A STATISTICAL ANALYSIS OF WORK HISTORY DATA IN A RETROSPECTIVE OCCUPATIONAL MORTALITY STUDY OF THE ASSOCIATION OF SOLVENTS WITH LEUKEMIA IN RUBBER WORKERS

by

Robert Spirtas

Department of Biostatistics
University of North Carolina at Chapel Hill

Institute of Statistics Mimeo Series No. 1049

January 1976

SPIRTAS, ROBERT. A Statistical Analysis of Work History Data in a Retrospective Occupational Mortality Study of the Association of Solvents with Leukemia in Rubber Workers. (Under the direction of LAWRENCE L. KUPPER.)

Does exposure to solvents used in the rubber tire industry lead to an increase in the risk of developing leukemia? This dissertation deals with data from work histories as a measure of time spent in solvent-exposed jobs. Several statistical analyses are carried out using two separate epidemiologic study designs, a case-control design involving sixty male cases of leukemia and 180 controls who have been individually matched on age-at-death, race and sex (three per case) and a hybrid design using the same set of cases with an age-stratified sample of the male Population at Risk (PAR) as the comparison group.

A computer program, called the Experience Transformation Algorithm, was developed which accumulates total time in prespecified groups of jobs. The output of this program is a multivariate continuous measure of exposure. Multivariate Linear Models were utilized to measure the differences in solvent exposure for cases and controls after adjustment for the effects of variables such as race, age-at-death and education. A new relative risk estimation procedure was employed which made use of the work experience information from the sample PAR.

There was little difference in solvent exposure for all leukemia cases and controls. This result held true for both study designs. However, when the study was restricted to cases of lymphatic leukemia (which had previously shown an elevated Standardized Mortality Ratio),

the results showed that cases tend to have spent more time, on the average, in solvent-exposed jobs than did controls. These results were not statistically significant for several of the analyses used, suggesting that:

(1) Solvent exposure may be more widely dispersed throughout the plant than originally estimated.

(2) The etiologic agent may not be a solvent.

Although the study findings are not totally consistent with the hypothesis, they suggest that care be used in handling solvents. The general application of the statistical procedures used is to the situation where a complex set of occupational environments can be categorized into mutually exclusive groups with or without knowledge of quantitative levels of environmental exposure to specific agents. Further research is underway to expand both the epidemiologic data base and the industrial hygiene measures of past and present solvent exposures.

A STATISTICAL ANALYSIS OF WORK HISTORY DATA IN A RETROSPECTIVE OCCUPATIONAL MORTALITY STUDY OF THE ASSOCIATION OF SOLVENTS WITH LEUKEMIA IN RUBBER WORKERS

by

Robert Spirtas

A Dissertation submitted to the faculty of the University of North Carolina in partial fulfillment of the requirements for the degree of Doctor of Public Health in the Department of Biostatistics, School of Public Health.

Chapel Hill

1975

Approved by:

_____
Adviser

_____
Reader

_____
Reader

DEDICATION

This dissertation is dedicated to the memory of my mother and father, Elizabeth and Abe Spirtas.

ACKNOWLEDGMENTS

The author is indebted to the members of his dissertation committee: Drs. Elizabeth J. Coulter, Robert L. Harris, Jr., Lawrence L. Kupper (Chairman), Anthony J. McMichael and Michael J. Symons. They all gave an extra measure of time and effort as well as many words of encouragement. Special thanks are due to Dr. Bernard G. Greenberg for enabling the author to participate in this study.

The author is most grateful to his wife, Joan, and his sons, Michael and John, who have said good-bye to their daddy on so many nights and week-ends.

Mrs. Anna Colosi wrote the computer program which transforms the work history. In addition, Mrs. Kay Fendt and several others in the OHSG computing staff helped the author in his encounters with the computer.

Mrs. Gay Hinnant typed the manuscript and made numerous helpful suggestions.

Support for the author's participation in this study was given by the Occupational Health Studies Group, University of North Carolina, under the direction of Drs. David A. Fraser and Robert L. Harris, Jr.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INTRODUCTION AND BACKGROUND

Summary

Work history information is used to discern associations between an excess of deaths due to leukemia and exposure to solvents among a cohort of rubber workers. Two separate study designs are considered: a case-control design involving three individually matched controls per case and a hybrid design using a stratified random sample of the cohort as the control group.

Statistical techniques are used which allow for the investigation of many hypotheses about the suspected associations. The utilization of complete work history information enables the exposure variable to be treated as multivariate and continuous, making for more sensitive and more powerful statistical analyses. The general application of these statistical procedures is to the situation where a complex set of occupational environments can be categorized into mutually exclusive groups with or without knowledge of quantitative levels of environmental exposure to specific agents.

Hypothesis

The hypothesis to be tested is that the risk of developing leukemia, among rubber workers, is significantly greater among those working in jobs entailing exposure to solvents, compared to those working

in other jobs. Further, the magnitude of this risk is hypothesized to increase with both increased exposure duration and increased exposure level.

Research Goal

The primary goal of this study is to analyze statistically the association between work experience and occurrence of leukemia using two retrospective study designs. The data base consists of records from ongoing studies of one rubber manufacturing company by the Occupational Health Studies Group, School of Public Health, University of North Carolina. To achieve this goal, the following two secondary goals will be pursued:

1. Develop a computer program capable of transforming complete work history data into a form amenable to statistical analysis.

2. Utilize multivariable statistical methods to elicit the relationship describing the inherent differences between cases and controls in terms of transformed work history data and other concomitant variables such as education. Study the importance of certain of these confounding variables in determining work-experience patterns.

Background of the Occupational Health Studies Group

Increasing emphasis has recently been given to the role of chemical exposure as an etiologic factor in the development of certain malignancies. One of the most difficult aspects of studies in this area has been the inability to retrospectively measure the environmental insult to the individuals under study.

For at least three reasons industrial populations serve as "natural experiments" for testing hypotheses about suspected chemical carcinogens:

1. Companies as well as unions often have good records on the "population at risk" allowing for relatively easy identification of historic cohorts. Often, these populations tend to be relatively stable with regard to in and out migration.

2. Companies sometimes have (at least qualitatively) retrospective information on chemical usage. In addition, they often have quantitative information on present usage.

3. Industrial populations are often exposed to much higher levels of toxic or potentially toxic chemicals than the general population, e.g., vinyl chloride. Occasionally, there is industrial hygiene analytical data documenting such exposure levels.

Occupational mortality studies of workers exposed to a single presumed carcinogen have begun to fill some of the gaps in knowledge by examining the work histories of workers exposed to these agents. The more difficult task is to investigate industries where workers are exposed to heterogeneous mixtures of chemicals, some of which have undergone chemical transformation, where little is known of toxicologic properties, and nothing of synergism or additive effects.

The Occupational Health Studies Group (OHSG) within the School of Public Health at the University of North Carolina, is currently engaged in a research program contracted between the university, management and the labor unions within the U. S. rubber industry. The OHSG is a multidisciplinary organization, drawing primarily upon the Departments of Biostatistics, Environmental Sciences and Engineering and Epidemiology.

It currently has a full-time professional research staff of 15 persons and an equal number of administrative, secretarial, data processing and analytical laboratory staff. The OHSG has an intimate knowledge of the data sources available from each of the four major rubber companies with which it is working, and has already transferred much of the relevant data into its own computer files. Three years of research work have already been carried out under this rubber industry contract, and the initial epidemiologic studies have been completed.

Results of Initial Cohort Study

In the initial mortality study done by the Occupational Health Studies Group, a cohort in a certain factory was reconstructed to include 8309 hourly rubber workers defined as active or pensioned as of January 1, 1964 (McMichael, Spirtas and Kupper, 1974). The breakdown by age, race and sex for this population is given in Table 1.1. Due to the small number of females in this population (767), the calculation of Standardized Mortality Ratios (SMR's) was limited to the male segment of the population.

The numerators for the SMR's consisted of the actual number of deaths in the study cohort for which the given cause was the underlying cause of death as determined by a nosologist from the National Center for Health Statistics. The denominator, consisting of the expected number of deaths, was computed using U. S. 1968 death rates for 5 year age intervals (U. S. Department of Health, Education, and Welfare, 1972a). The rates from the standard U. S. population were applied annually to the study population allowing for attrition due to death or loss to follow-up and also considering the aging of the study population.

TABLE 1.1

AGE-RACE-SEX DISTRIBUTION OF RUBBER WORKERS COHORT AS OF JANUARY 1, 1964

	Male				Female				
Age	White	Black	Race Unknown	Total Male	White	Black	Race Unknown	Total Female	Sex and Race Unknown
≤ 39	684	154	3	841	4	0	0	4	0
40-44	678	150	7	835	29	0	1	30	0
45-49	730	149	3	882	61	0	2	63	0
50-54	655	156	7	818	124	0	1	125	0
55-59	805	167	12	984	172	2	5	179	2
60-64	887	135	16	1038	132	0	7	139	2
65-69	864	96	14	974	109	0	6	115	1
70-74	627	44	8	679	79	0	2	81	0
75-79	334	16	9	359	20	1	1	22	0
80-84	94	5	2	101	7	0	0	7	0
≥ 85	26	0	0	26	2	0	0	2	0
Totals	6384	1072	81	7537	739	3	25	767	5

Grand Total = 8309

Source: OHSG Files

Table 1.2 presents SMR's for ten years of follow-up on the portion of this population with age forty to eighty-four at January 1, 1964. As can be seen there are more than the expected number of deaths for neoplasms of the lymphatic and hematopoietic system as well as stomach and prostate cancers. In addition there is an excess of deaths due to arteriosclerosis. Because of the excess of leukemia deaths coupled with knowledge of solvent use in the rubber industry, there was a natural tendency to follow up this initial finding with a more detailed analytic study (McMichael et al., 1975). The resulting case-control study of leukemia and solvent exposure, utilizing work histories as the measure of exposure, forms the basis for the research reported here.

Classification of Leukemia

Dorland's Illustrated Medical Dictionary (1965) defines leukemia as: ". . . a fatal disease of the blood-forming organs, characterized by a marked increase in the number of leukocytes and their precursors in the blood, together with enlargement and proliferation of the lymphoid tissue of the spleen, lymphatic glands and bone marrow." Leukemia has generally been considered to be a type of neoplasm. The Eighth Revision International Classification of Diseases Adapted for Use in the United States (u. S. Department of Health, Education, and Welfare, 1969), further subdivided leukemia into four cytologic groups; lymphatic, myeloid, monocytic and other and unspecified. Within each of these groups there is, essentially, a further breakdown by clinical course and characteristics into acute, chronic or unspecified. The only exception to this classification scheme is the addition of acute erythremia (ICDA code 207.2) as a subclassification under other and unspecified leukemias.

TABLE 1.2

AGE-RACE STANDARDIZED MORTALITY RATIOS (SMR) FOR COHORT OF RUBBER WORKERS--
SELECTED CAUSES (1964-73)--MALES

Age		40-64			65-84			40-84		
Cause	(ICDA Code)	Deaths Observed	Deaths Expected	SMR	Deaths Observed	Deaths Expected	SMR	Deaths Observed	Deaths Expected	SMR
All Causes		519	588.6	88	1463	1393.9	105	1982	1982.5	100
All Neoplasms	(140-239)	113	120.6	95	288	250.9	115	401	371.5	108
Malignant Neoplasms	(140-209)	110	119.0	92	281	248.0	113	391	366.9	107
Ca Stomach	(151)	10	6.5	155	30	17.0	177	40	23.4	171
Ca Large Intestine	(153)	9	8.8	103	33	25.6	129	42	34.3	122
Ca Rectum	(154)	3	3.6	84	6	9.1	66	9	12.7	71
Ca Pancreas	(157)	4	7.0	57	14	14.9	94	18	21.9	82
Ca Respiratory System	(160-163)	35	45.2	78	71	70.5	101	106	115.7	92
Ca Prostate	(185)	5	5.1	99	48	32.9	146	53	37.9	140
Ca Bladder	(188)	2	2.8	72	8	10.6	76	10	13.4	75
Ca CNS	(191-192)	1	3.6	28	3	2.6	114	4	6.2	65

TABLE 1.2 (Continued)

Age		40-64			65-84			40-84		
Cause	(ICDA Code)	Deaths Observed	Deaths Expected	SMR	Deaths Observed	Deaths Expected	SMR	Deaths Observed	Deaths Expected	SMR
Ca Lymphopoietic System	(200-209)	16	10.4	155	27	21.2	127	43	31.6	136
Lymphosarcoma, Hodgkin's	(200-201)	5	3.7	137	10	5.5	182	15	9.1	164
Leukemia	(204-207)	11	3.8	292	6	9.7	62	17	13.5	126
Diabetes	(250)	16	10.5	152	38	26.8	142	54	37.3	145
Ischemic Heart Disease	(410-413)	235	253.2	93	635	658.2	96	870	911.4	95
Cerebro-Vascular Disease	(440)	4	2.2	179	35	24.1	145	39	26.4	148
Chronic Respiratory Disease	(490-493)	16	16.6	96	45	51.7	87	61	68.3	89
Liver Cirrhosis	(571)	16	21.9	73	22	14.0	157	38	35.9	106
Suicide	(E950-959)	12	11.0	110	10	8.9	113	22	19.8	111
Other Accidents, Poison's, Violence	(E800-999 - E950-959)	22	44.7	50	29	40.9	71	51	85.6	60
All Other Causes		85	107.9	79	361	318.4	113	446	426.3	105

Source: OHSG Files

In Table 1.3 this cytological classification scheme is used to further break down the leukemia SMR's for the cohort under study. As Table 1.3 shows, the greatest excess is exhibited by lymphatic leukemia, especially in the forty to sixty-four year old age range (SMR = 7.93). This results, for which the chi-square test of association, $\frac{(\text{observed-expected})^2}{\text{expected}}$, is statistically significant ($p < .0001$), will be pursued in the subsequent analytical study.

According to Doll (1965) it is much more difficult to classify acute leukemias as to cell type than it is to so classify chronic leukemias. Since most of the cases in the current study were of the chronic type this should not pose much of a problem.

Epidemiology of Leukemia

In this study of the relationship between leukemia deaths among rubber workers in a certain factory and exposure to solvents, it is important to discuss the epidemiology of leukemia. Medical descriptions of the pathology, diagnosis and treatment of leukemia are given by Moore (1971) and Dameshek and Gunz (1964) among others.

Leukemia is a rare disease with a current world-wide crude death rate varying between two and eight per 100,000 population (Segi and Kurihara, 1966). With the exception of childhood leukemia (which is not of interest in this study of a cohort of workers), the age-specific mortality rates for leukemia show a generally increasing trend with age for U. S. whites. The increase in rates is slow until approximately age fifty when the rate begins to rise rapidly up to the upper age limit. For U. S. non-whites there appears to be an increase to a peak incidence at ages seventy to seventy-five followed by a general decline in

Table 1.3

SMR'S FOR CYTOLOGIC CATEGORIES OF LEUKEMIA FOR TOTAL MALE DEATHS 1964-1973

ICDA Category		Age at January 1, 1964					
		40-64		65-84		40-84	
		SMR	(Observed Deaths)	SMR	(Observed Deaths)	SMR	(Observed Deaths)
204	Lymphatic Leukemia	7.08	(8)	.77	(3)	2.19	(11)
205	Myeloid Leukemia	1.93	(3)	.82	(3)	1.15	(6)
206	Monocytic Leukemia	0.00	(0)	0.00	(0)	0.00	(0)
207	Other and Unspecified Leukemias	0.00	(0)	.97	(2)	.72	(2)

Source: OHSG Files

leukemia mortality. For both whites and non-whites the rates for those over age fifty is consistently higher for males than females (Burbank, 1971).

