

CHAPTER 21

INTERPRETIVE CONSIDERATIONS

INTRODUCTION

Careful consideration of bias, interactions, consistency, multiple testing, dose-response patterns and the exposure index, trends, power limitations, strength of association, and biologic credibility is essential to the interpretation of these data. Problems inherent in the evaluation of negative results and the summarization of these data should also be considered.

BIAS

At the 1987 followup examination, 995 of 1,188 eligible Ranch Hands (84%) and 1,299 of 1,729 eligible Comparisons (75%) were fully compliant. Therefore, differential compliance and the potential for compliance bias existed. The subcohorts of fully compliant participants have remained fairly stable across study examinations. The percentages of those who were fully compliant at the 1987 followup examination and at the Baseline examination were similar across groups (92% of Ranch Hands and 93% of Comparisons). Detailed analyses of available data indicate that those who participated did not differ from those who refused and these contrasts did not change with group membership. Thus, it is concluded that there is no detectable compliance bias in this study and this form of bias is excluded as an explanation of the results.

Information bias, represented by the possible overreporting of disease symptoms, was precluded by medical record verification of major disease conditions. The possibility still exists that Ranch Hand conditions may be more verifiable because they might tend to be seen by physicians more often than Comparisons; this would be revealed by group differences in the quantity and content of medical records. Since there is currently no way to quantify these aspects, this potential bias remains unexplored. Information bias due to errors in the data base introduced via data entry or machine error is negligible. All laboratory results were subject to strict quality control procedures and medical coding data were completely verified by medical record review. The misclassifications of a Ranch Hand by race and 13 participants by verified history of diabetes are inconsequential, as shown by repeated analyses of data with these mistakes corrected.

Misclassification bias is a definite possibility with regard to the classification of Ranch Hands according to the calculated dioxin exposure index. Recent, and as yet unpublished, serum dioxin assay results suggest that there is no relationship between current dioxin levels and the calculated index. Current dioxin levels are, however, strongly associated with occupation, with enlisted groundcrew having the highest and officers having the lowest levels; enlisted flyers have current dioxin levels lower than enlisted groundcrew and higher than the officers. Thirty percent of the flying officers and 76 percent of enlisted groundcrew have levels above background (10 ppt). Thus, the exposure index analyses presented in this and

previous reports may be biased toward finding no effect. The actual extent of this bias will be fully described in another report after all assay results are available.

Since 12 percent of assayed Ranch Hands (n=848) have current dioxin levels below 4 ppt, the approximate Comparison median (n=384), the group contrasts in this report may also be biased toward finding no effect. With 12 percent of Ranch Hands misclassified as exposed, a true relative risk of 2.0 would be estimated as approximately 1.1 and would thus be missed (assuming a disease prevalence of 5% in the Comparison group, equal sample sizes of 1,000 in each group, and a population probability of exposure of 2%). It is possible, however, that Ranch Hands having background levels today may have actually been exposed but their body burdens have decayed to the current level. If this is the case, there would be no misclassification bias regarding the group contrasts. Both cases will be addressed in a reanalysis of these data with the dioxin assay serving as the measure of exposure.

ADJUSTMENTS FOR COVARIATES AND INTERACTIONS

The matched design together with extensive covariate adjustments were implemented to preclude the possibility of confounding. Lack of adjustment for a confounder could hide an otherwise significant group difference or reveal a spurious difference. Adjusted and unadjusted results were presented to reveal the effect modification of the covariates. The presence of significant interactions with group, that is, a significant difference in the relative risk with levels of a covariate, precluded the presentation of an overall adjusted relative risk and, instead, a stratified analysis was conducted to describe the interaction. When the p-value was between 0.01 and 0.05, the data were analyzed with and without the interaction. If an interaction was significant at the 0.01 level or less, the analysis was stopped with a description of the corresponding stratified analysis. The large number of dependent variables in this study (approximately 300) and covariates produced many significant interactions, all of which were listed and summarized in each clinical chapter. Review of these interactions within and across clinical areas revealed no overall patterns. Additionally, since occupation is currently the best correlate with current dioxin levels, a difference in relative risk with levels of occupation (with relative risk among enlisted ground personnel being greater than the relative risk among officers) would support a dose-response effect. The lack of such an interaction would argue against a dose-response effect.

