

CHAPTER 21

INTERPRETIVE CONSIDERATIONS

This chapter reviews several scientific issues that should be considered when attempting to reach conclusions on a study of this size and complexity. These issues are critical to the interpretation of the data analyses in this report. Data patterns observed in many clinical chapters of this report are also summarized so that hypothesis testing of group differences may be placed in better perspective.

DIOXIN ENDPOINTS

Based upon data in this report, final conclusions on herbicide causality must consider results of the various clinical areas, reflected in the separate chapters. Each chapter introduction has attempted to highlight the major organ systems that are known or suspected to be significantly affected by the ingredients of Agent Orange with particular emphasis on the effects of dioxin. Categories of clinical endpoints and their generally accepted degree of association with dioxin are presented in Table 21-1. These associations are based on the scientific literature.

TABLE 21-1.

Summary Associations of Adverse Health Effects to
TCDD Exposure Reported in the Literature

Degree of Association by Clinical Chapter

Confirmed	Highly Suspected	Moderately Suspected	Negative or Weakly Suspected
Dermatology Neurology Hepatic	Malignancy	General Health Immunology	Psychology Cardiovascular Hematology Endocrine Renal Pulmonary

It is recognized that alternative conclusions based on these patterns of association are possible within the framework of current knowledge, particularly for the highly and moderately suspected areas (malignancy, general health, immunology). However, for illustrative purposes, two extremes are presented: multiple adverse findings in the Ranch Hand group for the areas

of dermatology, neurology, hepatic (discussed in Chapter 13), and cancer would suggest a case for TCDD causality, whereas multiple adverse findings in the weakly suspected areas, and not in any of the confirmed areas, would be difficult to ascribe to an overall TCDD causation.

The aspects of biological plausibility and specificity require balanced interpretation across clinical chapters, with careful attention placed on nonsignificant findings as well as significant findings. The chapters in this report should be viewed as artificial boundaries for convenience of presentation, and should not discourage consideration of their relatedness, or of the individual variables within them.

EXPOSURE

Approximately 600 exposure index analyses have been conducted in this study, underscoring attempts to associate increasing proportions of various abnormalities to estimates of increasing exposure.

To determine whether the results of the exposure analyses varied by chance, several perspectives were taken. Of the 255 adjusted exposure analyses (excluding 39 with interactions), 13 were statistically significant, a figure which is the expected number (based on $\alpha = 0.05$). It is recognized that this contrast is a crude yardstick, considering the relatedness of the dependent variables, statistical power, disproportionate representation of chapter variables, and the presence of interactions. The six possible patterns of exposure response (increasing, decreasing, V-shaped with fewer abnormalities at the low exposure level than the high exposure level, V-shaped with more abnormalities at the low exposure level than at the high exposure level, inverted V-shaped with fewer abnormalities at the low exposure level than the high exposure level, and inverted V-shaped with more abnormalities at the low exposure level than at the high exposure level) were tabulated (regardless of statistical significance) for the clinical chapters of dermatology, neurology, psychology, and renal. As noted in Table 21-1, two of these chapters contain clinical variables that have had confirmed associations to TCDD exposure, and two chapters have had negative or weakly suspected associations to TCDD. Of the 126 exposure analyses in these four chapters, 21 (or one-sixth) showed the primary pattern of interest, an increase--exactly the number expected. Taken together, these analyses suggest that statistically significant exposure analyses may have occurred due to chance among the data set, and that the pattern of dose-response may also have been random. These inferences, or that the exposure index was unrelated to actual exposure, together with the acknowledged limitations of the exposure index, indicate that estimated exposure may only be weakly relied upon to assert a causal relationship. Based upon the current exposure index calculations, either of the above inferential alternatives is possible.

The use of serum dioxin levels (see Chapter 23, Future Directions) in the next report will clarify the exposure calculations of this report and the Baseline Report. Thus, from an interpretive context, final conclusions on dose-response, and the implications to herbicide causation are based on current knowledge available for this report. These conclusions could change with future analyses using a factual exposure concept.

TYPES OF MEASUREMENTS

This report includes all types of measures traditionally used in morbidity followup epidemiologic studies, e.g., self-reports, structured interview responses, medical record data, physician findings, scalar measurements, biopsy results, laboratory determinations, morbidity indices, and mortality results. At many points in this report, various terms have been used to qualitatively describe the data and analyses arising from the measurement processes. In particular, the terms "subjective," "objective," "continuous," and "categorical," and "constructed indices" have been used to connote differences in data or data sets that are important in making statements of inference.

From the perspective of the Study Protocol, significant group differences for subjective historical variables, not mirrored by significant group differences in medical record findings or physician/laboratory testing, may be viewed as preliminary evidence of over-reporting by a group. The opposite finding of significant group differences for physical examination variables in the absence of reported symptoms may support the primary conclusion of significant subclinical group differences. Either of these alternatives may greatly affect an overall inference of herbicide causality. Hence, the descriptive phrases "subjective data" and "objective data" have not been used as value judgments of the worth of the data, but simply as inferential qualifiers.