In general, the sex ratio (ratio of total male deaths to total female deaths) for leukemia is greater than one for most countries of the world. The relatively high sex ratio (about 1.9) for deaths due to chronic lymphatic leukemia has led to the suggestion that environmental factors may play a role in the etiology of this type of leukemia (Kessler and Lilienfeld, 1969). The increase in the sex ratio over time for adult leukemia also suggests an environmental factor (Fraumeni and Wagoner, 1964). However, the present evidence linking leukemia with environmental factors is meager.

The changes in leukemia mortality rates over time have been discussed by several authors (Fraumeni and Miller, 1967; Kessler and Lilienfeld, 1969; Silverberg and Holleb, 1973). The generally upward trend for adult leukemia rates has continued over most of this century with a levelling off occurring in the 1960's. However, among the older age groups the trend continues upward. Part of this change may be explained by the change in definition of leukemia over time. Problems of classification probably resulted in an under-reporting of leukemia deaths in the early 1900's (Sacks and Seeman, 1947). The improvement in diagnostic accuracy has also resulted in attributing more deaths to leukemia. In addition the decrease in infectious disease mortality rates has removed or reduced certain competing risks, resulting in more deaths due to neoplasms as people's life expectancy has increased. However, even after considering all of these concomitant factors it is still obvious that there has been a great increase in the incidence of

leukemia over time. The principle increase has been in the acute leukemias. Court Brown and Doll (1959) postulated that an explanation for this increase over time may be the increased exposure to ionizing radiation.

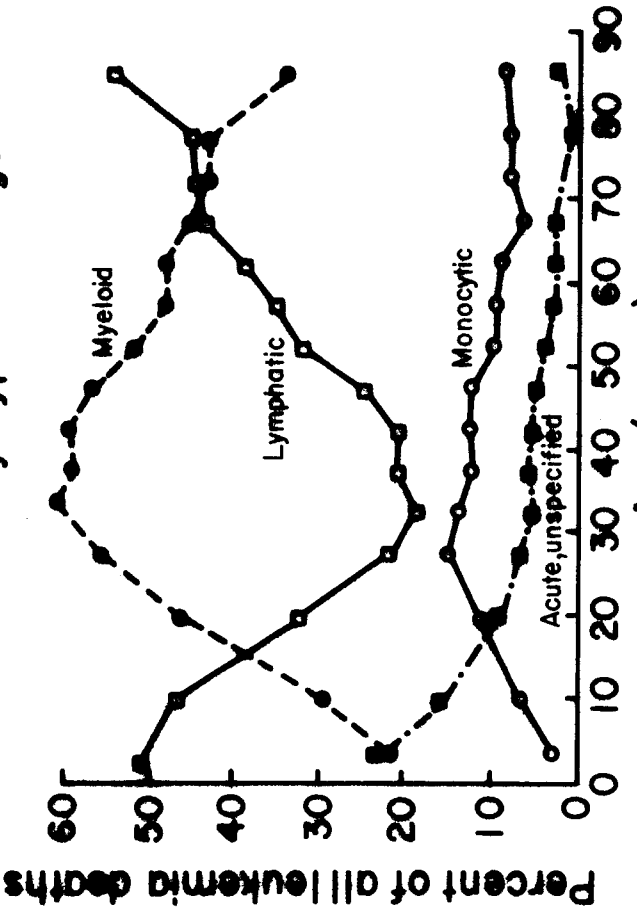
For the four cytologic types of leukemia deaths discussed in Table 1.3 the relative frequency in the U. S. is shown in Figure 1, which shows that the myeloid form is predominant in adults up to age seventy when the lymphatic form takes over. Doll (1965) surmised that the rapid increase in chronic lymphatic leukemia for older people is suggestive of a prolonged exposure to a weak carcinogen. The pattern exhibited in Figure 1 further emphasizes the findings in Table 1.3, where there were nine lymphatic leukemia deaths compared with three myeloid leukemia deaths in the age range forty to sixty-four.

There appears to be a difference in leukemia mortality rates due to race. Clarkson and Burchenal (1965) have shown that a racial difference exists for chronic lymphocytic leukemia among Japanese and certain other non-Caucasian groups, among whom it is extremely rare. However, McMahon and Koller (1957) were able to explain the difference in U. S. leukemia mortality rates among whites and blacks by the difference in social level.

The association of chromosomal aberrations with leukemia incidence has been cited often (Forni et al., 1971; Clarkson and Burchenal, 1965; Miller, 1966). The existence of a specific abnormality in the twenty-first chromosome, known as the Philadelphia or Ph^1 chromosome, is quite common in cases of chronic myelogenous leukemia, ranging up to ninety per cent by some estimates (Clarkson and Burchenal, 1965). A more

Figure 1

Relative Frequency of Leukemia Deaths by Type and Age



Source: Hewitt, 1955.

unusual defect of this same chromosome, which has been associated with chronic lymphatic leukemia, is known as the Christ Church (CH^1) chromosome (Gunz, Fitzgerald and Adams, 1962). However, since not all cases of leukemia exhibit chromosomal aberrations there must be at least one additional mechanism capable of causing the disease without affecting the structure of the chromosome. The often cited excess of leukemia among genetically damaged individuals, e.g., Down's Syndrome (mongolism), supports the hypothesis that genetic damage predisposes an individual to developing leukemia (Miller, 1966). Further support of the genetic link may be derived from the finding of an excess of leukemia deaths among Jews, who tend to intermarry less than other religious and ethnic groups (McMahon and Koller, 1957).

Blood group may play some role in susceptibility to leukemia. One hypothesis is that Type O blood may predispose toward serious blood changes (Rejsek and Rejsková, 1955).

Regional differences in leukemia rates exist in the United States with excesses in New York, Minnesota and California (Walter and Gilliam, 1956). However, the rate for Ohio, the state in which the study population lived and worked, was not significantly different from the rate for the entire United States.

There appears to be no clear cut urban-rural effect. Lilienfeld, Levin and Kessler (1972) have found a lower metropolitan rate for lymphatic leukemia and other and unspecified leukemia when compared with the nonmetropolitan rate. McMahon and Koller (1957) found only a slight association between urbanization and leukemia mortality. Hewitt (1960) found an association between leukemia and urbanization in England and Wales only among those over age sixty-five.

It is difficult to ascertain whether or not there is a seasonal nature for leukemia because of the insidious nature of its onset. In general, however, there has not been much evidence to support the hypothesis of seasonality (Kessler and Lilienfeld, 1969).

Education has been found to have an inverse relationship, in general, with mortality (Kitagawa and Hauser, 1973). However, there is little evidence to support this hypothesis for leukemia deaths. It is still important to determine whether education may be indirectly related to the incidence of leukemia by influencing job patterns which have varying environmental hazards.

Socio-economic status may be associated with incidence of leukemia. Sacks and Seeman (1947) found an association with economic level. Lilienfeld, Levin and Kessler (1972) found a higher mortality rate for U. S. native whites than for foreign born whites.

Smoking has been conclusively found to cause lung cancer and to be less strongly associated with laryngeal, esophageal and bladder cancer (U. S. Department of Health, Education, and Welfare, 1964). Kahn (1966) reported an SMR of 1.41 for leukemia in current smokers but, in general, there is little to suggest a direct association between smoking and leukemia.

The incubation or latent period for leukemia ranges from one to ten years (Cobb, Miller and Wald, 1959; Cronkite, 1961; Lilienfeld, Pederson and Dowd, 1967; Polednak, 1974). The modal average appears to be approximately five years. There is some speculation that there may be an inverse relationship between the strength of leukemogenic agents and the latent period, but this point has not been well established (Cobb, Miller and Wald, 1959).

Varying between zero to ten per cent, the ten year survival rate for the different forms of leukemia insures that a study of deaths due to leukemia will include most of the cases which have occurred in the cohort under study (U. S. Department of Health, Education, and Welfare, 1972b).

Even though the focus of this study is on leukemia deaths it is worthwhile to indicate work that has been done on leukemia morbidity. Most of the work in this area has been done by National Cancer Institute personnel (Dorn and Cutler, 1959; Haenszel, Marcus and Zimmerer, 1956; Cutler, 1973). Levin et al. (1974) state that morbidity and mortality patterns are similar by age and sex. Other information has come from Tumor Registries such as the one in Connecticut (Connecticut State Department of Health, 1966).

Etiologic Agents

The three classes of agents most often implicated in the etiology of leukemia are radiation, chemicals and viruses. It is generally accepted that ionizing radiation can cause leukemia. The association between radiation and myeloid leukemia has been cited by several authors (Fraumeni and Miller, 1967; Kessler and Lilienfeld, 1969; Bizzozero, Johnson and Ciocco, 1966; Court Brown and Doll, 1957). The accepted fact that ionizing radiation can cause chromosomal damage is thought to be one possible etiologic link in the causation of leukemia (Fraumeni and Miller, 1967; Kessler and Lilienfeld, 1969; Forni et al., 1971; Miller, 1964).

The case for chemical leukemogens, although not as well accepted as the radiation theory, is still quite strong. The chemical agent

most often associated with leukemia is benzene, which has the chemical designation C_6H_6 (Cronkite, 1961; Pagnotto et al., 1961; Vigliani and Saita, 1964). The familiar hexagonal benzene ring was first determined in 1866 (Hunter, 1969). During the late nineteenth century it was reportedly tested for use as an anesthetic (Hamilton, 1931). Its toxic property has been known since 1897 (Hunter, 1969). Benzene was known at that time to have a leukotoxic effect, but it was not understood to have a role in the causation of leukemia. In fact, it was used at one time as a medication to reduce the white cell count in leukemic patients (Hamilton, 1931). "Benzene Poisoning" was a well-known phenomenon in the beginning years of this century, but the link with leukemia did not come until 1928 (Vigliani and Saita, 1964). In the United States in 1939 the publication of papers by Mallory, Gall and Brickley (1939) and Erf and Rhoads (1939) were milestones in establishing this relationship. Later studies by Tough and Court Brown (1965) and Forni et al. (1971) have indicated that benzene may be able to cause chromosomal damage. Aksoy et al. (1974) have been able to associate benzene exposure with Hodgkin's disease, which is cytologically similar to leukemia. Most of the cited cases of benzene leukemia have had the myelogenous form of the disease, although instances of lymphatic leukemia have been reported (Falconer, 1933; Tareef, Kontchalovskaya and Zorina, 1963; Rejsek and Rejsková, 1955; Kieć and Kuński, 1965).

In most of the previously cited studies the suspected agent was the commercial grade of benzene, often called benzol, which contains impurities such as toluene, xylene, carbon disulphide and phenol (Hunter, 1969). It is conceivable that one or more of these contaminants has caused or contributed to the leukemogenic etiology. However,

toxicologic studies with relatively pure benzene strengthen the suspicion that benzene, itself, is the etiologic agent (Schrenk et al., 1940).

The recent publication of a benzene criteria document (U. S. Department of Health, Education, and Welfare, 1974) recommends that the concentration of benzene on a time weighted average be kept below ten parts per million parts of air (ppm) with a ceiling, never to be exceeded of twenty-five ppm. In the previously cited studies which included industrial hygiene measurements for benzene, the levels were usually much higher than these recommended limits.

In addition to benzene, other chemicals suspected of being leukemogenic include: chloramphenicol (Brauer and Dameshek, 1967; Fraumeni and Miller, 1967), phenylbutazone (Bean, 1960; Jensen and Roll, 1965), 7-12 dimethylbenz [a] anthracene (Sugiyama, Kurita and Nishizuka, 1967) and arsenic (Kjeldsberg and Ward, 1972). A list of drugs and chemicals associated with blood dyscrasias is given in Table 1.4.

The hypothesis that viruses can cause leukemia is not strongly supported by the literature, although many scientists still support this theory. Proponents of the viral theory point to time-space clusters of leukemia cases and to the high occurrence of multiple cases in some families to support their hypothesis (Fraumeni and Miller, 1967; Kessler and Lilienfeld, 1969). The highest familial incidence has been found for lymphatic leukemia (Clarkson and Burchenal, 1965). The counter argument to this theory is usually the statistical law that chance alone will dictate a certain number of multiple cases (Mantel, 1967).

Certain occupations have been cited in case-history studies for their associations with cases of leukemia, e.g., anesthesiologists

TABLE 1.4

DRUGS OR CHEMICALS SHOWN BY DIRECT OR CIRCUMSTANTIAL EVIDENCE TO BE ASSOCIATED WITH BLOOD DYSCRASIAS

*Acetanilid	Gold Salts	Quinacrine
Acetazolamide	Imipramine	Quinidine
*Acetophenetidin	Lead	Quinine
Allylisopropylacetylurea	Mepazine	Ristocetin
Aminopyrine	Meprobamate	Stibophen
*Aminosalicylic Acid	Methimazole	Streptomycin
Arsphenamine	Methylphenylethyl Hydantoin	*Sulfacetamide
Benzene	*Naphthalene	Sulfadiazine
Carbutamide	*Nitrofurantoin	*Sulfamethoxypyridazine
Chloramphenicol	*Pamaquin	*Sulfanilamide
Chlordane	Phenindione	Sulfisoxazole
Chlorothiazide	Phenylbutazone	*Sulfoxone
Chlorpromazine	*Phenylhydrazine	*Thiazolsulfone
Chlorpropamide	*Primaquine	Thiobarbital
Colchicine	Primidone	Thiouracils
Diphenylhydantoin Sodium	*Probenicid	Tolbutamide
Dipyrone	Promazine	Trimethadione
Gamma Benzene Hexachloride	Pyrimethamine	Trinitrotoluene

*These drugs have been associated with induction of hemolytic anemia principally in patients with glucose-6-phosphate dehydrogenase-deficient cells.

Source: Erslev and Wintrobe, 1962

(Bruce et al., 1968), canners (Falconer, 1933), chemical workers (Tareef, Kontchalovskaya and Zorina, 1963), painters (Viadana and Bross, 1972), pesticide workers (Kjeldsberg, 1972), printers (Greenburg et al., 1939), rubber workers (Forni and Luciano, 1967) and shoe factory workers (Aksoy et al., 1971). Epidemiologic studies have shown the same results for farmers (Milham, 1971), radiologists (March, 1950; Matanoski et al., 1975), and rubber workers (McMichael, Spirtas and Kupper, 1974; McMichael et al., 1975; Monson and Nakano, 1974; Fox, Lindars and Owen, 1974). For all of these studies, excepting the one involving radiologists, the suspected etiologic agent has been a chemical.

Although the etiology of leukemogenesis is presently unknown, it is felt likely to require a complex set of events involving human susceptibility, possibly of a genetic or biologic nature, as well as an oncogenic stimulus (Peterson, Cooper and Good, 1965). The stimulus is most likely to be radiation or a chemical agent.

Occupational Epidemiologic Studies

Most of the previously cited studies relating leukemia to prior benzene exposures have been of a case-history nature; that is, individual cases of leukemia are described in an anecdotal fashion with little knowledge of the underlying population at risk. While such studies are valuable in pointing out potential health hazards, they do not allow one to quantify the magnitude of the hazard in terms of relative or attributable risk.

Until recently there has been a paucity of studies relating chronic low level industrial exposure to increased leukemia incidence. Ishimaru et al., (1971) have shown that probable industrial exposure to

benzene or medical x-rays results in a relative risk of getting leukemia of 2.5. Two recent studies of the rubber industry have found excessive incidence of leukemia but have been unable to associate these excesses with specific jobs or exposures (Monson and Nakano, 1974; Fox, Lindars and Owen, 1974). A study of petroleum workers exposed to benzene found no excess of leukemia deaths (Thorpe, 1974). However, this study has been questioned as to its methodology (Brown, 1975). McMichael et al. (1975) have examined the association of leukemia with specific jobs involving solvent exposure in the rubber industry. The subject of this dissertation involves statistical considerations related to this last study. A recent compendium of studies on rubber workers has strengthened the argument that there are excessive numbers of leukemia deaths occurring in men who have been employed in this industry (McMichael, Andjelkovic and Tyroler, in press).

