second follow-up examination in 1987.

At the outset, the authors of the Protocol⁽¹⁾ identified many complications that would inhibit the detection of an effect if one did indeed exist. The sample size was limited to 1261 Ranch Hands. Thus, statistical power is fixed by nature, precluding study of rare diseases such as specific types of cancer and, especially, soft tissue sarcoma. They anticipated overt and subtle reporting biases that could, if not identified and circumvented, invalidate the results. No known disease endpoint was prespecified. Veteran complaints covered a broad range of medical and psychological conditions as well as a variety of adverse reproductive effects. The physical examinations, interviews, and laboratory assays are therefore wide ranging, producing hundreds of analyzable endpoints, each with its own set of risk factors. The reproductive effects investigation is based on the medical record verification of birth defects in every child fathered by the study participants; it also includes analyses of stillbirths, abortions, infant and prenatal mortality, and physical and mental impairments.

In the hypothetical case that there is no herbicide effect on health, about 5 percent of the many hundreds of statistical tests of hypothesis applied on the same data arising from this study will reject (produce p-values less than 0.05). This is known as the multiple testing artifact and it is common to all large studies. Unfortunately, there is no known statistical procedure that can distinguish between significant group differences that arise due to the multiple testing artifact and those which may arise due to a true herbicide effect. To guard against misinterpreting the multitude of findings, each analysis is interpreted with prior knowledge, concomitant information, and causality criteria.

The latency periods of adverse health effects, if they exist, are also unknown. Animal experiments have produced results sometimes conflicting with veteran complaints, complicating the interpretation of study results. The study is necessarily long (20 years) to ensure that a latent effect will not be missed if one exists.

Since there was no dosimetry for the Ranch Hands during their tours in Vietnam, there is no direct way to assess their exposure to herbicides or dioxin. Instead their exposure was indirectly approximated with an index based on work history data, following the example of other classic epidemiologic studies. The inadequacy of that index is now being realized.

Statistical Associations and Causality Criteria

Large epidemiological studies are necessarily statistical. Without a well-defined endpoint, investigators must compare exposed and control cohorts on dozens or even hundreds of medical conditions. Statistical analyses produce measures of association between exposure status (exposed, control) and each endpoint. Additionally, analyses are adjusted for covariates to reduce bias and variance. Analyses will be biased if certain covariates, termed confounders, are not taken into account. The inclusion of covariate information in an analysis also allows the investigation of the change in the exposure versus endpoint association with a covariate.

Due to the multiple testing artifact, investigators must assess many statistical associations to determine which are suggestive of a causal relationship between exposure and health effects and which ones are not. Among those that are statistically associated with exposure, some may be noncausally associated. Causal associations may be indirect or direct. An indirect causal association between a medical condition and an exposure occurs when the

exposure causes a change in the intermediate condition and that change causes the medical condition of interest to become manifest. For example, it may be conjectured that exposure to TCDD is indirectly causally related to heart disease through its ability to increase levels of cholesterol.

Interpretations require the combined efforts of medical doctors, statisticians, and subject matter specialists. A thorough interpretation will assess significant associations; the directionality of the findings, regardless of statistical significance; and changes in directionality or association with covariate information. Causality criteria have been widely discussed in the literature; see Kleinbaum, Kupper, and Morgenstern,⁽⁷⁾ for example. A minimal set of criteria is 1) time sequence, 2) strength of association, 3) dose-response, and 4) consistency. To support a causal argument, the exposure must have occurred earlier in the time sequence than the medical condition of interest. Even though a study will have identified a group of exposed individuals, the exposure may not have occurred during a fixed time period for some subjects or the medical condition may have precursors that occurred before the exposure. Causal associations may be stronger than noncausal associations, although strength of association will not be a reliable guide when the exposure is heterogeneous, of short duration, or expressed only after a long latency period. A dose-response relationship between exposure and a specific medical condition is sought via the development of an exposure index. Individuals with no exposure should experience fewer conditions than those subjects with low exposure and these, in turn, should have fewer conditions than heavily exposed subjects. In the absence of individual dose information, as is usually the case in studies of occupational exposure, studies rely on indirect indices of exposure, such as cumulative time on the job, to assess the dose-response relationship. Finally, if the association is to support a causal argument, it should be consistent with existing subject matter knowledge, usually derived from animal and laboratory experiments.