CONSISTENCY

Ideally, an adverse health effect in Ranch Hands attributable to herbicide or dioxin exposure would be revealed by internally and externally consistent findings. A finding would be regarded as internally consistent if it did not contradict prior information, other findings, or medical knowledge. For example, the finding of significantly increased femoral pulse abnormalities is not consistent with the lack of increased posterior tibial pulse abnormalities in Ranch Hands. Further, the lack of interaction with occupation is not consistent with known patterns of dioxin levels in Ranch Hands. A finding would be externally consistent if it had been previously

established either in theory or empirically as related to exposure. The observed excess of basal cell carcinoma in Ranch Hands is externally inconsistent since there is no prior evidence that basal cell carcinoma is related to dioxin or herbicide exposure. It is also internally inconsistent because there is no evidence that basal cell carcinoma relative risk is greater among enlisted ground personnel than the relative risk among officers.

MULTIPLE TESTING

The lack of a predefined medical endpoint has necessitated the consideration of literally hundreds of dependent variables. Each dependent variable is analyzed many different ways to accommodate covariate information and different statistical models. In the hypothetical case that Ranch Hand physical health is the same as that of the Comparisons, about 5 percent of the many statistical tests of hypotheses shown in this report should be expected to detect a group difference (produce p-values less than 0.05). The observation of significant results due to multiple testing, even when there is no group difference, is known as the multiple testing artifact and is common in large studies. Unfortunately, there is no statistical procedure available to distinguish between those statistically significant results that arise due to the multiple testing artifact and those that may be due to a bona fide herbicide effect. Instead, the authors have relied on reasoned consideration of strength of association, consistency, dose-response patterns, and biologic credibility to weigh and interpret the findings.

DOSE-RESPONSE PATTERNS AND THE EXPOSURE INDEX

Ideally, a dose-response effect would be revealed by a regression of disease prevalence on exposure. The most obvious effect would be represented by an increasing trend in disease prevalence from a low rate among Ranch Hands with low exposure to a high rate among Ranch Hands with a high exposure. A dose-response effect may be expected to occur regardless of the presence or absence of a group difference.

Epidemiologic studies of health effects after environmental or occupational exposure to toxic chemicals or substances have generally relied upon indirect measures of exposure, termed exposure indices, to assess dose-response. For example, Lee and Fraumeni¹ studied respiratory cancer mortality in Montana smelter miners exposed to airborne arsenic trioxide and sulfur dioxide. The exposure index for an individual miner was simply the number of years of employment. With it, a statistically significant dose-response effect was demonstrated. In the aborted Centers for Disease Control (CDC) study of health effects in U.S. Army troops potentially exposed to Agent Orange in Vietnam, study investigators derived several exposure indices in terms of troop locations, known half-lives of dioxin in soil and on plant leaves, and the dates and spray paths of Ranch Hand aircraft. The study was canceled after their exposure indices failed to correlate with current dioxin levels in assay study subjects. In the Air Force Health Study (AFHS), each Ranch Hand's dioxin exposure was metricized as the product of the gallons of herbicide sprayed during his tour and the dioxin concentration of that herbicide divided by the number of Ranch Hands in his job category during his tour. This exposure index has so far failed to reveal consistent dose-response effects and is not correlated with current dioxin body burden in Ranch Hands. It has also failed to correlate with the extrapolated Vietnam dioxin dose in Ranch Hands assuming first order kinetics and a half-life in humans of 7.1 years.²

The AFHS exposure index was based on the best information available during the design phase of this study. The gallons sprayed, dioxin concentrations, and personnel figures are considered accurate. The index is based on the logic that exposure should increase with increased spraying or if fewer men in an occupational category became available to do the work. Similarly, it was reasoned that exposure should decrease as spraying decreased or as more men became available to do the work. The validity of this index is limited, however, since the gallons sprayed and personnel figures are not specific to an individual Ranch Hand's assigned base in Vietnam or to his specific daily work schedule. The AFHS exposure index is probably more accurate than the indices attempted by the CDC because the Ranch Hands were much closer to the herbicide than the Army and recorded troop locations were somewhat inaccurate for the individual soldier. Indirect exposure indices based on work history and demographic information have demonstrated significant dose-response effects in studies of long-term occupational exposure with moderate to high relative risks. Such indices have failed to demonstrate significant effects or have failed to correlate with direct measures of exposure, such as the dioxin assay, when exposures are short in duration, are of less than industrial intensity, or when the relative risk is small.