The results of several case history studies (Cronkite, 1961; DeGowin, 1963; Erdogan and Aksoy, 1973; Erf and Rhoads, 1939; Falconer, 1933; Forni and Luciano, 1967; Jedlicka et al., 1958; Tareef, Kontchalovskaya and Zorina, 1963; Vigliani and Saita, 1964) suggest that workers exposed to benzene may be at excess risk of developing leukemia. The study of Rejsek and Rejsková (1955), which is based on medical examination of 4538 persons, lacks a stated study design and also lacks any analysis of rates, thus making the results difficult to compare with more formal epidemiologic studies. The study of Ishimaru et al., (1971) is the only epidemiologic study found to specifically implicate benzene. Dinman (1974) takes the position that no epidemiologic evidence exists to substantiate this association. A National Cancer Institute Report (U. S. Department of Health, Education, and Welfare, 1972c)

takes an equivocal stand. Thus the association of benzene with leukemia can only be regarded as suspected at the present time.

CHAPTER II

METHODOLOGY

Retrospective Epidemiologic Study Design

Because of the existence of eight years of death certificates for the Population at Risk (PAR) at the onset of this study and the relative rarity of leukemia as a cause of death, a retrospective case-control study design was chosen to investigate possible associations of specific jobs with deaths from or with leukemia. The subsequent collection of two more years of death certificates for the study population brought the total number of male leukemia deaths up to twenty-three (twenty with leukemia as the underlying cause and three with leukemia as a contributory but not underlying cause).

It is assumed that deaths from leukemia are representative of incident cases. As indicated in Chapter I, leukemia is almost invariably a fatal disease. In addition, there is no reason to believe that potential cases (who have contracted leukemia but are still alive) would be a biased group with respect to their work histories.

Two sets of controls were chosen for this study: an individually-matched set of controls and a group-matched set. These two groups were chosen independently from different sources. The former set of controls was restricted to the deceased portion of the cohort. In addition the study designs are different. The results from the respective statistical analyses, therefore, will not be directly comparable. The

individually-matched controls have been analyzed both by taking the matching into account and by ignoring the matching. A discussion of the question of matching in retrospective studies can be found in an article by Miettinen (1970a) in addition to a recent article by McKinlay (1975).

In the individually-matched phase of this study, each case was matched with three controls on the basis of age-at-death, race and sex. Thus within each quartad all the individuals would be of the same race and sex. In addition, the case would have the youngest age-at-death. Further, each of these three controls had died during the same calendar period as the respective case (1964-73) without evidence of a malignant neoplasm or non-malignant blood disorder on the death certificate. All cases and controls were active or retired hourly workers in the same plant at the onset of the study. To insure against a potential competing risk, each control was chosen with the criterion that age at death was zero to five years greater than the corresponding case. The controls were chosen, within the given constraints, by the following procedure:

1) The entire PAR was listed alphabetically in a computer printout.

2) For each case three sets of two random numbers were chosen to designate a starting place for page and line in the alphabetized listing.

3) From this random starting point on the computer printout, one worked down the printout until a suitable control was found based on sex, race, age-at-death and causes of death.

The initial results of this phase have recently been reported (McMichael et al., 1975).

For the group-matched phase of this study the same case group of twenty-three leukemia deaths was used as in the individually-matched phase. The control group, however, consisted, essentially, of a

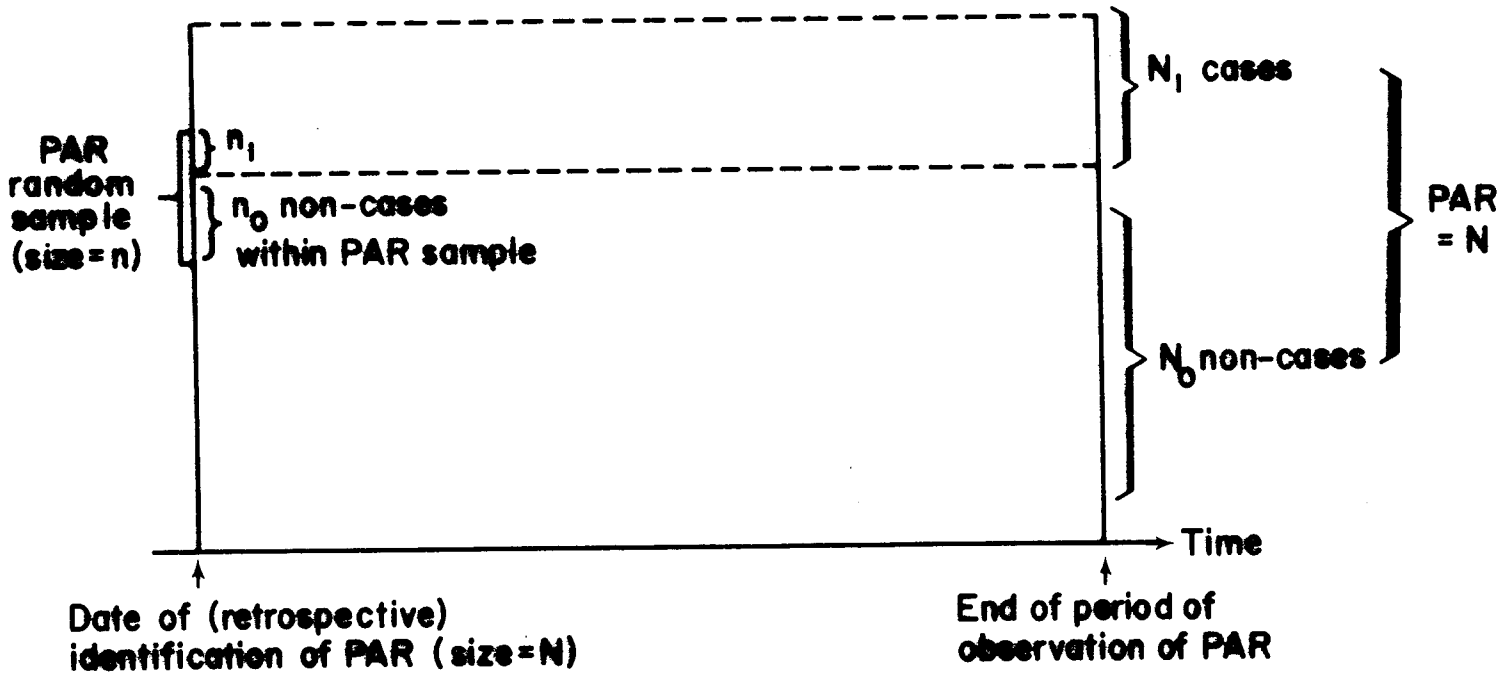
stratified random sample of the original male cohort. Thus, approximately five hundred controls were chosen from each of the age-at-January 1, 1964 groups forty to fifty-four, fifty-five to sixty-four, sixty-five to eighty-four. This represented an approximate twenty-two per cent sample of the total population at risk in the age range forty to eighty-four. However, the sampling strategy resulted in a disproportionately large number of controls in the older age groups in order to more nearly match the age at January 1, 1964 of the cases. Naturally, some of the members of the sample were cases and were retained because of the study design. In addition, controls dying of or with the other diseases including neoplasms were included in the sample in order to allow an accurate estimation of the probability of exposure. A more detailed discussion of the criteria for control selection for Phase II is given in Chapter V. Figure 2 illustrates schematically the study design for the group-matched phase.

Estimates of Relative Risk

One of the most important measures calculated in epidemiologic studies is an estimate of relative risk. In retrospective studies, where the probability of death from the disease of interest is very low, Cornfield's estimate of relative risk is usually given (Cornfield, 1951). In the first phase of the present study, the additional knowledge of the individual matching can be utilized in a measure of relative risk described by Miettinen (1970b).

For the second phase of this study it is possible to construct an estimate of relative risk which makes use of the hybrid nature of the study design shown in Figure 2. This approach, which involves elements

Figure 2
Diagrammatic Representation of Study Design



Source: Kupper, McMichael and Spirtas, 1975.

of cohort as well as case-control study designs, has first been used by McMahon (1962) and the statistical aspects have been discussed by Kupper, McMichael and Spirtas (1975). The estimate of relative risk for a simple random sample can be expressed as a ratio of conditional probabilities:

$$RR = \frac{P(D/E)}{P(D/\bar{E})} = \frac{P(E/D)}{1-P(E/D)} \left[\frac{1}{P(E)} - 1 \right] \qquad (2.1)$$

where

$P(D/E)$ = Probability of disease in the subset of exposed individuals

$P(D/\bar{E})$ = Probability of disease in the subset of nonexposed individuals

$P(E/D)$ = Probability of exposure in the subset of diseased individuals

$P(E)$ = Unconditional probability of exposure in the overall population.

Here, exposure is used in the broad sense to mean working in any job where the worker may come in contact with the suspected etiologic agent.

Assumptions:

$P(E/D)$ is known a priori

$P(E)$ can be estimated from the PAR sample as

$P(E) = \frac{x}{n}$, where x is the number of exposed workers and

n is the size of the simple random sample.

Figure 2 schematically illustrates a PAR of N persons consisting of N_1 cases with a given disease and N_o controls (non-cases). A sample of size n is taken and later found to contain n_1 cases and n_o controls (n_1 and n_o are not known prior to sampling). Work history (or other exposure) information will be collected for the n_o controls. Thus, the study design assumes that exposure information will be collected for all

persons with diseases of interest, in addition to comparable exposure information on a sample of the non-diseased population. Since similar exposure information will be available for the n_1 cases, it will be unnecessary to collect it again for the n_1 cases chosen in the sample. Moreover, it is possible to construct an exact confidence interval (although not necessarily symmetric) for the estimate of relative risk.

For the situation where concomitant variables (such as age) could be exerting a potentially confounding effect, it may be useful to stratify the population. Using the usual formulas for a stratified random sample (Cochran, 1967), yields:

$$RR_i = \frac{P_i(E/D)}{1-P_i(E/D)} \left[\frac{1}{P_1(E)} - 1 \right]$$

where $P_1(E) = \frac{x_1}{n_{(i)}}$, i.e. the ratio of exposed workers in the $i^{\underline{th}}$ stratum of the sample to the total in the $i^{\underline{th}}$ stratum (note that the notation $n_{(i)}$ is used to avoid confusion with case-control designation). A method for standardizing the relative risk estimate on the confounding factor has been given by Miettinen (1972). The adaptation of this technique, given by Kupper, McMichael and Spirtas (1975) will be used in this phase of the study.

Analytical Model

Because of the existence of several potentially confounding variables in this study as well as the wide range of putative exposures involved in different jobs, it was felt appropriate to consider multivariate statistical models for the analysis of the data. The use of multivariate statistical methods for differentiating between populations dates back to the work of Mahalanobis (1930), Hotelling (1931), and

Fisher (1936). The basic motivations for these investigators were completely different. Mahalanobis was interested in a generalized measure of distance between two populations. Hotelling's desire was to generalize "students" t-test to the multivariate case. Fisher was trying to devise a rule for classification. That all three approaches lead to essentially the same result is not only surprising but comforting in the realization that three interpretations can be made from one set of calculations.

In more recent years, there has been a growing interest in the use of discriminant function analysis in the field of biological sciences. Mather (1949) discussed its use in the field of genetics. Radhakrishna (1946) discussed its use in the medical areas of clinical trials, predicting treatment effects, selection of patients for a particular operation and diagnosis. Application of discriminant function analysis to psychometric problems was suggested by Lubin (1950). The technique is discussed by several standard textbooks in the field of biostatistics (Armitage, 1971; Kendall, 1961; and Anderson, 1958).

As a means of conceptualizing the statistical models mentioned above, assume there are two populations (A and B) where p measurements are taken for each individual, (x_i, i=1, ..., p). Assume that x is distributed as the multivariate normal and that the covariance matrices for the two populations are equal. Then there are essentially two standard methods one can base the calculation of the discriminant function upon: the eigenvector approach (following Fisher's derivation) and dummy dependent-variable regression. It has been shown that the two computational approaches yield essentially the same results (Anderson, 1958).

Figure 3 illustrates how the two models nominally differ. Both models are, of course, just computational algorithms leading to the same solution. From an etiological point of view, the regression model is suggested; i.e., in the chain of causality the exposure precedes or is hypothesized to lead to a particular disease. From a statistical point of view, however, the investigator controls or chooses the number of people dying from a certain cause along with an arbitrarily large group of controls in a retrospective study. The disease is an independent or control variable while, on the other hand, exposure is allowed to fluctuate randomly. Thus, exposure is properly a dependent or response variable in this study (in a purely statistical sense). Aesthetically, the eigenvector approach is more appealing.

The argument deciding the issue of which model to use involved the associated concomitant variables. In the individually-matched phase of the study, age-at-death, race and sex were matching variables. Since these variables have been controlled by the matching process and no longer allowed to fluctuate as random variables, they are more properly treated as control variables (in the statistical sense) and can be treated as covariables on the right hand side of the standard statistical equation $Y = X\beta + \varepsilon$. But since the retrospective nature of the epidemiologic design insures that the variable giving information on disease state may be also a control variable, it is therefore necessary to use the eigenvector model in order that all control variables be treated alike.

The Multivariate General Linear Model of Starmer and Grizzle (1968) offers an elegant computer algorithm for formulating the desired problem. Proper construction of the design matrix will yield the desired

FIGURE 3

CONCEPTUAL DIFFERENCES IN STATISTICAL MODELS

Variable / Method	Independent	Dependent
Eigenvector	Indicator of disease	Exposure
Regression	Exposure	Indicator of disease

statistical model, while proper choice of contrast matrices will generate Fisher's Linear Discriminant Function and test other hypotheses of interest, e.g. significance of covariates. The model is of the form:

$$Y_{nxp} = X_{nxq} \quad \beta_{qxp} + \varepsilon_{nxp}$$

Assumptions: Each row of $\varepsilon \sim N_p(\underline{0}, \ddagger)$

where n = total number of cases plus controls

p = number of Occupational Title Groups

q-2 = number of covariates (leaving 1 degree of freedom each for overall mean and effect of case-control designation)

The dependent variable, Y, is the measure of exposure in this model, while the matching variables as well as nuisance variables such as age-at-death and education are treated as covariates. Additional computations are provided by the Statistical Analysis System procedure REGR (Service, 1972).

Analysis of Work History Data

If there existed prior knowledge of which jobs were more hazardous in terms of specific exposures leading to specific diseases, then it might be possible to divide the cohort into various exposure groups and move directly into an historic prospective study. The rubber industry, however, is a complex chemical industry, undergoing continuous change in its raw chemical usage as well as production methods and environmental controls. Past studies have indicated that some chemicals in this industry are carcinogenic (Mancuso, Ciocco and El-Attar, 1968; Case and Hosker, 1954; Boyland, 1954; Mancuso and Brennan, 1970).

Previous epidemiological studies of the rubber industry have attempted to make use of various classifications of rubber workers in

several ways. A study of Social Security records by Guralnick utilized a simple dichotomy--rubber worker or non-rubber worker (Guralnick, 1963). Mancuso et al. (1968) used cross-sectional information on department as a means of separating their study population into five groups. Parkes (1966) designed a 15-group categorization of current departments for his study of sickness absence in the British rubber industry.