Proof of Safety Versus Proof of Hazard

In 1985, Bross presented minimal sample size criteria for proof of safety and for proof of hazard in studies of environmental and occupational exposures.⁽⁸⁾ His work is directed at rectifying widespread misconceptions about proof of safety that are prevalent in government agencies, in the medical and scientific establishments, and in other groups involved in occupational and public health and safety. He cites the erroneous notion that a failure to obtain statistically positive results in an epidemiologic study warrants a claim of safety, such as in the Environmental Protection Agency (EPA) interpretations of Love Canal data.⁽⁹⁾ The conclusion of his work is that it is far more difficult to provide a valid scientific proof of safety than to provide a corresponding proof of hazard. He shows that the quantity of data required for a valid assurance of safety is on the order of 30 times greater than that required for a valid proof of hazard. In fact, the size of the sample needed so far exceeds what is ordinarily available in epidemiologic studies, that assurances of safety given on the basis of such studies have no scientific validity. Bross's work was later refined and extended by Millard.⁽¹⁰⁾

Bross's work, summarized here in terms of relative risk, requires the simplifying assumptions that a specific change occurs in the environment or workplace at a known time in a given place within a stable population. The change might be an accident or a technological innovation in the workplace. The population at risk is assumed to be observed for equal time intervals before and after the event or, in studies with a control group, that the

person-times of follow-up in the two groups are equal. Let the adverse health effects be called "deaths." Let the number of deaths in the "before" period be x and in the "after" period be y . In controlled studies, y is the number of deaths in the exposed and x is the number of deaths in the control cohort. Let $z = x + y$ be the total deaths.

The usual statistical measure of the health effect of the workplace or environmental change would be the relative risk of death (y/x). Let the observed or sample value of the relative risk be RR and true value be T . Hence, if $T = 1$, the site or workplace would be safe, or as safe as it was originally. If there is hazard, T will be greater than 1. For example, a doubled risk would be given by $T = 2$.

Let A denote an "acceptable" relative risk, greater than 1.0, that would be permitted to declare an environment safe. There is general agreement that A should be about 1.10, indicating a 10 percent increase in deaths among the exposed. The choice of A is a societal and legal one; the value 1.10 is, according to Bross, founded in tort law and established scientific practice.

A standard statistical method to control false positives is to use the estimator RR to set a 95 percent confidence interval for the parameter T . With this method, we can be 95 percent sure that T lies in a specified range. If L is the lower limit of this interval and U is the upper limit, we can be 95 percent confident that $L < T < U$.

To demonstrate safety, we would want to argue that it is very unlikely that the true relative risk is greater than the acceptable relative risk A . In these terms, safety would be (statistically) proved if $L < T < U < A$.

To demonstrate hazard, we would want to argue that it is very likely that the true relative risk is greater than the acceptable relative risk A . In these terms, hazard would be (statistically) proved if $A < L < T < U$.

The minimal statistical requirement for a valid proof of safety is that the square root of z , \sqrt{z} , be at least as large as the right-hand side of Equation 1.

$$\sqrt{z} = (RR + 1)(A + 1)/(A - RR), \quad RR < A \quad (1)$$

while the corresponding requirement for a valid proof of hazard is that \sqrt{z} be at least as large as the righthand side of Equation 2.

$$\sqrt{z} = (RR + 1)(A + 1)/(RR - A), \quad RR > A \quad (2)$$

The two requirements are symmetric. The requirement that RR be less than A for the application of the requirement for safety agrees with common sense in that one would not be interested in proving safety when the observed relative risk was greater than the acceptable relative risk. Similarly, one would not want to prove hazard when the observed relative risk was less than the acceptable value.

While the value of RR depends on the particular study, we can get an idea of the order of the magnitude of z by using the numerical value of T as a surrogate for RR in these equations. Substituting $A = 1.10$ and $RR = 1.0$ in Equation 1 gives $\sqrt{z} = 42$, or $z = 1764$. Thus, if the observed relative risk is less than the acceptable relative risk A , one would require at least 1764 deaths to be 95 percent confident that the true relative risk is less than the acceptable relative risk A . Substituting $A = 1.10$ AND $RR = 2.0$ in Equation 2 gives $\sqrt{z} = 7$ or $z = 49$. Hence, if $RR = 2$ one would require at least 49 deaths to be 95 percent confident that the true relative risk exceeds the acceptable relative risk A .