This report contains numerous comments on the differences in results between analyses of continuous versus categorical data from the same variable (exclusively laboratory data). Because the statistical power is stronger for detecting mean shifts than categorical differences, it was anticipated that very small mean shifts might be more easily discerned than differences in proportions of abnormalities between the two groups. Both methods of examining the data reveal important aspects of the distribution. Inferentially, when both types of analyses were done, greater weight has been given to significant group differences when analyses of both data forms agree. Lesser weight was given to significant differences seen in only one analysis, and least weight to significant shifts in means if both group means were within normal range, and the mean difference was not supported by other statistical findings in related variables (e.g., hepatic test battery). Consistent patterns of findings within an organ system, or between related organ systems, is required to strongly suggest an inference of causality.

Several summary indices were constructed in this report, e.g., dermatology index, cranial nerve function index, and anatomic categories of abnormal peripheral pulses, and are similar to some indices in the 1984 Baseline Report. They were formed by summing or grouping related abnormalities for the purposes of assessing increased numbers and/or showing group directionality of overall results. They should not be strongly considered in final inferences because they are artificially derived.

BASELINE-FOLLOWUP EXAMINATION DIFFERENCES

A common difficulty of followup studies is the inherent variation in measurement systems from one observation period to the next. To the maximum extent possible, the USAF has restricted clinical variation by requiring the use of identical laboratory equipment for most clinical chemistries, by the

use of 50 samples from the Baseline serum bank to evaluate interexamination laboratory differences, and by the use of carefully prescribed written clinical procedures that allow little room for variation. Nonetheless, some interexamination variability must be expected, but in the presence of blindness to group membership, there is no reason to expect biases in the results with respect to either the Ranch Hand or Comparison groups.

This report has cited classical longitudinal analyses to assess changes in variables between the examinations by group. Of 21 variables examined, 5 showed statistically significant group differences in the changes between examinations. Four of these significant results were attributed to actual changes over time, while the other (e.g., sedimentation rate) was believed due to a change in laboratory methodology.

Other less refined longitudinal contrasts consisting of narrative discussions of Baseline results versus followup results have been presented in all chapters. Interpretive caution is required in assessing examination similarities or differences because of the slight changes in cohort composition between the examinations (see Chapter 2, Population), the use of slightly different statistical models and modeling strategy (see Chapter 7, Statistical Methods), and sometimes the use of the Original Comparison group. The relative contribution of these changes was not explored mathematically, but is believed to have played a minimal role in accounting for any large group shifts between examinations.

In the context of comparing results between examinations, there has been a subtle but consistent observation that group differences have substantially narrowed over the 3-year period, either by decreased findings in the Ranch Hands, increased findings in the Comparisons, or a combination of both mechanisms. In general, several broad interpretations are possible: any bona fide herbicide effect decreases over time, that the convergence is largely attributable to unquantifiable factors, that both examinations have produced chance results, or that these observations have been affected by the slight shifts in cohort composition and modeling strategy.

Several segments of this report have noted marked differences in the prevalence rates of abnormalities found at the Baseline and followup specialty examinations, e.g., the dermatology and neurology clinical assessments. The followup dermatological examination detected substantially more abnormalities than the Baseline examination, whereas far greater numbers of neurological abnormalities were noted at the Baseline examination than at the followup for some variables. These examination variances were affected by differences in "clinical sensitivity" between the examining teams, although clearly other factors (such as a true change in disease-abnormality status or slight cohort differences) contributed. The phrase "clinical sensitivity" refers to the inherent differences in clinical styles and interpretations of possible abnormalities that often prevail. Because of examiner blindness to exposure status, and because of the judgment that the interexamination variation was within the artful bounds of accepted medical practice, no bias was thought to have resulted from this inherent variation.

STUDY BIASES

Each reviewer of this report must reach a conclusion on whether the results of this study have been seriously flawed by the design, the operation

of significant biases, or both. The Protocol authors believe that the comprehensive multifaceted design is the chief strength of this study, although it is recognized that each and every published phase of the study must invite renewed inspection of fundamental scientific aspects of the study design.

It is believed that, with the exception of skin test readings, all data in this study were collected accurately and validly, and that blindness to group membership was well maintained throughout the collection process. This opinion is important from an inferential perspective in that both misclassification of data (tending to dilute true group differences) and bias in data (creating a false group effect) most likely did not occur appreciably in this study. Thus, it is believed that both the magnitude and direction of the group results found in this study reflect truth to the maximum degree possible, within the inherent boundaries of statistical models to account for all important adjusting variables.