Fortunately, the development of the serum dioxin assay and its application to Ranch Hands and Comparisons will obviate the concern about the calculated exposure index.

TRENDS

An assessment of consistent and meaningful trends is an essential element of the interpretation of any large study with multiple endpoints, clinical areas, and covariates. However, caution must be exercised in the interpretation of trends.

Increased abnormalities or adverse means for the Ranch Hands across medically related variables within a clinical area might indicate an exposure effect. In this case, it is important to note that there is moderate to strong correlation between endpoints. Hence, the strength of the group differences must also be considered in assessing the extent of the suspected exposure effect.

Based on preliminary results, current dioxin levels are strongly associated with occupation. Thus, strong, statistically significant differences between groups in means or percent abnormalities for different occupations (i.e., group-by-occupation interactions) would be indicative of a dose-response effect. In this situation, one would expect to see a steadily increasing relative risk or difference between means as occupational exposure increased (i.e., officers less than enlisted flyers less than enlisted groundcrew). Under these assumptions, significant group-by-occupation interactions would be expected for clinical endpoints affected by dioxin exposure. The lack of a significant interaction with occupation could be due to the absence of a true effect, or the power limitations of the statistical test for interactions.

An increasing trend in differences between groups in means or disease rates with levels of a covariate (other than occupation) could also indicate an exposure effect. For example, an increased relative risk for hepatic disease with increased levels of alcohol consumption could be due to an indirect causal relationship between exposure and hepatic disease through alcohol consumption. In assessing potential indirect causal relationships, it is important to consider the strength of the group differences and consistency of both the results with related endpoints and findings over time (i.e., 1982 Baseline, 1985 followup, 1987 followup examinations).

Based on the calculated exposure index, increasing trends in Ranch Hand disease rates with increasing levels of exposure within occupational category would be expected in the presence of an exposure effect. However, preliminary results of serum dioxin assays of the Ranch Hands indicate that the calculated exposure index is not a good measure of actual dioxin exposure. Thus, the results of the exposure index analysis should be interpreted with caution.

POWER LIMITATIONS

The fixed size of the Ranch Hand cohort limits the ability of this study to detect group differences. This limitation is most obvious with regard to specific types of cancer, such as soft tissue sarcoma and non-Hodgkin's lymphoma, which are so rare that fewer than one case is expected in each group and, therefore, this study has virtually no statistical power to detect low to moderate group differences regarding them. On the other hand, these sample sizes are sufficient to detect very small mean shifts in the continuously distributed variables. For example, with regard to IgG, this study has approximately 90 percent power to detect a mean shift of 1 percent. The detection of significant mean shifts without a corresponding indication of increased Ranch Hands abnormalities or disease is considered to be of little importance or an artifact of multiple testing. This study has good power to detect relative risks of 2.0 or more with respect to diseases occurring at prevalences of at least 5 percent in the Comparison group, such as heart disease and basal cell carcinoma.

In an attempt to overcome the lack of power to detect group differences for specific types of systemic cancer, all types of systemic cancer were combined into a single variable. It is still possible, however, that an increased risk could exist for a particular rare type of cancer and that increased risk would be missed in this study.

STRENGTH OF ASSOCIATION

Ideally, an adverse effect, if it exists, would be revealed by a strong association between group and a disease condition, that is, by a statistically significant relative risk greater than 2.0. Statistically significant relative risks less than 2.0 are considered of less importance than larger risks because relative risks less than 2.0 can easily arise due to unperceived bias or confounding; relative risks greater than 5.0 are less subject to this concern. Statistically significant relative risks greater than 5.0 were not found in this study.