Studies of other industries have generally been concerned with exposure to a known or suspected hazardous agent (Lee and Fraumeni, 1969). One of the more sophisticated approaches has been the categorization of steel workers by Lloyd et al. (1970). They classified their population at risk into 54 work areas at initial observation and by areas in which at least five years had been spent.

An even more powerful statistical procedure was utilized by Kramer and Mutchler (1972) in a study where quantitative environmental measurements were available, allowing the estimation of time-weighted yearly average concentrations of vinyl chloride. Their study made use of "step-wise multiple linear regression analysis." Such an accumulation of historical environmental data is extremely rare, although these authors demonstrate the utility of such information. Case et al. (1954) were able to study several cohorts simultaneously by establishing exposure categories for certain chemicals.

The method used in the present study for categorizing the study population was first used by McMichael et al. (1975). The conceptual model has been described by Gamble and Spirtas (in press) and is discussed in detail in Chapter III. Complete work histories are used, resulting in measurements of several distinctly different environmental

experiences for each worker, thus suggesting the use of a multivariate statistical model.

For the present study, over one thousand jobs were classified into seventy-seven Occupational Titles (OT's). While Redmond et al. (1975) in the steelworkers study broke down duration of exposure to a single hazard into six temporal categories, the present study treats time spent in arbitrarily-chosen groups of Occupational Titles as continuous variables. The means of creating such a measurement via the "Experience Transformation Algorithm" (ETA) are discussed more fully in Chapter III. In the situation where potential hazards are not clearly delineated prior to the execution of the study, it is very useful to have the ability to analyze the exposure variable in multivariate continuous form.

The present study is an extension of work done in OHSG projects. This dissertation will examine in detail various statistical models, based on considerations of fit and parsimony. Parametric as well as nonparametric techniques will be considered in an attempt to discover the most sensitive statistical procedure which is appropriate for the analysis of work history data.

Collection of Data

The source of information on cause of death and cytological type of leukemia was the death certificate. All death certificates were coded by a nosologist at the National Center for Health Statistics using the eighth revision of the ICDA Code (U. S. Department of Health, Education, and Welfare, 1969). Underlying as well as multiple causes of death were recorded. Since adult leukemia is usually accurately

diagnosed and followed by the physician until death, the information on specific type of leukemia would usually be available at the time the death certificate was written. Thus, the information on cause of death should be reliable.

The source of environmental information is the work history record. This information is designed for use in the company's personnel department, which makes it a secondary data source. However, because of the financial importance of maintaining precise information on worker skills as well as union interest in accuracy of seniority data, these records are felt to be of sufficient accuracy for determining qualitative differences in the work-place environment.

Photostatic copies of the front covers of personnel folders were made for all cases and controls. In some instances, two or more jacket fronts were stapled together for individuals with a long work history. The upper portion of the front cover gives information on name, social security number, sex, marital status, place of birth and education. The lower portion contains job history information for each job held including the date into job and department number along with a word description of the specific job. In the margin OHSG personnel have added a four-character alphabetic abbreviation called the Occupational Title (OT) code, using an OT dictionary (see Chapter III). In addition, a two-digit line number has been placed after each OT code. This coding was done completely "blind": the coder did not know if the work history belonged to a case or control. A transcription of a personnel folder cover is found in Chapter III.

Coded work histories as well as death certificates were entered into computer-readable form via Optical Character Recognition (OCR)

typing. This involved the use of specially equipped IBM Selectric typewriters. Each record was typed once by each of two typists to minimize typist bias. The OCR sheets were then scanned by a Scandata 2250 Scanner controlled by a PDP-8 Computer System, which entered the raw data onto computer tape. The double records were then compared and any differences were printed out for correction. In this way a high degree of quality control was maintained on the accuracy of the typed information.

CHAPTER III

EXPERIENCE TRANSFORMATION ALGORITHM

Flow of Data

In order to place the Experience Transformation Algorithm (ETA) in perspective to the whole study a flow diagram (Figure 4) has been prepared which indicates the dynamic process which the work history data undergo.

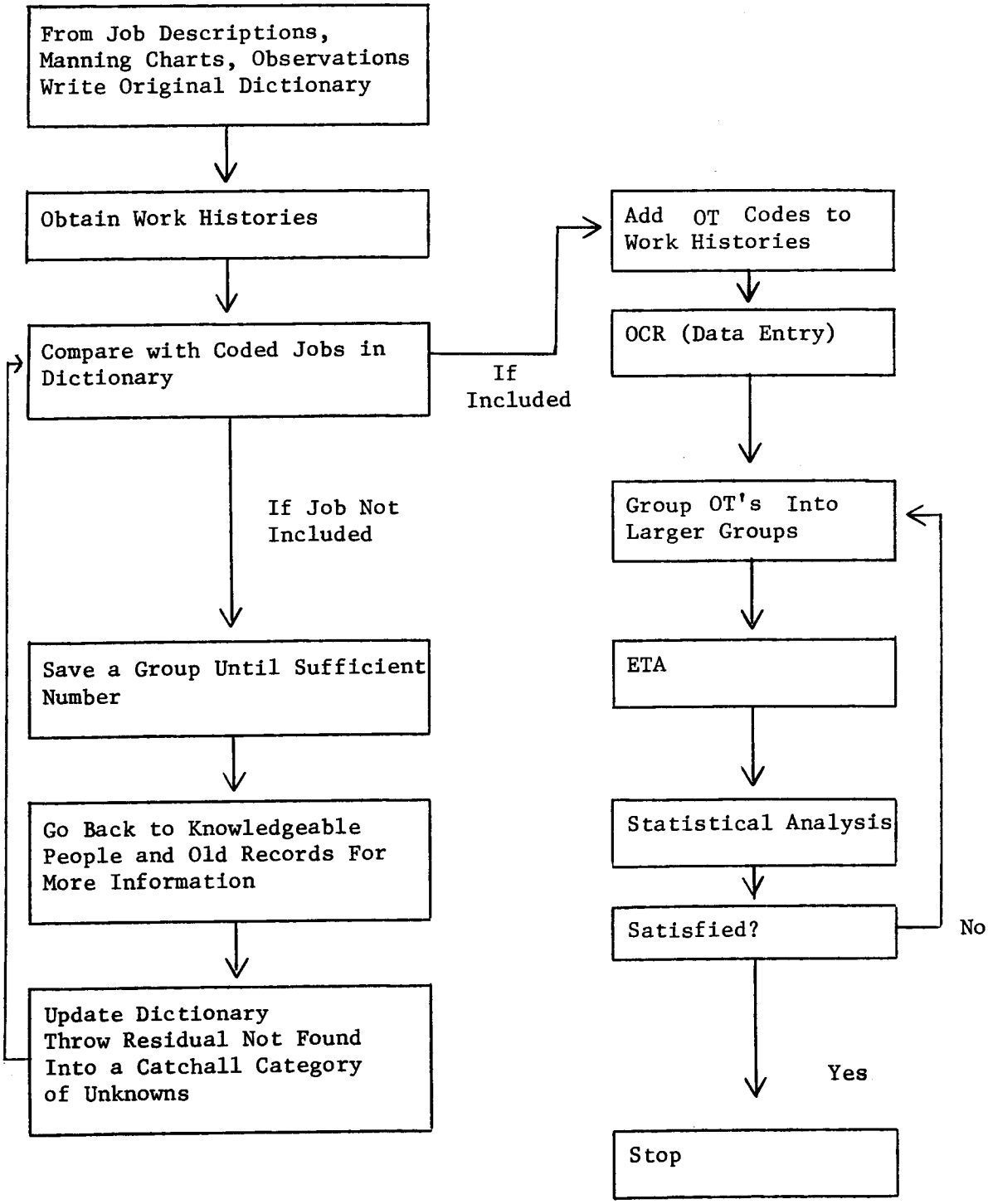
The first essential element in processing work histories is to write a dictionary which translates company-derived job descriptions into some standard format which the computer can easily recognize. The computer code can be numeric or alphabetic, but for ease of handling should be of fixed length. Table 3.1 is a copy of the Occupational Title Dictionary used in this study.

Once the desired work histories are located it is possible to transcribe them onto coding forms on site at the company or to photocopy them for later coding. At this point it is helpful to keep the dictionary open-ended so that additions can still be made. Only after processing a large percentage of the data is it possible to finally complete the OT dictionary.

After adding an OT code and line number to each line of the work history (see Figure 5) the data must be entered into the computer. At OHSG this task is accomplished by Optical Character Recognition (OCR) typing as explained in Chapter II.

FIGURE 4

FLOW DIAGRAM OF OT-ETA PROCEDURE



Source: Gamble and Spirtas (in press).

TABLE 3.1

OCCUPATIONAL TITLE DICTIONARY

Occupational Title	Description
ACTV	Active
BEBL	Bead Building
BPMC	Batch Preparation--Miscellaneous
BPTR	Batch Preparation--Tires
BPTU	Batch Preparation--Tubes, Flaps Bladders
CALO	Calendar Operation
CALT	Calendar Tending
CEMD	Cementing Treads
CEMG	Cement Mixing
CLRK	Clerical
CMTR	Cutting and Milling--Tires
CMTU	Cutting and Milling--Tubes
CRBA	Curing--Bladders
CRFB	Curing--Flaps (black)
CRFW	Curing--Flaps (white)
CRSX	Curing--Miscellaneous
CRTR	Curing--Tires
CRTU	Curing--Tubes
CRVV	Curing--Valves
DEAD	Dead
FINX	Finish and Repair--Flaps, Bladders, Sleeves
FIRA	Finish and Inspect--Tires
FIRB	Repair Tires
FITU	Finish and Repair--Tubes and Airbags
FUSP	Tube and Flap Building
INAC	Inactive
INRE	Inspect and Repair--Green Tires
LOFF	Layoff
MECH	Mechanical Building
MIFB	Milling--Flaps and Bladders
MIMG	Mill Mixing
MIMS	Milling--Miscellaneous
MIPL	Milling--Plystock
MITD	Milling--Treads
MITU	Milling--Tubes
MNCP	Carpentry
MNEC	Mechanic
MNEL	Maintenance--Electrical
MNMA	Machinist
MNMS	Maintenance--Miscellaneous
MNPT	Maintenance--Painting
MNSM	Maintenance--Sheet Metal
MNWL	Maintenance--Welding
MNWR	Maintenance--Millwright

TABLE 3.1 (Continued)

Occupational Title	Description
OTHR	Other jobs
PALI	Paint and Line--Green Tires
PIBE	Pigment Blending
PLHA	Plystock Handling
PWPT	Power Plant
RCDV	Reclaim--Devulcanizing
RCMI	Reclaim--Milling
RCPR	Reclaim--Preparation
RETI	Retired
SALA	Salaried
SBPR	Service--Batch Preparation--Tires
SBPU	Service--Batch Preparation--Tubes
SICK	Sick Suspense
SJAN	Janitor
SLIN	Liner Service
SMOL	Mold Cleaning and Repair
SPCP	Special Products (Defense)
SREC	Reclaim--Service
SRLC	Roll Changing
STBB	Service--Tire and Bead Building
STOD	Tuber Service--Tread Tuber
STOR	Shipping and Receiving
STOX	Tuber Service--Flaps and Bladders
STTU	Tuber Service--Tubes
SYNT	Synthetic Plant
TCKR	Trucking (General)
TEST	Quality Control Testing
TRBL	Tire Building
TUBC	Tuber Operation--Flaps and Bladders
TUBD	Tuber Operation--Treads
TUBU	Tuber Operation--Tubes
UNKN	Unknown
VVMA	Valve Preparation

Source: OHSG Files

FIGURE 5

TRANSCRIBED WORK HISTORY

Name:	John Doe
Social Security Number:	123-21-1975
Birth Date:	05-18-23
Sex:	Male
Marital Status:	M
Place of Birth:	Cleveland, Ohio
Education:	4 yrs. H.S.

O. T.	Line #	Date In	Department
SPCP	01	7-10-41	8-U
SPCP	02	7-22-42	8-U
SPCP	03	9-01-44	Li-38
STBB	04	12-06-44	17-B
STBB	05	12-12-49	131
TRBL	06	10-12-57	131
SALA	07	8-09-65	G.O.
TRBL	08	12-28-65	131
SALA	09	5-01-72	G.O.
ACTV	10	9-01-73	--

The next step is to decide on how to group the individual OT's into larger aggregates suitable for statistical analysis. This step requires a great deal of judgment and is heavily dependent on the hypothesis to be tested. Thus for the current study of the association between solvent exposure and lymphatic leukemia the ten Occupational Title Groups (OTG's) shown in Table 3.2 have been created. This grouping sets up three levels of solvent exposure as well as five process-oriented categories, an "Inactive" category and an "All other" category.

The basic idea of the ETA is to take raw data from individual work histories and somehow manipulate it into a form suitable for statistical analysis. This process will be described in detail in the next section. Thus, the ETA is a computer program which accumulates time as a continuous variable in prespecified OT's. (See Appendix for Program Listing.)

Finally the transformed data are subjected to statistical analysis procedures such as regression analysis or discriminant function analysis. In addition by grouping the data after transformation, it is possible to compute estimates of relative risk or perform other categorical procedures.

The exposure pattern for a specific physico-chemical agent may be intermittent as in the example in Figure 5 and/or low-level in nature. Thus, flexibility in approach is required, emphasizing the need for the data set to come as close as possible to portraying the work history as a dynamic sequence of events.

Computational Procedure

Most work histories have more job entries than the example in Figure 5, averaging fifteen to twenty separate jobs; but the example

TABLE 3.2

A PRIORI OCCUPATIONAL TITLE GROUPS CHOSEN FOR LEUKEMIA STUDY

Definition	OTG	Occupational Titles
High Solvent	1	CALT, CEMD, CEMG, MNPT
Medium Solvent	2	CALO, FIRA, FIRB, FITU, INRE, MNEC, MNEL, MNMA, MNWR, PALI
Low Solvent	3	BEBL, PLHA, STOD, TRBL, VVMA
Compound and Mixing	4	BPTR, BPTU, CMTR, CMTU, MIMG, PIBE, SBPR, SBPU
Milling	5	MIFB, MIMS, MIPL, MITD, MITU
Extrusion	6	STOX, STTU, TUBC, TUBD, TUBU
Curing	7	CRBA, CRFB, CRFW, CRTR, CRTU, CRVV
General Service	8	PWPT, SJAN, STOR, TCKR, TEST
Inactive	9	INAC, LOFF, RETI, SICK
All Other	10	ALLO

Source: Personal Communication--John Gamble, May 30, 1974

used is sufficiently complex to illustrate the utility of the ETA. The individual jobs have been reclassified into OT's by writing in a four letter acronym to the left of each job line. In addition to this OT code, chronological job number, date into the job and department number are entered into a computer record for each line of work history. Information on name, social security number, sex, marital status, date of birth, place of birth and education level are also taken from the work history for entry into the computer. Information on race and age-at-death is entered from other data sources and merged with information from the work history.

From an examination of Figure 5 it is apparent that a raw work history cannot be easily summarized. Workers tend to move from job to job, spending varying amounts of time at each job. The ETA accomplishes the task of aggregating the data by accumulating in the computer continuous measurements of time spent in prespecified groups of OT's (OTG's). The mechanism for accumulating time is, essentially, to subtract the "Date In" from one line of the work history from the "Date In" for the previous line to get time spent (in years and fractions of years) in the previous job. In Figure 5 the worker was a tire builder from 10-12-57 to 8-09-65 and again from 12-28-65 to 5-01-72. The computer program would accumulate 14.3 years in OTG 3 and 17.9 years in OTG 10. For the other eight OTG's the program output would indicate no time had accumulated (see Figure 6).