GROUP INTERACTIONS: PATTERN RECOGNITION

Many of the adjusted analyses in this report have demonstrated significant group-by-covariate interactions, requiring stratified analyses to determine the nature of significant group differences. All significant two- and three-factor interactions have been included in the main text or in appendices. The analysis of followup data has found substantially more interactions than the analysis of Baseline data, due primarily to the larger number of covariates used in the followup analyses.

Several related viewpoints have aided in the overall interpretation of group-by-covariate interaction in the report. In the presence of a significant interaction, a direct conclusion on main group effects cannot be made, and the focal point of interpretation resides with the covariate stratum containing the significant group effect (or a reversal in nonsignificant group effects across strata). Past this point, however, there appears to be little consensus in how to best place the interaction into inferential context. Further interpretations appear to be largely individualistic.

No consistent pattern has emerged to support a finding of impairment in the Ranch Hands for any specific stratum of one or more covariates. In fact, of all the two- and three-factor interactions encountered, only one was thought to have possible biologic relevance. Other interactions may have such relevance, but the reason was not apparent. As with tests of group differences, significant interactions may occur by chance, but the method to calculate an expected number of group-by-covariate interactions, unfortunately, remains an open research question.

Because of the possible diverse interpretations of interactions, all significant two- and three-factor interactions involving group with statistically significant strata are presented in Table 21-2 for detailed inspection. No particular covariate or group pattern is noted, although the variables in psychology and gastrointestinal showed Ranch Hands at a relative detriment, while the interactions in the cardiovascular chapter indicated detrimental findings in the Comparisons.

Most variables without interactions in this report have shown remarkable concordance between unadjusted and adjusted results, both in terms of absolute value of relative risk and of statistical significance.

TABLE 21-2.

Summary of Significant Covariate Strata (or Covariate Level Difference)
 Found Within Significant Two- and Three-Factor Group-by-Covariate Interactions
 by Clinical Chapter and Dependent Variable
 (Group Direction and p-Value)

Clinical Chapter	Dependent Variable	Covariate Stratum	RH>C	C>RH	p-Value
General Health	Self-Perception of Health	Enlisted Groundcrew	*		0.003
Malignancy	Basal Cell Carcinoma (Verified Interval)	Enlisted Flyer	*		0.019
		Systemic Cancer (Verified plus Suspected, Interval)	Enlisted Flyer	*	
	Basal Cell Carcinoma (Verified plus Suspected, Lifetime)	Intermediate Skin Reaction to Sun	*		0.038
	Systemic Cancers (Verified, Lifetime)	Enlisted Flyer	*		0.019
	Systemic Cancer (Verified plus Suspected, Lifetime)	Enlisted Flyer	*		0.004
	Neurology	Pin Prick	Impaired (Diabetic Class)		*
Psychology	Paranoia	Born Before 1942	*		0.027
		Schizophrenia	High School	*	
	Social Introversion	Combat Index--Low	*		0.002
	Validity	Black		*	0.038
	Total CMI	High School	*		<0.001
Gastrointestinal	SGOT	1-4 Drinks per Day	*		0.010
	Alkaline Phosphatase	Exposed to Ind. Chems.	*		<0.001
	Direct Bilirubin	Exposed to Ind. Chems.	*		0.035
	Triglycerides (cont.)	Born In or Before 1922	*		0.039
	Triglycerides (disc.)	Officer	*		0.035
	Uroporphyrins	BUN<14		*	<0.001

TABLE 21-2. (continued)

Summary of Significant Covariate Strata (or Covariate Level Difference)
 Found Within Significant Two- and Three-Factor Group-by-Covariate Interactions
 by Clinical Chapter and Dependent Variable
 (Group Direction and p-Value)

Clinical Chapter	Dependent Variable	Covariate Stratum	RH>C*	C>RH	p-Value
Dermatology	Dermatology Index	Pre-SEA Acne: 1 vs. 0		*	0.004
Cardiovascular	Systolic Blood Pressure	Black/53 Yrs Old		*	0.006
	ECG (Overall)	0 Pack-years		*	0.038
	ECG (Arrhythmia)	7 Pack-years/10% Body Fat		*	0.018
	Posterior Pulses (Manual)	Enlisted Flyer	*	*	0.032
	Leg Pulses (Manual)	Officer/21% Body Fat		*	0.026
	Peripheral Pulses (Manual)	Officer		*	0.030
Hematology	WBC	Nonblack/30 Pack-years/ 35 Yrs Old	*		<0.001
	WBC	Black/Officer/35 Yrs Old		*	0.003
	WBC	Black/EFL/35 Yrs Old		*	0.050
	PLT	Nonblack/30 Pack-years and 1 pack/day	*		0.014
	PLT	Black/30 Pack-years and 1 pack/day	*		0.007
Renal	Urinary Protein	Normal (Diabetic Class)	*		0.018
	Urinary WBC	Nonblack/Born In or After 1942	*		0.001
	BUN	Black		*	0.017
	Urine Specific Gravity	Nonblack/Enlisted Groundcrew	*		<0.001
Endocrinology	Testosterone	<10% Body Fat	*	*	0.012
	Testosterone	10-25% Body Fat		*	0.023
	Differential Cortisol	Black/Born In or After 1942		*	0.003