BIOLOGIC CREDIBILITY

The assessment of biologic credibility requires consideration of the question: In biologic terms, is it understood how the exposure under study could produce the effect of interest? While lack of biologic credibility or even a contradiction of biologic knowledge can sometimes lead to dismissal of a significant result as spurious, the failure to perceive a mechanism may reflect only ignorance of the state of nature.³ On the other hand, it has proven all too easy to propose credible biological mechanisms relating most exposures to most cancers. Thus, while pertinent, the response to this question is not especially convincing one way or the other.³

INTERPRETATION OF NEGATIVE RESULTS

In 1985, Bross presented minimal sample size criteria for proof of safety and for proof of hazard in studies of environmental and occupational exposures.⁴ 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 public health and safety. He cites the erroneous notion that failure to obtain statistically significant results in an epidemiologic study warrants a claim of safety, such as in Environmental Protection Agency interpretations of Love Canal data.⁴ His work concludes 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 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' work was later refined and extended by Millard.⁵ Michalek⁶ has recently applied Bross' methods to demonstrate that the AFHS is large enough to demonstrate hazard (for disease prevalences on the order of 5%), but not large enough to prove safety.

SUMMARIZATION OF RESULTS

Many readers will attempt to tally statistically significant results across clinical areas and study cycles. A study of this scope having a multitude of endpoints and no prescribed strength of association to declare an effect meaningful demands, and at the same time defies, meaningful summary tabulation. Such summaries are misleading because they ignore correlations between the endpoints, correlations between study cycle results, and the nonquantifiable medical importance of each endpoint. In fact, many endpoints are redundant (e.g., psychological scales, indices developed from combining multiple variables) in an effort not to "miss" anything. Additionally, such tabulations combine endpoints that are not medically comparable. For example, sense of smell is of less medical importance than the presence of malignant neoplasm. Statisticians attempt to summarize multidimensional repeated measures data with growth curve analyses; these methods have not been applied in this study because they apply only to continuously distributed data, do not account for medical importance, and reduce the data "too much."

Nevertheless, given the lack of adequate summary statistics, the tally of significant results will occur. Such summaries can be misleading and must be carefully interpreted.

OTHER ANALYTICAL STRATEGIES

The analytical plan for this report was written before Ranch Hand 2,3,7,8-tetrachlorodibenzo-p-dioxin (TCDD) results became available. Other analyses, such as restriction to enlisted groundcrew, were not carried out, although such analyses appear now to be well motivated in view of the TCDD concentrations in that occupation. The analytical strategy for this and previous reports was conceived during protocol development with no knowledge of the relative exposures of the three occupational categories of Ranch Hands. At that time, some investigators speculated that the enlisted flyers were the most heavily exposed to TCDD. The accomplishment of within occupational strata analyses at this time would constitute another attempt, as was our inspection of group-by-occupation interactions, to use occupation as a surrogate for TCDD exposure. The next report, already in progress, will show the results of analyses of all health conditions against current TCDD concentrations in Ranch Hands. Current health in Ranch Hands will also be assessed relative to the extrapolated Vietnam TCDD dose using a first-order kinetic assumption. Additionally, Ranch Hands having high current TCDD concentrations will be contrasted with Comparisons having background TCDD levels. Therefore, continued analysis of these data without accounting for TCDD concentrations is not warranted.

CONCLUSION

The interpretation of the AFHS requires careful consideration of potential biases, interactions, consistency of results, the multiple testing artifact, dose-response patterns and the exposure index, trends, power limitations, strength of association, and biologic credibility. Additionally, any assurances of safety drawn from these data are not scientifically valid and should be avoided. The AFHS is large enough to establish hazard (for disease prevalences on the order of 5%), but is not large enough to establish safety. Simple tabulations of positive results can be misleading.

CHAPTER 21

REFERENCES

1. Lee, A.M., and J.F. Fraumeni, Jr. 1969. Arsenic and respiratory cancer in man: an occupational study. JNCI (42):1045-1052.
2. Pirkle, J.L., W.H. Wolfe, D.G. Patterson, L.L. Needham, J.E. Michalek, J.C. Miner, M.R. Peterson, and D.L. Phillips. 1989. Estimates of the half life of 2,3,7,8 tetrachlorodibenzo-p-dioxin in Vietnam veterans of Operation Ranch Hand. Journal of Toxicology and Environmental Health (27):165-171.
3. Breslow, N.E., and N.E. Day. 1980. Statistical methods in cancer research, volume 1. Lyon, IARC.
4. Bross, I.D. 1985. Proof of safety is much more difficult than proof of hazard. Biometrics (41):785-793.
5. Millard, S.P. 1987. Proof of safety versus proof of hazard. Biometrics (43):719-725.
6. Michalek, J.E. 1989. The value of epidemiologic studies. Applied Industrial Hygiene (In press).