As a final option the ETA program allows for temporal grouping. The previously mentioned dynamic nature of the chemical usage and manufacturing processes in many industries, including tire manufacturing, suggest that the exposure (even within any given job) may change over

time. Thus, for example, if time spent before a certain date was known to be an important exposure determinant, the program could slice the work history data into time periods. Figure 6 indicates schematically, with a dotted line, a time slice after January 1, 1945. It is relatively easy to examine specific historical periods in investigating chronic diseases such as cancer in order to utilize available information on latency periods. By slicing the work history on the basis of calendar time it becomes possible to narrow down the period under observation to the desired etiologically meaningful period. Since age at exposure may be important the ETA has been designed to slice on age. In addition, to cover the possibility that workers may become sensitized or acclimated to an exposure an option has been included to look at a certain fraction of the work history (e.g., the first one-third).

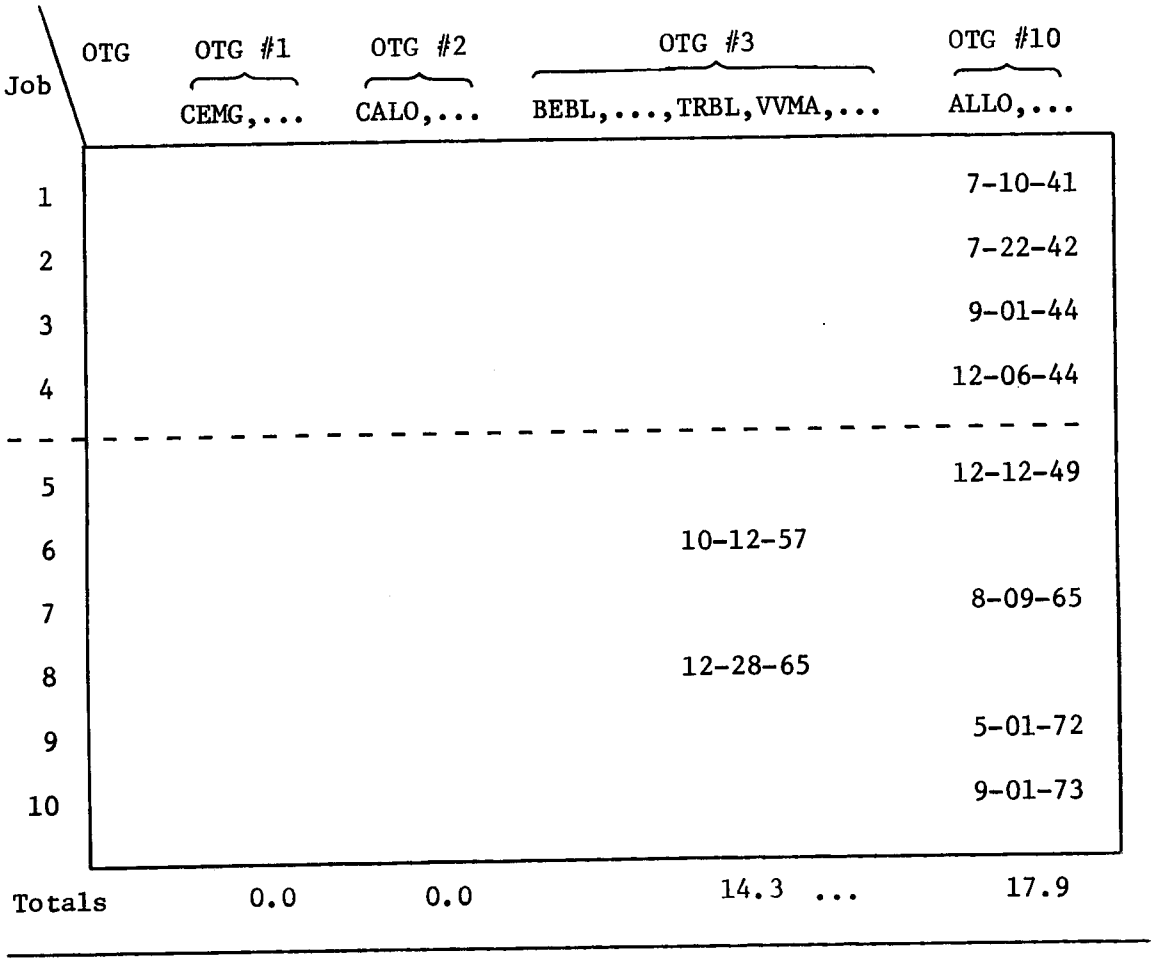
However, the cost of this time slicing is a reduction in the total time measured for each worker. For workers who spent very little time in a hazardous area the use of the time-slicing option may result in output which shows no exposure. In addition, the data may become more skewed in a statistical sense. Thus, a certain amount of care must be used in doing any time slicing.

Summary

The ETA is designed to accommodate the situation where nothing is known about physico-chemical exposures. In such a situation OTG's can still be created using knowledge of materials handled and occupations. If, however, there is some information on exposure it can be utilized by the ETA as in the present study of the relationship of solvent exposure to leukemia deaths. If more quantitative information existed

FIGURE 6

CONCEPTUAL FRAMEWORK OF EXPERIENCE TRANSFORMATION ALGORITHM



on quantitative levels of exposures, it would be possible to modify the program to generate exposure indices rather than simply accumulating time spent on each OTG. The rationale and conceptualization behind this approach to work history utilization are given by Gamble and Spirtas (in press).

The unique aspect of this approach to studying occupational health is the use of complete work history information in measuring the complex experience of each worker under study. The main dependent variables, length of time in prespecified Occupational Title Groups, are recorded in the computer as continuous variables which allows more flexibility in the statistical analysis.

For the present study it has been difficult to obtain quantitative environmental data which shows, historically, solvent exposure by job (or work area) and by time period. It is even more difficult to determine the chemical composition of the solvents used over time. There is at least hearsay evidence that indicates the benzene content of those solvents has been greatly reduced over time. Attempts to obtain such historical data by OHSG industrial hygienists are continuing.

CHAPTER IV

ANALYSIS OF MATCHED QUADRUPLET DATA (PHASE I)

Introduction

As mentioned in previous chapters the twenty-three male cases dying of or with leukemia have been separately matched with two comparison groups: first on an individual basis where three controls have been matched with each case on age-at-death, race and sex and second, on a group basis where a stratified (three date-of-birth strata) random sample has been taken from the entire male Population at Risk (PAR) for use as the control group.

The purpose of this chapter is to explore several statistical approaches to analyzing the work history data after transformation by the Experience Transformation Algorithm (ETA), as well as to present the results of the leukemia case-control study as defined by the former study design. These results are presented in the context of hypothesis testing, i.e., the specific question being asked is whether time spent in solvent-exposed jobs is greater for cases than controls. Several statistical models will be utilized to analyze the data within the matched quadruplet design. Thus, while the results are presented in terms of statements of inference their proper interpretation is more in the vein of hypothesis generating.

Demographic Characteristics

In order to place the cases and controls from the matched quadruplet study design into proper historical perspective, Table 4.1 gives their respective age distributions as of January 1, 1964, the date when the study cohort was defined. The conscious effort to constrain age-at-death to be at least as great for controls as cases in the matched quadruplet design (discussed in Chapter II) has resulted in making this control group slightly older than the cases; i.e., the average age at January 1, 1964 was 64.4 for the controls versus 61.5 for the cases.

The racial breakdown for cases and controls is given in Table 4.2. Because of the matching on this variable, the relative frequency of whites and blacks is identical for the cases and matched controls. However, race, as well as age, could have been a determinant in job placement and mobility as well as in mortality distribution. Thus, even though matching on these variables, it will be necessary to treat them as potential confounding factors and to include them in the statistical analysis.

Since only male cases and controls are considered in this study, sex will not be a variable. It is recognized that leukemia rates differ by sex and that, in the past at least, sex was an important consideration in job placement--certain jobs were considered too dirty or too arduous for women to handle. It will be necessary to pool data from several rubber plants in order to have sufficient data to study the sex variable as well as possible race-sex interactions.

Marital status is given in Table 4.3. This table shows a similar pattern for cases and controls with approximately ninety per cent married workers in each of the two study groups. It is interesting to

TABLE 4.1

AGE DISTRIBUTIONS FOR LEUKEMIA CASES AND CONTROLS IN MATCHED QUADRUPLET STUDY DESIGN, RUBBER WORKERS 1964-1973

	Cases		Controls	
Age at January 1, 1964	Number	(Per Cent)	Number	(Per Cent)
≤ 39	0	(0.0)	2	(2.9)
40 - 44	1	(4.3)	1	(1.4)
45 - 49	1	(4.3)	4	(5.8)
50 - 54	4	(17.4)	8	(11.6)
55 - 59	4	(17.4)	10	(14.5)
60 - 64	4	(17.4)	8	(11.6)
65 - 69	5	(21.7)	12	(17.4)
70 - 74	3	(13.0)	12	(17.4)
75 - 79	1	(4.3)	10	(14.5)
80 - 84	0	(0.0)	2	(2.9)
Totals	23	(100.0)	69	(100.0)

Source: OHSG Files

TABLE 4.2

RACIAL DISTRIBUTIONS FOR LEUKEMIA CASES AND CONTROLS IN MATCHED QUADRUPLET STUDY DESIGN, RUBBER WORKERS 1964-1973

	Cases		Controls	
Race (Code*)	Number	(Per Cent)	Number	(Per Cent)
White (1)	20	(87.0)	60	(87.0)
Black (2)	3	(13.0)	9	(13.0)
Totals	23	(100.0)	69	(100.0)

*Code used in originally entering data into computer.

Source: OHSG Files.

note that none of the cases was single, although this could be only a chance occurrence.

Education, as depicted in Table 4.4a, appears to differ slightly for cases compared with controls. It appears that the cases, on the average, had more education than the controls. In order to create sufficient cell sizes to test this apparent difference, educational level was combined into two cateogries: Low Education Level (< High School) and High Education Level (at least some High School). For the analysis, the cases and controls for whom educational level was missing or unknown, were excluded. The results of the standard Pearson Chi-Square test (without continuity correction) showed a chi-square value of 4.08 which has a p-value of 0.043 (Table 4.4b). If education is taken to be an indicator of socio-economic status, the results of the present study can be said to be in concordance with the work of Sacks and Seeman (1947) showing that socio-economic status is associated with leukemia rates. Since the study population consists entirely of blue-collar workers, one would not expect to see a large gradient for this variable. However, it may be that workers with higher education levels choose to work in the areas of the plant which have greater leukemogenic hazard(s). In order to account for the possible separate effects of socio-economic status and work environment it will be necessary to include education as a variable in the multivariate statistical analyses.

Place of birth, as shown in Table 4.5a, appears to differ for cases and controls. A larger proportion of the cases were U.S. born than the comparable proportion of the individually matched controls.

TABLE 4.3

MARITAL STATUS DISTRIBUTIONS FOR LEUKEMIA CASES AND CONTROLS IN MATCHED QUADRUPLET STUDY DESIGN, RUBBER WORKERS 1964-1973

		Cases		Controls	
Marital Status	(Code*)	Number	(Per Cent)	Number	(Per Cent)
Married	(1)	21	(91.3)	62	(89.9)
Single	(2)	0	(0.0)	5	(7.2)
Divorced/ Separated	(3)	2	(8.7)	1	(1.4)
Widowed	(4)	0	(0.0)	1	(1.4)
Totals		23	(100.0)	69	(100.0)

*Used in originally entering data into computer.

Source: OHSG Files.

TABLE 4.4a

EDUCATION DISTRIBUTIONS FOR LEUKEMIA CASES AND CONTROLS IN MATCHED QUADRUPLET STUDY DESIGN, RUBBER WORKERS 1964-1973

		Cases		Controls	
Education Level	(Code*)	Number	(Per Cent)	Number	(Per Cent)
No High School	(1)	10	(43.5)	37	(53.6)
1-3 Years High School	(2)	7	(30.4)	10	(14.5)
4 Years High School	(3)	3	(13.0)	1	(1.4)
1-3 Years College	(4)	1	(4.3)	3	(4.3)
Missing/Unknown	(9)	2	(8.7)	18	(26.1)
Totals		23	(100.0)	69	(100.0)

*Used in originally entering data into computer.

Source: OHSG Files.

TABLE 4.4b

CHI-SQUARE TEST OF ASSOCIATION ON EDUCATION

	Cases	Controls	
Low Education Level (< High School)	10	37	47
High Education Level (All Other)	11	14	25
	21	51	72

$$\chi_1^2 = \sum_{i=1}^{4} \frac{(O_i - E_i)^2}{E_i} = 4.08$$

Tabulated $_{.05}\chi_1^2 = 3.84 < 4.08$

$p = 0.0434$

Source: Table 4.4a.

TABLE 4.5a

PLACE-OF-BIRTH DISTRIBUTIONS FOR LEUKEMIA CASES AND CONTROLS IN MATCHED QUADRUPLET STUDY DESIGN, RUBBER WORKERS 1964-1973

	Cases		Controls	
Place-of-Birth	Number	(Per Cent)	Number	(Per Cent)
U.S.A.	22	(95.7)	55	(79.7)
Canada	0	(0.0)	2	(2.9)
Europe	1	(4.3)	12	(17.4)
Totals	23	(100.0)	69	(100.0)

Source: OHSG Files.

TABLE 4.5b

CHI-SQUARE TEST OF ASSOCIATION ON PLACE-OF-BIRTH

	Cases	Controls	
U.S.A.	22	55	77
Foreign Born	1	14	15
	23	69	92

$$\chi_1^2 = \sum_{i=1}^{4} \frac{(O_i - E_i)^2}{E_i} = 3.21$$

Tabulated $_{.05}\chi_1^2 = 3.84 > 3.21$

Source: Table 4.5a.

This agrees with the results of the study by Lilienfeld, Levin and Kessler (1972) showing higher leukemia rates for U.S. born than for immigrants. By lumping together the categories Canada and Europe it is possible to test the hypothesis of association. The resulting value of the chi-square statistic (without continuity correction) was 3.21. This corresponds to a posterior level of significance (p-value) of 0.073. Thus, there is insufficient data to claim that there is a significant difference in place of birth for cases and controls.

Statistical Analysis of Matched Quadruplet Data

Group Matched Analysis

In order to take account of the several potential confounding factors as well as to make full use of the experience of each of the cases and controls, the following statistical model was chosen:

$$\underline{Y}_{92\times5} = X_{92\times5}\,\underline{\beta}_{5\times1} + \underline{\varepsilon}_{92\times1} \qquad (4.1)$$

where $\underline{Y}_{92\times1}$ is the vector of total time in solvent-exposed areas computed by the ETA program for the work histories of the twenty-three cases and sixty-nine controls.

$\underline{X}_{92\times5}$ is the design matrix defined by columns as follows:

$X_1 = 1$ to determine overall mean

$$X_2^{\dagger} = \begin{cases} .18058, \text{ if case} \\ -.06019, \text{ if control} \end{cases}$$

$$X_3^{\dagger} = \begin{cases} -.10201, \text{ if education} < \text{High School} \\ .10655, \text{ otherwise} \end{cases}$$

$$X_4^{\dagger} = \begin{cases} -.26919, \text{ if race is black} \\ .04038, \text{ if race is white} \end{cases}$$

$X_5^{\dagger\dagger}$ = normalized age-at-death.

For all of the parametric models in Chapters IV and V the raw data have been transformed before analysis. There were two purposes for making these transformations:

(1) For the dummy variables (case-control code, education code, race code) it was desirable to have an objective coding scheme chosen without prior knowledge of the data.