An appreciation of the sample sizes required to produce 1764

deaths can be gained from data derived from the AFHS. In the recently released 1989 mortality update,⁽¹¹⁾ the overall cumulative death rate in both Ranch Hands and Comparisons combined was about 2.8 deaths per 1000 person-years. The observed overall relative risk, RR , was 1.0. Suppose one wanted to design a new study of these populations to demonstrate safety. Bross's minimal requirement is 1764 total deaths with $RR = 1.0$ and $A = 1.1$. Let N denote the total person-years of follow-up required to yield 1764 deaths in both groups. One would then have a $2.8 \cdot N/1000 = 1764$ or $N = 630,000$ person-years of follow-up and, in a study with equal group sizes, $630,000/2 = 315,000$ person-years of follow-up per group. Since the average time since Vietnam exposure is 17 years, the resultant minimal sample size per group would be $315,000/17 = 18,529$. Thus, to make assurances of safety with 95 percent confidence, having observed $RR = 1.0$, one would require at least 18,529 Ranch Hands and an equal number of Comparison subjects. This is an impossibility since there are only 1261 Ranch Hands.

The sample size requirement for demonstration of hazard is far less severe, as can be seen by repeating the previous example with $z = 49$, assuming $RR = 2$ was of interest. In that case, the minimal requirement is 515 subjects per group, which is, of course, exceeded in the AFHS. Thus, with regard to overall mortality, the AFHS is large enough to prove hazard but not large enough to prove safety.

Exposure Misclassification and its Consequences

In the absence of dosimetric data, epidemiologic investigators have generally used work history information to index exposure. For example, in a mortality study of male workers exposed to airborne arsenic trioxide and sulfur dioxide in a Montana smelter, Lee and Fraumeni⁽¹²⁾ used the number of years worked in moderate and heavy arsenic areas to index exposure. Similar indices have been used in studies of asbestos and chemical exposures. Such indices, although crude because they ignore individual variation and work habits, can suffice to demonstrate a dose-response effect, as was the case with Lee and Fraumeni and many other studies of occupational exposures to toxic substances and chemicals.

Following these and other examples, the AFHS indexed Ranch Hand exposure to TCDD by E , given by $E = C \cdot G/P$, where C was the concentration of TCDD in the herbicides sprayed during the subject's tour and P was the number of personnel in the subject's job specialty during his tour. This index was prescribed in the study Protocol as the best index, given available data. Self-reported exposures have been avoided to preclude the possibility of reporting bias.

An assessment of the validity of E as a measure of TCDD exposure has recently become possible since the development of the serum TCDD assay at the CDC. The ongoing CDC assay of AFHS participants allowed a display of the relationship between E and current TCDD body burden. Additionally, the half-life estimate together with known times since tour and the assumption of exponential decay permits a study of the relationship between E and the estimated initial Ranch Hand TCDD dose.

The assay results indicate that, as a group, the Ranch Hands have been exposed to TCDD and that, as a group, the Comparisons are unexposed (Table I).

In Table I, current TCDD results are shown for each of five occupational strata as well as for all assayed Ranch Hands and Comparisons. All but 2 of 385 assayed Comparisons have a current TCDD body burden less than 15 parts per trillion (ppt), levels

TABLE I. Serum TCDD Results

Occupational Stratum	Ranch Hand			Comparison		
	Sample Size	Median*	Range*	Sample Size	Median*	Range*
Flying Officers (pilot)	157	7	0-43	94	4	2-13
Flying Officers (navigator)	39	9	1-36	22	5	2-8
Nonflying Officers	14	7	3-25	3	4	4-5
Flying Enlisted	98	16	1-127	68	4	1-13
Nonflying Enlisted	312	23	0-313	198	4	0-26
All Personnel	620	13	0-313	385	4	0-26

*In parts per trillion.

that are considered background trace amounts; the Comparison median is 4 ppt. In contrast, 44.8 percent of 620 assayed Ranch Hands have current values above 15 ppt; the Ranch Hand median is 13 ppt. If the threshold for background exposure is taken as 10 ppt, as suggested by this and the CDC ground troop study,⁽¹³⁾ 59.8 percent of Ranch Hands and 2.6 percent of assayed Comparisons have current TCDD levels above background.

However, a plot of E versus current TCDD body burden in the 620 assayed Ranch Hands (Figure 1) shows no association; correlation = -0.07. Further, no association is seen between E and extrapolated Vietnam TCDD dose, correlation = -0.05, or between the logarithms of these quantities.