TABLE 21-2. (continued)

Summary of Significant Covariate Strata (or Covariate Level Difference)
 Found Within Significant Two- and Three-Factor Group-by-Covariate Interactions
 by Clinical Chapter and Dependent Variable
 (Group Direction and p-Value)

Clinical Chapter	Dependent Variable	Covariate Stratum	RH>C*	C>RH	p-Value
Immunology	Total T Cells	Black		*	0.039
	B Cells	Nonblack/0 Pack-years		*	0.004
	Monocytes	Enlisted Groundcrew/ 4 Drinks/Day	*		0.003
Pulmonary	Pleurisy Tuberculosis X-ray	1-10 Pack-years	*		<0.001
		1-10 Pack-years	*		0.020
		0 Pack-years		*	0.030
	Total Interactions: 43		26	17	

*Relative risk greater than one, or Ranch Hand mean greater than Comparison mean.

CLASSICAL COVARIATES

Many of the dependent variables in this report are known to be significantly affected by risk factors also measured in this study. The use of these covariates in the adjusted analyses has served to clarify Ranch Hand-Comparison group differences in the presence of significant covariate group differences. Such adjustments, whether by a single covariate, multiple covariates, or covariate interactions, have given results on group differences generally quite similar to the unadjusted analyses both in terms of relative risk and statistical significance. In fact, in only one instance in this report has an unadjusted result of $p \geq 0.10$ changed to a value of $p \leq 0.05$ in the adjusted analysis. The covariates used in this study were not effect modifiers (which may be synergistic with exposure and also be equally distributed between groups). Consistent effects were observed for almost all of the classical covariates of age, race, occupation, education, alcohol, smoking, percent body fat, and glucose tolerance. In only a few instances were unexpected effects noted, e.g., personality type, wine consumption, and a few smoking and alcohol "inversions."

The overall covariate effects observed in this study indeed reflect the mainstream of results found in well-conducted epidemiologic studies, and lend credence to the validity of the clinical endpoints and covariate values in this report.

MULTIPLE COMPARISONS

As noted in Chapter 7, Statistical Methods, the problem of multiple comparisons is complex and not easily adjudicated because of the total number of statistical tests, the number of tests performed on each dependent variable, and the biologic relatedness of many of the variables. A conscious effort has been made to expand inferential interest to borderline group associations ($0.05 < p \leq 0.10$) thereby increasing the probability of the acceptance of a false association. Each chapter summary has carefully flagged all borderline associations to provide expanded summary statements for possible inclusion in deriving final conclusions. Additional confidence in the final acceptance or rejection of an overall herbicide effect would be warranted if the majority of borderline associations were in the same consistent direction as the significant associations.

Multiple analyses on the same variable have been conducted in this report. Continuous and categorical data have been subjected to both unadjusted and adjusted analyses, and multiple adjusted analyses were sometimes conducted with different covariates or slightly different covariate sets. The question arises as to which results best reflect the truth when different results are found. In general, the following approach has been followed: the statistical significance of both continuous and categorical analyses is convincing, while significance for only the continuous analysis must be viewed in terms of the biologic relevance of the mean shift detected.

Overall, the multiple comparison issue is due to repeated hypothesis testing for group, exposure, and interaction strata differences. The calculation of expected numbers of significant associations for these tests is difficult (if not impossible) because of the relatedness of the dependent variables, the relatedness of the covariates, and the often difficult analytic decisions that arise in a "step-down, best model" strategy. Thus,

the final assessment of whether the frequency of significant associations does not meet, or exceeds expectation, must remain an interpretive judgment of each reader.

CAUSALITY

The AFHS is an inferential assessment of observed group differences. The inference of herbicide causality will be determined by a balanced judgment of the following factors: biological plausibility, consistency, specificity, coherence, time relationships, and strength of association. Except for aspects of association strength, most of these causality factors have been discussed in the preceding sections of this chapter. Nearly every statistically significant group difference in this report has only been of moderate to weak strength. Highly significant p-values ($p < 0.001$) were not found for main group associations, but were observed for covariate tests. A few strata in the group interactions were highly significant. Most of the statistically significant estimated relative risks were below the value of 2.0 (a traditional boundary of interest in epidemiology). The few relative risks above 2.0 generally had very wide confidence intervals due to low proportions of detected abnormalities. Weakly significant associations, in particular, are cause to reassess the element of chance and the possible presence of other causality factors before a final conclusion of cause and effect is determined.