(2) Since there was a mixture of dummy variables with a continuous variable (age-at-death) it was useful to have all the variables transformed to approximately the same order of magnitude to avoid roundoff error in the matrix inversion process (Draper and Smith, 1968).

---

†Assuming there are n_1 people in category 1 and n_2 in category 2 for any of the variables case-control code, education code or race code, the transformed variable will be Z_i where:

$$Z_i = \begin{cases} \dfrac{-n_2}{\sqrt{n_1 n_2 (n_1+n_2)}}, & \text{if category 1} \\ \dfrac{n_1}{\sqrt{n_1 n_2 (n_1+n_2)}}, & \text{if category 2} \end{cases}$$

††The original age-at-death variable has been transformed by the standard normal transformation:

$$Z_i = \frac{X_i - \bar{X}}{s}$$

where $\bar{X}$ is the sample mean and s is the sample standard deviation for age-at-death.

These goals were accomplished by adding the following constraints:

(1) The dummy variables were recoded in such a way as to be orthogonal to the "βo" column and to have sum of squares equal to unity.

(2) The values for age-at-death were "normalized" by subtracting the sample mean age-at-death from every value and then dividing by the sample standard deviation.

In this way all of the independent variables were objectively recoded in such a way that the majority of their values were between -1.0 and 1.0.

$\underline{\beta}_{5\times1}$ is the vector of unknown parameters

$\underline{\varepsilon}_{92\times1}$ is the error vector

The assumptions for this model are that:

(1) $\varepsilon \sim N(0,\sigma)$

(2) There are no interaction effects

(3) The effects of the covariates are linear.

Thus, the Analysis of Covariance model contains length of service in solvent exposed jobs as the dependent variable; designation as case or control as the independent variable; and education, race and age-at-death as covariables. The logic behind the above-mentioned designation of model (4.1) is given in Chapter II.

The Analysis of Variance (ANOVA) for model (4.1) is given in Table 4.6a. The effects of education and age-at-death were not statistically significant (see Table 4.6b). However, the effect of race was significant; the β coefficient was in the direction which indicated that whites tended to spend more time, after adjusting for the effects of other variables, in solvent-exposed areas than blacks. In addition,

TABLE 4.6a

ANOVA FOR MODEL (4.1)

Source	D.F.	M.S.	F	P
Case-Control Code	1	264.40	3.38	0.070
Education Code	1	9.34	.12	0.731
Race Code	1	510.00	6.51	0.013
Normalized Age-at-Death	1	3.95	.05	0.823
Error	87	78.31		

$R^2 = 0.11$

Source: OHSG Files.

TABLE 4.6b

REGRESSION COEFFICIENTS FOR MODEL (4.1)

Term	β Value	T	P (Two-sided)
Intercept	6.43	6.97	< 0.001
Case-Control Code	16.40	1.84	0.035*
Education Code	3.18	.35	0.731
Race Code	23.06	2.55	0.013
Normalized Age-at-Death	.21	.22	0.823

*One-sided test of the hypothesis that the regression coefficient for the case-control code was greater than zero.

Source: OHSG Files.

the cases spent more time, on the average, than controls in solvent-exposed areas after adjusting for the effects of education, race and age-at-death ($p = 0.035$). Recalling that the controls have been matched on age-at-death, race and sex, the seeming anomaly of the race effect being statistically significant is due to the fact that the dependent variable in this model is the exposure variable (which is often formulated as an independent variable). Thus, while the proportion of white (black) cases is the same as the proportion of white (black) controls; the Analysis of Covariance model indicates that job patterns have been related to race.

In order to determine whether the difference in time spent in solvent-exposed OT's for cases versus controls was the same for cytologic subgroups, the statistical analysis was rerun separately for the cases of lymphatic and myelogenous leukemia. (There were no cases of monocytic leukemia and only two cases of other and unspecified leukemia.)

The statistical model was essentially the same as in the preceding section except for a decrease in the number of cases (and respective controls) and a slight change in the transformation of the dummy variables. For lymphatic leukemia the Analysis of Covariance model for fourteen cases and forty-two controls was:

$$\underset{\sim}{Y}_{56\times1} = X_{56\times5}\,\underset{\sim}{\beta}_{5\times1} + \underset{\sim}{\varepsilon}_{56\times1} \qquad (4.2)$$

The Analysis of Variance for model (4.2) is given in Table 4.7a. Again, from Table 4.7b, there is no statistical significance for the effects of education and age-at-death. Race, again has a significant effect with whites spending more time, on the average, in solvent-exposed jobs than

blacks. In addition the case-control regression coefficient was highly significant, with cases spending more time on solvent-exposed areas than controls after adjusting for the effects of education, race and age-at-death ($p < 0.001$).

For myelogenous leukemia the comparable model was:

$$\underset{28\times1}{Y} = \underset{28\times5}{X}\ \underset{5\times1}{\beta} + \underset{28\times1}{\varepsilon} \qquad (4.3)$$

The Analysis of Variance for model (4.3) is given in Table 4.8a. As in the previous models (4.1) and (4.2) the effects of education and age-at-death are not statistically significant and the effect of race is significant with whites consistently spending more time, on the average, in solvent-exposed OT's than blacks. However, contrary to the previous results, the cases of myelogenous leukemia spent <u>less</u> time, on the average, in the solvent-exposed jobs than did their matched controls after adjusting for the effects of education, race and age-at-death (see Table 4.8b). This inconsistency also comes out in comparing the values of R^2 for the three models. For models (4.2) and (4.3) the values are 0.26 and 0.36 respectively compared with 0.11 for the model (4.1). Lumping together all leukemia deaths in model (4.1) resulted in a poorly fitting model.

Although no formal statistical analysis is presented for the two cases with cytologic definition as other and unspecified leukemia, it is interesting to note that, on the average, these cases spent more time in solvent-exposed areas than did their controls. It would be of interest, if possible, to re-examine any pathology material available for these two cases to try to determine more precisely the cell type since both were classified by ICDA Code 207.9, which is the designation for unspecified leukemia.

TABLE 4.7a

ANOVA FOR MODEL (4.2)

Source	D.F.	M.S.	F	P
Case-Control Code	1	880.30	12.27	0.001
Education Code	1	4.73	.07	0.798
Race Code	1	384.68	5.36	0.025
Normalized Age-at-Death	1	111.70	1.56	0.218
Error	51	71.75		

$R^2 = 0.26$

Source: OHSG Files.

TABLE 4.7b

REGRESSION COEFFICIENTS FOR MODEL (4.2)

Term	β Value	T	P (Two-sided)
Intercept	6.18	5.46	< 0.001
Case-Control Code	29.84	3.50	< 0.001*
Education Code	2.27	.26	0.798
Race Code	20.03	2.32	0.025
Normalized Age-at-Death	1.48	1.25	0.218

*One-sided test of the hypothesis that the regression coefficient for the case-control code was greater than zero.

Source: OHSG Files.

TABLE 4.8a

ANOVA FOR MODEL (4.3)

Source	D.F.	M.S.	F	P
Case-Control Code	1	397.01	5.97	0.023
Education Code	1	114.13	1.72	0.203
Race Code	1	311.15	4.68	0.041
Normalized Age-at-Death	1	203.09	3.05	0.094
Error	23	66.50		

$R^2 = 0.36$

Source: OHSG Files.

TABLE 4.8b

REGRESSION COEFFICIENTS FOR MODEL (4.3)

Term	β Value	T	P (Two-sided)
Intercept	6.84	4.44	< 0.001
Case-Control Code	-21.20	-2.44	0.989*
Education Code	11.76	1.31	0.203
Race Code	20.67	2.16	0.041
Normalized Age-at-Death	-3.21	-1.74	0.094

*One-sided test of the hypothesis that the regression coefficient for the case-control code was greater than zero.

Source: OHSG Files.

It thus seems possible, based on the above findings, to attribute most of the association found between the twenty-three leukemia deaths and solvent-exposed OT's to the strong association between the fourteen lymphatic leukemia cases and these OT's. The fact that this cytologic subgroup also showed the highest SMR in Table 1.3 strengthens the consistency of this finding. Henceforth, this phase of the study will be restricted to cases of lymphatic leukemia and their respective controls. Table 4.9 presents the transformed work history data for the fourteen cases of lymphatic leukemia.

The review of literature in Chapter I indicates that the majority of benzene-induced leukemias in previous studies have been of the myelogenous type. This raises the possibility that the etiologic agent in the present study was some solvent other than benzene, such as toluene or xylene. Even though overt benzene usage has been halted in the rubber industry, it would be prudent to continue following this population of rubber workers to examine the work histories of leukemia deaths for two reasons:

(1) Benzene is still present as a contaminant of other solvents.

(2) One of the solvents used to replace benzene may itself predispose to the excess of lymphatic leukemia.

Although the traditional Analysis of Covariance model assumes no interaction effects it was decided to investigate the analogous regression model containing all first order interactions. The reasons for this exercise were to try to obtain a better fitting model and to check the assumption of no interactions in the Analysis of Covariance model. Here interactions between dichotomous variables can be thought of in the same sense as the interaction in an experimental design with two qualitative levels for two factors. Thus, the following model was chosen:

TABLE 4.9

OUTPUT FROM ETA ON FOURTEEN LYMPHATIC LEUKEMIA CASES USING OTG's FROM TABLE 3.1

Case Number*	Marital Status	Birth Year	Place of Origin	Education Code	Study Cause of Death (ICDA)	Race	Age at Death	Time Spent in each O.T. Group									
								1	2	3	4	5	6	7	8	9	10
1	1	22	1	2	204.1	1	43	--	--	--	4.8	--	--	--	--	.3	18.9
4	1	05	1	1	204.0	1	64	--	31.5	.5	--	--	--	--	10.1	.7	--
9	1	89	1	1	204.1	1	74	2.4	.8	10.9	--	--	--	.1	1.5	--	6.1
12	1	06	1	3	204.1	1	60	--	.8	15.5	--	--	--	--	7.4	.2	23.1
17	1	87	1	1	204.9	1	84	--	--	--	--	--	--	--	--	20.1	26.5
18	3	11	1	1	204.0	1	59	6.8	--	--	--	--	--	--	--	1.6	18.1
26	1	06	1	3	204.1	1	58	--	3.1	15.4	--	--	5.0	--	--	6.1	5.9
31	1	17	1	2	204.1	1	53	--	--	2.6	--	--	--	--	--	1.2	29.0
40	1	07	1	1	204.1	2	59	--	--	--	7.6	18.5	--	--	--	11.0	1.7
44	1	11	1	1	204.0	1	53	.2	.2	--	.3	1.7	--	--	.3	4.6	6.2
45	1	92	1	1	204.9	1	77	--	35.6	--	--	--	--	--	3.1	15.3	1.6
53	1	03	1	2	204.1	1	65	--	--	35.6	--	--	--	--	--	2.2	--
54	1	95	1	2	204.9	2	77	--	--	--	--	--	--	--	31.1	12.8	--
55	1	97	1	3	204.1	1	75	.1	17.8	--	--	--	--	.1	--	15.5	5.6

*Note: There were originally sixty total cases of leukemia. The fourteen cases presented here are the male lymphatic leukemia cases only.

Source: OHSG Files.

$$\underline{Y}_{56\times1} = X_{56\times11}\,\underline{\beta}_{11\times1} + \underline{\varepsilon}_{56\times1} \qquad (4.4)$$

where $\underline{Y}_{56\times1}$ is the vector of total time spent in solvent-exposed areas computed by the ETA for each of the fourteen cases of lymphatic leukemia and their forty-two controls.

$X_{56\times11}$ is the design matrix defined by columns as follows:

X_1 = 1 to determine overall mean

$$X_2^{\dagger} = \begin{cases} .23146, \text{ if case} \\ -.07715, \text{ if control} \end{cases}$$

$$X_3^{\dagger} = \begin{cases} -.12440, \text{ if education} < \text{High School} \\ .14354, \text{ otherwise} \end{cases}$$

$$X_4^{\dagger} = \begin{cases} -.32733, \text{ if race is black} \\ .05455, \text{ if race is white} \end{cases}$$

$X_5^{\dagger\dagger}$ = normalized age-at-death

X_6 = interaction between case-control code and education code = $X_2 * X_3$

X_7 = interaction between case-control code and race code = $X_2 * X_4$

X_8 = interaction between case-control code and normalized age-at-death = $X_2 * X_5$

X_9 = interaction between education code and race code = $X_3 * X_4$

X_{10} = interaction between education code and normalized age-at-death = $X_3 * X_5$

X_{11} = interaction between race code and normalized age-at-death = $X_4 * X_5$

$\beta_{11\times1}$ is the vector of unknown parameters.

$\varepsilon_{56\times1}$ is the error vector.

The assumption for this model is that $\varepsilon \sim N(0,\sigma)$.

Three criteria were used in comparing statistical models:

(1) The value of R^2, the amount of variation explained by the model.

(2) The value of the Mean Square Error, the variation associated with the residuals.

(3) Parsimony in choosing the number of terms to include in the model.

As these criteria are dependent on each other (in fact, they are often competitive) the decisions on which model to choose are somewhat arbitrary. The exercise proceeded by stepwise backward elimination of the term with the least significant partial F value in the model--being careful to retain all lower order effects associated with any interaction term still in the model.

The Analysis of Variance for model (4.4) is given in Table 4.10.

---

†For case-control code, education code and race code the original variables have been transformed as in model (4.1).

††For age-at-death the original variable has been transformed by the standard normal transformation as in model (4.1).

As can be seen by comparing these results with model 4.2, the improvement in fit (R^2 = .34 for model 4.4 versus .26 for model 4.2) has been achieved with only a slight increase in the Mean Square Error (72.22 versus 71.75 respectively).

After going through the above mentioned stepwise backward elimination procedure the final model selected as being the best compromise in meeting the three criteria was model (4.5) as given in Table 4.11a. This model has a smaller Mean Square Error than either model (4.2) or (4.4) and yet has a value for R^2(.32) which is almost as high as that for model (4.4). This is achieved while decreasing the number of terms in the model from eleven [in model (4.4)] to six as follows:

$$\underset{\sim}{Y}_{56\times1} = X_{56\times6}\,\underset{\sim}{\beta}_{6\times1} + \underset{\sim}{\varepsilon}_{56\times1} \qquad (4.5)$$

where the parameters are defined as in model (4.4).

$X_{56\times6}$ is defined by columns as follows:

X_1 = 1 to determine overall mean

$$X_2 = \begin{cases} .23146, & \text{if case} \\ -.07715, & \text{if control} \end{cases}$$

$$X_3 = \begin{cases} -.32733, & \text{if race is black} \\ .05455, & \text{if race is white} \end{cases}$$

X_4 = normalized age-at-death

X_5 = interaction between case-control code and race code = $X_2 * X_3$

X_6 = interaction between case-control code and normalized age-at-death = $X_2 * X_4$

TABLE 4.10

ANOVA FOR MODEL (4.4)

Source	D.F.	M.S.	F	P
Case-Control Code	1	977.58	13.54	< 0.001
Education Code	1	4.59	0.06	0.802
Race Code	1	310.01	4.29	0.044
Normalized Age-at-Death	1	118.46	1.64	0.207
Case-Control * Education	1	26.55	0.37	0.547
Case-Control * Race	1	121.75	1.69	0.201
Case-Control * Age-at-Death	1	190.62	2.64	0.111
Education * Race	1	40.00	0.55	0.461
Education * Age-at-Death	1	75.93	1.05	0.311
Race * Age-at-Death	1	0.66	0.01	0.924
Error	45	72.22		

$R^2 = 0.34$

Source: OHSG Files.