These results will be fully described when all Ranch Hands have been assayed. The lack of association between E and current TCDD body burden may be due to the short duration of exposure, about one-year for most Ranch Hands, and variation in individual work habits and duty. These aspects are currently under investigation.

The TCDD assay results so far indicate that E is not a valid measure of current or extrapolated initial TCDD body burden in Ranch Hands, diminishing the validity of all previous attempts to detect a dose-response relationship with E. The entire study

will be reanalyzed with the TCDD assay results, and the extrapolated Vietnam TCDD dose, as the indicators of exposure. This reanalysis is scheduled to begin in September 1989. The results will be released at the conclusion of a one-year analysis and report writing period.

About 40 percent of assayed Ranch Hands have current TCDD levels below 10 ppt, a level that may be regarded as an upper limit for background exposure. Without additional data, we can only assume that these Ranch Hands were not exposed in Vietnam or that, in the worst case, they were exposed and their body burdens have decayed to background levels. The reanalysis of study data will take both possibilities into account. If they were, in fact, not significantly exposed in Vietnam, current estimates of relative risk in the AFHS are biased toward finding no effect.

The magnitude of the bias due to misclassifying exposed subjects can be assessed in terms of the bias of estimated odds ratio, a quantity sometimes estimated by statisticians instead of the relative risk. The odds ratio approximates the relative risk for rare diseases. In the case that only about 60 percent of the Ranch Hands were exposed to TCDD in Vietnam, if the true odds ratio or relative risk were 2, one would estimate an odds ratio of about 1.1 and thus miss finding the health effect, assuming 1000 subjects in each group, a disease prevalence of 5 percent in the Comparison group and an exposure prevalence of 2 percent. If the true effect were a tripling of disease prevalence, an odds ratio or relative risk of 3, the estimated value would be about 1.2. Thus, with misclassification as high as 40 percent, a doubling or tripling of disease prevalence could be missed in a study as large as the AFHS. These bias estimates and their consequences are being avoided in the AFHS via the introduction of assay results as the exposure index.

Conclusion

In the context of occupational exposures to advanced composites, epidemiologic studies are statistical investigations of health effects in human beings that can complement animal experiments in the resolution of health complaints. The prospective AFHS has been discussed as an exemplary study of health effects in a cohort

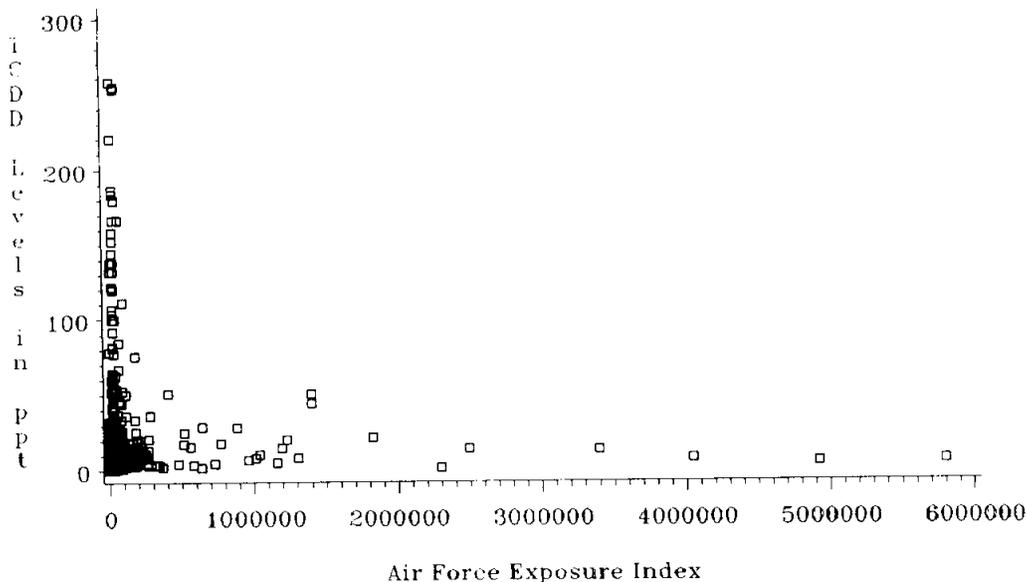


FIGURE 1. Current serum TCDD levels versus Air Force Exposure Index (N = 620) Ranch Hand personnel.

occupationally exposed to herbicides and their contaminant (TCDD).