TABLE 4.11a

ANOVA FOR MODEL (4.5)

Source	D.F.	M.S.	F	P
Case-Control Code	1	957.09	14.14	< 0.001
Race Code	1	403.43	5.96	0.018
Normalized Age-at-Death	1	105.99	1.57	0.217
Case-Control * Race	1	174.56	2.58	0.115
Case-Control * Age-at-Death	1	141.90	2.10	0.154
Error	50	67.70		

$R^2 = 0.32$

Source: OHSG Files.

TABLE 4.11b

REGRESSION COEFFICIENTS FOR MODEL (4.5)

Term	β Value	T	P (Two-sided)
Intercept	6.35	5.74	< 0.001
Case-Control Code	31.34	3.76	< 0.001*
Race Code	20.29	2.44	0.018
Normalized Age-at-Death	1.41	1.25	0.217
Case-Control * Race	99.80	1.61	0.115
Case-Control * Age-at-Death	12.17	1.45	0.154

*One-sided test of the hypothesis that the regression coefficient for the case-control code was greater than zero.

Source: OHSG Files.

Neither regression coefficient for the two interactions in model (4.5) is statistically significant. However, their inclusion improves the fit of the model. Thus, the assumption of no interactions in the Analysis of Covariance model (4.2) has not been violated. However, the improvement in fit from the inclusion of these interaction terms in model (4.5) does call for a more detailed investigation of their influence on the main effects. The β regression coefficients for model (4.5) are given in Table 4.11b.

In order to facilitate the discussion of model (4.5), Figure 7 has been constructed. In addition to plotting all fifty-six data points, the expected values (of time in solvent OT's) are shown as regression lines for the four groups: white cases, white controls, black cases and black controls. Several points can be observed:

(1) White cases tended to spend more time in solvent-exposed O.T.s than white controls.

(2) The explanation for the case-control by race interaction is that none of the eight blacks (cases or controls) spent any time in solvent-exposed OT's (see Table 4.12), while most of the white cases had some exposure.

(3) The explanation for the case-control by age-at-death interaction is that the differential between cases and controls tends to increase with increasing age-at-death. Thus, the difference between cases and controls will not be statistically significant if tested at a low enough age (between ages fifty-five and sixty).

(4) The fact that half of the cases died at age $\leq$ sixty suggests that the results presented in Tables 4.11a and 4.11b may be

FIGURE 7
GRAPHICAL DEPICTION OF MODEL (4.5)

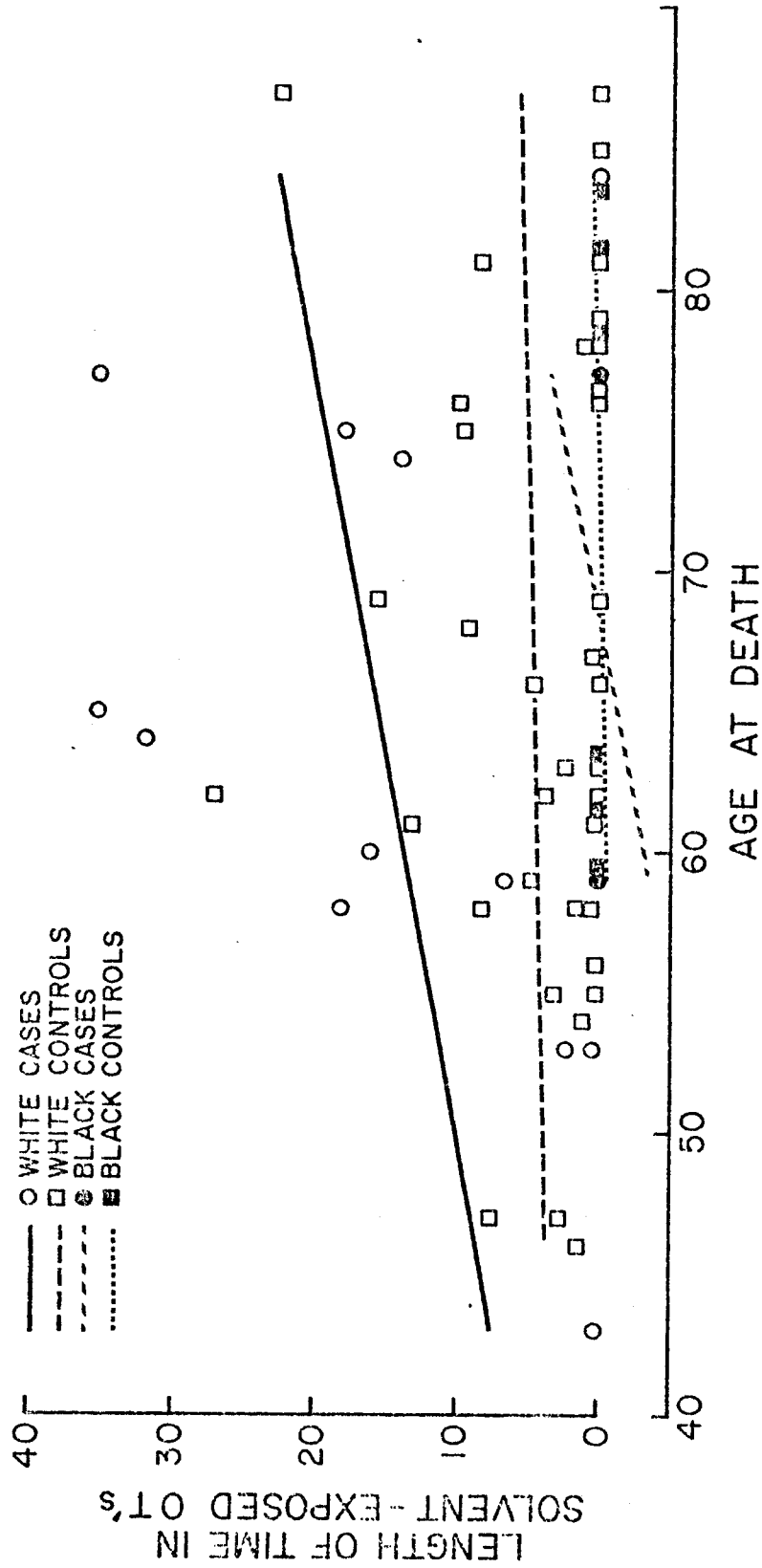


TABLE 4.12

CROSS TABULATION OF EXPOSURE BY RACE
IN MATCHED QUADRUPLET STUDY DESIGN

	Whites	Blacks	
Exposed	33	0	33
Not Exposed	15	8	23
	48	8	56

$$\chi_1^2 = \sum_{i=1}^{4} \frac{(O_i - E_i)^2}{E_i} = 13.39$$

Tabulated $_{.05}\chi_1^2 = 3.84 < 13.39$

Source: OHSG Files.

misleading. When interpreted in the light of the results shown in Figure 7, it appears that the younger cases of lymphatic leukemia (who had the higher SMR in Table 1.3) did not differ appreciably from their controls in terms of solvent exposure.

Of the two black male deaths from lymphatic leukemia one had another cause as the underlying cause of death. Thus, it is difficult to interpret the value of 2.00 for the black male leukemia SMR. Two possible explanations are:

(1) The number of death(s) is (are) due to chance.

(2) Another agent, besides solvents, is involved in at least some of the deaths from (or with) lymphatic leukemia.

In order to overcome the problem of interpreting the data in the presence of the case-control by race interaction in model (4.5), the analysis was subsequently restricted to whites only. This involved excluding the two black male cases and their associated six matched controls, resulting in model (4.6).

$$\underline{Y}_{48\times1} = X_{48\times4}\,\underline{\beta}_{4\times1} + \underline{\varepsilon}_{48\times1} \qquad (4.6)$$

where $X_{48\times1}$ is defined by columns as:

X_1 = 1 to determine the overall mean

$$X_2 = \begin{cases} .25000, \text{ if case} \\ -.08333, \text{ if control} \end{cases}$$

X_3 = (Age-at-death - 65.8)/11.4

X_4 = interaction between case-control code and normalized age-at-death = $X_2 * X_3$

Note: The same logic was used in constructing this model as in model (4.4).

TABLE 4.13a

ANOVA FOR MODEL (4.6)

Source	D.F.	M.S.	F	P
Case-Control Code	1	1108.59	14.52	< 0.001
Normalized Age-at-Death	1	116.18	1.52	0.224
Case-Control * Age-at-Death	1	154.33	2.02	0.162
Error	44	76.34		

$R^2 = 0.27$

Source: OHSG Files.

TABLE 4.13b

REGRESSION COEFFICIENTS FOR MODEL (4.6)

Term	β Value	T	P (Two-sided)
Intercept	7.40	5.84	< 0.001
Case-Control Code	33.71	3.81	< 0.001*
Normalized Age-at-Death	1.58	1.23	0.224
Case-Control * Age-at-Death	12.51	1.42	0.162

*One-sided test of the hypothesis that the regression coefficient for the case-control code was greater than zero.

Source: OHSG Files.

As shown in the Analysis of Variance (Table 4.13a) for model (4.6) the interaction effect is close to achieving statistical significance so it is retained in the model. It is of interest to note that model (4.6) does not fit the data (for whites only) as well as does model (4.5) (for blacks as well as whites). In fact not only is the value for R^2 lower (.27 versus .32), but in addition the mean square error has increased (76.34 versus 67.70). Thus in a purely statistical sense model (4.5) is better. But in light of the fact that no blacks spent any time in solvent-exposed jobs, model (4.6) is felt to be more appropriate. The latter model is more parsimonious and continues to show the significance of the case-control code as a predictor of solvent exposure (see Table 4.13b). However, the inclusion of the case-control by age-at-death interaction term can be interpreted, as in Figure 7, to weaken the solvent exposure hypothesis. For the remainder of this chapter (except model 4.8) the emphasis will be placed on analyzing solvent exposure for white males only.

The data for white males only were next examined using a non-parametric test. The Mann-Whitney U Test (Siegel, 1956) was chosen as an appropriate distribution-free two-sample test of the difference in medians. Table 4.14 gives the data for the length of time in solvent-exposed jobs for the twelve white cases of lymphatic leukemia and their matched controls. In addition, the data have been ranked with ties receiving the average rank. Thus, the fifteen persons with no time spent in the solvent-exposed areas each received an average rank of 8.0. Because of the relatively large sample sizes ($n_1 = 12$, $n_2 = 36$), the large sample formulae were utilized. By setting the sum of the ranks

TABLE 4.14

TIME (IN YEARS) IN SOLVENT-EXPOSED OTG's AND ASSOCIATED RANKS* FOR TWELVE WHITE CASES OF LYMPHATIC LEUKEMIA AND MATCHED CONTROLS IN MATCHED QUADRUPLET STUDY DESIGN

Case Number**	Time Spent by Case	Rank	Time Spent by Control 1	Rank	Time Spent by Control 2	Rank	Time Spent by Control 3	Rank
1	0	8	2.82	26	7.56	31	1.49	23
4	31.94	46	0	8	10.77	37	9.22	34
9	14.15	39	9.74	35	0	8	10.16	36
12	16.29	41	13.32	38	.14	16	2.54	24
17	0	8	0	8	0	8	22.88	44
18	6.78	30	4.85	29	0	8	27.02	45
26	18.45	43	4.07	28	0	8	.78	21
31	2.57	25	.52	19	8.35	32	3.37	27
44	.43	18	1.15	22	0	8	0	8
45	35.63	48	0	8	8.67	33	0	8
53	35.60	47	0	8	.68	20	15.92	40
55	17.93	42	0	8	0	8	.25	17
Σ Ranks		395		237		217		327

*For the fifteen subjects with no time in solvent-exposed OTG's an average rank of 8 has been assigned.

**From prior OHSG Study.

Source: OHSG Files.

of the larger group (controls) equal to R_2, the value of U was computed as:

$$U = n_1 n_2 + \frac{n_2(n_2 + 1)}{2} - R_2 = 317.$$

Then using the notation of Siegel (1956) the value of Z is calculated where:

$$Z = \frac{U - \frac{n_1 n_2}{2}}{\sqrt{(\frac{n_1 n_2}{N(N-1)})(\frac{N^3-N}{12} - \Sigma T_i)}} \quad .$$

Here $N = n_1 + n_2 = 48$

$T = \frac{t^3-t}{12}$ for t = number of ties at a given rank

Since the only tie occurred at the score of 0.0,

the value of $\Sigma T = T = \frac{(15)^3-15}{12} = 280.$

Thus, the value of Z = 2.442 (which corresponds to a p-value of 0.007 for the one-tailed test using a table of the normal distribution) was found using the test statistic. It can therefore be concluded, without assumptions as to the distribution of the data, that the white cases spent more time than the white controls, on the average, in solvent-exposed OT's. The disadvantage of this approach is that it does not account for the effect of age-at-death or the interaction of case-control code by age-at-death on the length of time spent in these jobs.

An estimate of relative risk can be made from this retrospective data by first arraying the data as in Table 4.15 according to whether or not cases and controls ever worked in a solvent-exposed OTG. Then using Cornfield's formula (1951) the odds ratio ($\psi = \frac{ad}{bc}$) can be computed

TABLE 4.15

SOLVENT EXPOSURE AMONG WHITE MALE LYMPHATIC LEUKEMIA CASES AND CONTROLS IN MATCHED QUADRUPLET STUDY DESIGN

	Cases	Controls	
Exposed	10	23	
Not Exposed	2	13	
Totals	12	36	48

Source: OHSG Files.

TABLE 4.16a

EXPOSURE TO OTG's 1, 2 AND 3 AMONG WHITE LYMPHATIC LEUKEMIA CASES AND CONTROLS IN MATCHED QUADRUPLET STUDY DESIGN

		Cases	Controls	
OTG 1 High Solvent	Exposed	4	3	
	Not Exposed	8	33	
	Totals	12	36	48
		Cases	Controls	
OTG 2 Medium Solvent	Exposed	7	16	
	Not Exposed	5	20	
	Totals	12	36	48
		Cases	Controls	
OTG 3 Low Solvent	Exposed	6	14	
	Not Exposed	6	22	
	Totals	12	36	48

Source: OHSG Files.

TABLE 4.16b

RELATIVE RISK ESTIMATES AND NINETY-FIVE PER CENT CONFIDENCE INTERVALS FOR OTG's 1, 2 AND 3

OTG		$\hat{RR}$	Lower C.I.	Upper C.I.
1	High Solvent	5.50	1.13	29.64
2	Medium Solvent	1.75	.47	6.57
3	Low Solvent	1.57	.42	5.84

Source: OHSG Files.

to be 2.826. The use of this estimate of relative risk is generally acceptable provided that the disease under study is relatively rare, which is certainly true for lymphatic leukemia. By using the approximate formula for variance of the odds ratio (Armitage, 1971 formula 16.6) it is possible to estimate $\text{Var}[\ln \psi] = \frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d} = .7204$. The ninety-five per cent confidence interval for $\ln \psi$ is calculated as $\ln \psi \pm 1.96 \sqrt{\text{Var}(\ln \psi)}$, which yields (L,U). To get the corresponding confidence limits for ψ, it is necessary to transform the data back to get (e^L, e^U) as the approximate ninety-five per cent confidence interval for $\psi = (.54, 14.91)$. Thus, it appears that there is insufficient evidence to claim that the relative risk is significantly different from 1.0. Further, because of the small numbers involved, it does not seem useful to further subdivide the population on the basis of age-at-death. However, it is possible to get separate estimates of relative risk for each of the three individual solvent OTG's. The appropriate data are given in Table 4.16a and the corresponding estimates of relative risk along with approximate confidence intervals are given in Table 4.16b. Even though the only estimate to differ significantly from 1.0 was the one for the High Solvent OTG, the gradient of relative risk for the three OTG's strengthens the hypothesis that the risk of lymphatic leukemia is related to the level of solvent exposure within the plant.