The interpretation of the many statistical associations that can arise in an epidemiologic study requires careful consideration of the multiple testing artifact and established causal criteria. In studies with many endpoints, such as the AFHS, interpretation is challenging and not always conclusive due to conflicting prior knowledge and unknown latency periods.

Bross's⁽⁸⁾ calculations show that epidemiologic studies designed to demonstrate safety are, in practice, not feasible. Further, an observed relative risk of 1.0 in a study designed to detect hazard is not a valid basis for assurances of safety.

Work history indices of exposure, while sufficient to detect a dose-response effect in past studies of occupational exposures to toxic chemicals, are subject to error when the exposure is weak or the period of exposure is short. Additionally, exposures can be highly heterogeneous, as was TCDD exposure among Ranch Hands, and this can lead to a strong bias toward finding no effect.

This discussion has been centered around the prospective AFHS as the example. Case-control studies focused on a single disease endpoint and a single exposure are less prone to the multiple testing artifact, but they are still subject to issues of exposure index error. Bross's calculations apply to case-control studies as well as to prospective studies.

References

1. Lathrop, G.D.; Wolfe, W.H.; Albanese, R.A.; Moynahan, P.M.: The Air Force Health Study. An Epidemiologic Investigation of Health Effects in Air Force Personnel Following Exposure to Herbicides: Study Protocol. National Technical Information Service AD A 122 250. NTIS, Springfield, VA (1982).
2. Lathrop, G.D.; Wolfe, W.H.; Albanese, R.A.; Moynahan, P.M.: The Air Force Health Study. An Epidemiologic Investigation of Health Effects in Air Force Personnel Following Exposure to Herbicides: Baseline Morbidity Study Results. National Technical Information Service AD A 138 340. NTIS, Springfield, VA (1984).
3. Lathrop, G.D.; Machado, S.; Grubbs, W.; Karrison, T.; et al: The Air Force Health Study. An Epidemiologic Investigation of Health Effects in Air Force Personnel Following Exposure to Herbicides: First Follow-up Examination Results. National Technical Information Service AD A 188 262. NTIS, Springfield, VA (1987).
4. Patterson, Jr., D.G.; Needham, L.L.; Pirkle, J.L.; Roberts, D.W.; et al: Correlation between serum and adipose tissue levels of 2,3,7,8-tetrachlorodibenzo-p-dioxin in 50 persons from Missouri. *Arch. Environ. Contam. Toxicol.* 17:139-143 (1988).
5. Wolfe, W.H.; Michalek, J.E.; Miner, J.C.; Peterson, M.R.; et al: Serum 2,3,7,8-tetrachlorodibenzo-p-dioxin Levels in Air Force Health Study Participants—Preliminary Report. *Morbidity and Mortality Weekly Report* 37:309-311 (1988).
6. Pirkle, J.L.; Wolfe, W.H.; Patterson, Jr., D.G.; Needham, L.L.; et al: Estimates of the Half Life of 2,3,7,8-tetrachlorodibenzo-p-dioxin in Vietnam Veterans of Operation Ranch Hand. *J. Toxicol. Environ. Health* 27:165-171 (1989).
7. Kleinbaum, D.G.; Kupper, L.L.; Morgenstern, H.: *Epidemiological Research: Principles and Quantitative Methods*. Life-time Learning Publications, Belmont, CA (1982).
8. Bross, I.D.: Why Proof of Safety is Much More Difficult than Proof of Hazard. *Biometrics* 41:785-793 (1985).
9. U.S. Environmental Protection Agency: *Environmental Monitoring at Love Canal, Vol 1*. U.S. Government Printing Office, Washington, DC (1982).
10. Millard, S.P.: Proof of Safety Versus Proof of Hazard. *Biometrics* 43:719-725(1987).
11. Wolfe, W.H.; Michalek, J.E.; Miner, J.C.: The Air Force Health Study. An Epidemiologic Investigation of Health Effects in Personnel Following Exposure to Herbicides. Mortality Update. National Technical Information Service, Springfield, VA (1989).
12. Lee, A.M.; Fraumeni, Jr., J.F.: Arsenic and Respiratory Cancer in Man: An Occupational Study. *J. Natl. Cancer Inst.* 42:1045-1052 (1969).
13. Centers for Disease Control: Serum Dioxin in Vietnam-era Veterans—Preliminary Report. *Morbidity and Mortality Weekly Report* 36:470-475 (1987).