Matched Quadruplet Analysis

The results of the previous section showed an association between solvent exposure and lymphatic leukemia. However, the statistical analysis ignored the individual matching of each case with three controls. Thus, although the previous model did take race, education

and age-at-death into account as covariates, it did so only on a group basis.

The purpose of the present section is to analyze the data utilizing the matching framework and taking into account as many covariables as possible. Within each quartad there will be one case and three controls--all of the same race. The following model was used for the matched quadruplet analysis:

$$(Y-\bar{Y}) = \beta_0^* + \beta_1(X_1-\bar{X}_1) + \beta_2(X_2-\bar{X}_2) + \beta_3(X_1-\bar{X}_1)(X_2-\bar{X}_2) + \varepsilon \qquad (4.7)$$

where $Y-\bar{Y}$ = length of time spent in solvent-exposed jobs by the case minus the average time spent by the three matched controls.

$(X_1-\bar{X}_1)$ = difference in normalized age-at-death** between case and the average of three matched controls. (This difference will always be negative due to the matching constraints.)

$(X_2-\bar{X}_2)$ = difference in education codes*** between case and the average of three matched controls.

$(X_1-\bar{X}_1)(X_2-\bar{X}_2)$ = interaction between difference in age-at-death and difference in education codes.

---

*Note: Here β_0 can be interpreted as the case-control effect.

**Normalized age-at-death = (Age-at-death - 65.8)/11.4.

***Education code = $\begin{cases} -.14434, & \text{if < High School} \\ .14434, & \text{otherwise.} \end{cases}$

TABLE 4.17a

ANOVA FOR MODEL (4.7)

Source	D.F.	M.S.	F	P
Difference in Age-at-Death	1	1.29	.01	0.80
Difference in Education	1	17.03	.07	0.94
Interaction (Age * Education)	1	.55	.00	0.96
Error	8	246.61		

$R^2 = 0.10$

Source: OHSG Files.

TABLE 4.17b

REGRESSION COEFFICIENTS FOR MODEL (4.7)

Term	β Value	T	P (Two-sided)
Intercept	11.85	.62	0.28*
Difference in Age-at-Death	5.90	-.07	0.94
Difference in Education	29.21	.26	0.80
Interaction (Age * Education)	25.63	-.05	0.96

*One-sided test of the hypothesis that the value of β_o in model (4.7) is significantly greater than zero.

Source: OHSG Files.

TABLE 4.18

DIFFERENCES IN TIME IN SOLVENT-EXPOSED OTG's AND ASSOCIATED SIGNED RANKS FOR TWELVE WHITE CASES OF LYMPHATIC LEUKEMIA AND MATCHED CONTROLS

Case Number	Difference*	Signed Rank
1	-3.955	-4
4	25.276	10
9	7.516	5
12	10.960	7
17	-7.628	-6
18	-3.842	-3
26	16.830	8
31	-1.515	-2
44	0.044	1
45	32.744	12
53	30.067	11
55	17.849	9
	Σ negative ranks	-15

*Difference computed as time in solvent OTG's for case less the average time in solvent OTG's for the respective three matched controls.

Source: OHSG Files.

Here again the β's represent the unknown parameters and ε is the residual error term (Assumption: $\varepsilon \quad N[0,\sigma^2]$).

The ANOVA for model (4.7) is given in Table 4.17a. The β regression coefficients are given in Table 4.17b. By the process of backward elimination it was found that none of the terms in the model had significant regression coefficients. In addition, the low value of $R^2(0.10)$ in Table 4.17a suggested that the differences in time spent by the cases and their matched controls $(Y-\bar{Y})$ be examined via the standard "paired-t" test. The raw data are given in Table 4.18. The calculated value for $t_{11} = 2.67$ corresponds to a p-value of 0.011 for the one-sided hypothesis that the average difference between a case and the average of the three matched controls was positive.

This result is concordant with the grouped analysis of model (4.2). However, taking the matching into account reduces the number of degrees of freedom for error to eight. Thus, the difference between cases and controls as depicted by the intercept term in Table 4.17b, while in the same direction, is not statistically significant in model (4.7).

The "paired-t" test is robust against non-normality. However, in order to provide a measure similar to the nonparametric test used in the group-matched analysis, the Wilcoxon Signed-Rank Test (Bradley, 1968) was used on the data in Table 4.18. The resulting sum of the ranks of the negative differences totaled fifteen, corresponding to a posterior confidence level of approximately 0.03. This indicates results similar to the above "paired-t" test; namely, that the cases spent significantly more time in the solvent-exposed areas than did the controls.

An estimate of relative risk was made using the data from the matched quadruplet design. This estimate differs from the estimate of relative risk in the previous section in that it treats each quartad of case and matched controls as the unit of observation. The method of computing relative risk and an approximate confidence interval are given by Mantel and Haenszel (1959) and Miettinen (1970b), respectively. The raw data are given in Table 4.19. Then, using essentially the notation in Miettinen (1970b),

R = Number of matched controls per case = 3

S_m = Number of entries in Table 4.19 whose index values total m

$\hat{\rho}$ = Estimate of relative risk

$$I(\hat{\rho}) = \sum_{m=1}^{R} \frac{S_m(R+1-m)/m}{\hat{\rho}[\hat{\rho}+(R+1-m)/m]^2} .$$

Then from equations (4.3), (4.6) and (4.7) in Miettinen (1970b) estimates can be computed for $\hat{\rho}$, $\underline{\rho}$ and $\bar{\rho}$ which correspond with the point estimate of relative risk and the lower and upper approximate confidence limits.

$$\hat{\rho} = \sum_f \sum_g Z_{fg} f(R-g) / \sum_f \sum_g Z_{fg}(1-f)g$$

Here Z_{fg} are the entries in Table 4.16 where

$$f = \begin{cases} 0, & \text{if case never exposed} \\ 1, & \text{if case ever exposed} \end{cases}$$

$$g = \begin{cases} 0, & \text{if no controls exposed} \\ 1, & \text{if 1 control exposed} \\ 2, & \text{if 2 controls exposed} \\ 3, & \text{if 3 controls exposed} \end{cases}$$

TABLE 4.19

CROSS TABULATION OF CASES AND CONTROLS USING EACH QUARTAD AS THE UNIT OF OBSERVATION IN THE MATCHED QUADRUPLET STUDY DESIGN

Cases Ever Present in Any Solvent-Exposed OTG		Number of Matched Controls Ever Present in Any Solvent-Exposed OTG				Total Number of Matched Quadruplets
		3	2	1	0	
Yes	(1)	2	5	3	0	10
No	(0)	1	0	1	0	2

Source: OHSG Files.

From this formula the point estimate of relative risk is $\hat{\rho} = 2.75$.

To compute the estimate for the approximate lower bound for the ninety-five per cent confidence interval the following formula is used:

$$\underline{\rho} \doteq \exp\{\ln \hat{\rho} - 1.96/\hat{\rho}[I(\hat{\rho})]^{\frac{1}{2}}\}$$

$$\underline{\rho} = .53.$$

The corresponding upper bound is computed as

$$\bar{\rho} \doteq \exp\{\ln \hat{\rho} + 1.96[\ln\{1 + \underline{\rho}^2/\hat{\rho}^4 I(\hat{\rho}^2/\underline{\rho})\}]^{\frac{1}{2}}\}$$

$$\bar{\rho} = 22.59.$$

Therefore, by taking the matching into account the point estimate of relative risk is computed to be 2.75 with a ninety-five per cent confidence interval of [.53,22.59]. While the point estimate and lower confidence interval estimate compare favorably with the previous results which ignored the matching, the upper confidence interval is much greater for the present estimate when compared with the previous estimate (22.59 versus 14.91, respectively).

Multivariate Response Model

Finally, it was decided to expand the analysis of model (4.2) to the multivariate response case for expository purposes. Here the model is said to be multivariate in the sense that there are now ten response variables corresponding respectively to the amount of time each person spent in each of the ten OTG's. This model is expressed by the following matrix equation:

$$Y_{56\times10} = X_{56\times5}\,\beta_{5\times10} + \varepsilon_{56\times10} \qquad (4.8)$$

The observations are taken from the fourteen cases of lymphatic leukemia and their respective forty-two matched controls. The least squares estimates $\hat{\beta}$ are given in Table 4.20. This disaggregation of the

solvent-exposed jobs into OTG's 1, 2 and 3 will result in variables with even more skewness than in the univariate model. Thus, while inferential statements are not warranted, it is possible to make some observations regarding the regression coefficients.

The regression coefficients for the effect of case-control designation are positive for all three solvent-exposure OTG's implying that, on the average, cases spent more time in each of these OTG's than did controls. The only other OTG's with a positive coefficient for the case-control effect were General Service and Inactive. After reexamining the individual jobs included in the General Service OTG it was discovered that one of the jobs, janitoring, may well have involved exposure to solvents. Thus, the multivariate model supports the findings of the prior univariate models. The existence of positive coefficients, for the education effect, in Low Solvent and General Service can be explained by the fact that tire builders and skilled craftsmen respectively make up a large per cent of these two OTG's. Both of these jobs have traditionally required a relatively high educational level for admittance. The race effect coefficients indicate that whites spent relatively more time in solvent-exposed jobs in addition to jobs in curing as well as the miscellaneous All Other category, while blacks spent relatively more time in Compounding and Mixing, Milling and General Service which have been regarded as the dirtier jobs with generally higher dust exposure. The age-at-death coefficient indicates that older workers tend to have spent more time in inactive status (sick, laid off, etc.) over their entire careers, which is intuitively reasonable.

TABLE 4.20

REGRESSION COEFFICIENTS FOR MODEL (4.6)

Time in OTG		Overall Mean	Case-Control Code	Education Code	Race Code	Normalized Age-at-Death
High Solvent	(1)	.18	2.15	-1.49	.75	-.07
Medium Solvent	(2)	3.73	12.88	-5.28	13.74	1.37
Low Solvent	(3)	2.27	14.81	9.05	5.53	.19
Compound/ Mix	(4)	.96	-.32	-3.72	-11.78	-.32
Milling	(5)	1.62	-.72	-4.58	-23.97	-.21
Extrusion	(6)	.39	-.59	2.33	.35	-.43
Curing	(7)	1.77	-7.63	2.41	5.08	.09
General Service	(8)	2.24	7.17	11.67	-22.72	.99
Inactive	(9)	7.91	1.90	7.39	-6.81	4.41
All Other	(10)	1.39	-16.50	-6.24	23.81	-.82

Source: OHSG Files.

Summary

In summary, because of the great deal of job mobility and the large number of jobs, there appears to be a correspondingly large variance in length of time spent in specific areas of the rubber tire plant under study. When this variability in job patterns is added to the inherent difference in individual susceptibility, any etiological relationship will be difficult to determine. In other words, statistical models can only explain a certain percentage of the variability in such a situation. While standard univariate nonparametric procedures can be utilized, they do not take into account covariates or multivariate responses. Aside from this procedural statistics problem, the results tend to confirm the hypothesis that exposure to solvents was associated with the excess of lymphatic leukemia deaths in this plant. However, these results must be considered in light of the case-control by age-at-death interaction, as well as the fact that neither black case spent any time in a solvent-exposed OT.

CHAPTER V

ANALYSIS OF GROUP-MATCHED DATA (PHASE II)

Introduction

The discussion in this chapter will, in general, be parallel with the discussion of Chapter IV. Thus, there will be information presented on the demographic characteristics of both the cases and the comparison group, results of statistical modeling and an estimate of relative risk. However, the results in this chapter differ, for the following reasons, from the results in Chapter IV:

(1) The comparison group of controls is an age-stratified random sample from the male PAR. Thus, the controls are, in effect, matched with cases on sex (study subjects are all male) and plant, but not on race or age. One result of the stratification procedure used was a closer age distribution of cases and controls.

(2) It is assumed that for purposes of computing relative risk, the only stochastic element is the random sampling of the PAR (i.e. the comparison group). This assumption is different from the one made in most retrospective studies (which assume that $P(E|D)$ is stochastic) and results in shorter confidence intervals.

Thus, the comparison of Phases I and II can be done on a qualitative basis only, since the statistical assumptions and comparison groups are

different for the two phases.

Once again, the work histories of all twenty-three cases dying of or with leukemia have been utilized. The data are analyzed both in an hypothesis-testing framework using statistical models similar in nature to ones used in Chapter IV as well as in an hypothesis-generating mode to explore the data for individual jobs which might be especially hazardous. By the limitation of the study design for this phase, the data will be analyzed only as a comparison between two groups.

Demographic Characteristics

The decision to take a stratified sample of the PAR was based on knowledge of age and length of service. It was known that a group of cases dying from diseases of adulthood would tend to be older than the rest of a cohort of workers defined as alive at some previous point in time. It was also desirable to avoid the potential dilution effect of recent hirees, i.e., employees who would tend to be much younger (at the point in time when the cohort was defined) than the cases. Table 5.1 illustrates the cross tabulation of age on January 1, 1964 by employment year for the study cohort. The large number hired during the period 1940-1944 represents workers hired for war-time production during World War II.

The stratified random sample from the PAR consisted of subsamples of 496, 498 and 505, respectively, from the age strata forty to fifty-four, fifty-five to sixty-four and sixty-five to seventy-nine. This represented sampling fractions of 19.7%, 24.8% and 24.3%, respectively. The stratification was done prior to and independent of the decision to study the association of leukemia with solvent exposure. A decision

TABLE 5.1

JANUARY 1, 1964 AKRON MALE RUBBER WORKERS COHORT

Age at 1/1/64	Employment Year												
	1905 -1909	10-14	15-19	20-24	25-29	30-34	35-39	40-44	45-49	50-54	55-59	60-64	Totals
40-44							2	530	125	154	10	10	831
45-49						11	122	453	134	143	9	3	875
50-54					28	75	123	377	102	95	13	3	816
55-59				20	214	189	174	245	78	58	3	1	982
60-64			24	151	294	158	116	187	59	35	1		1025
65-69		6	62	209	250	130	90	165	44	9			965
70-74	1	22	69	150	169	78	40	103	29	4	1		666
75-79	1	23	59	78	85	31	14	50	7				348
80-84	1	5	25	17	28	7	1	16					100
Totals	3	56	239	625	1068	679	682	2126	578	498	37	17	6608
		(1%)	(4%)	(9%)	(16%)	(10%)	(10%)	(32%)	(9%)	(8%)	(1%)		
Service dates awaiting correction, or missing													70
													6678

Source: OHSG Files.

was made to try to dampen the discrepancy in age between cases and the PAR by taking a stratified random sample with proportionately larger samples from the older age groups. The consequence of this procedure can be seen in examining Table 5.2. The average age at January 1, 1964 for the stratified sample was 60.0 compared with 60.6 for the cases.

The racial distributions for cases and controls are given in Table 5.3. Even though race was not a matching variable in Phase II, it can be seen that there is little difference between cases and controls for this variable.

Since only males were eligible for control selection, sex was not a variable in this study. This was also true in Phase I. It is recognized that sex may be quite important when determining job pattern, but because of the relatively small number of female cases with many of the diseases under study in this cohort, it was decided to limit the controls to males only in the hybrid study design discussed in Chapter II.

Marital status is given in Table 5.4. Here, as in Phase I, the overwhelming majority of controls, as well as cases, were married as of January 1, 1964. Given the age and stable nature of the PAR sample (most of whom had been employed at the same plant for over ten years), it is not surprising to see such a large married proportion.

The distributions for education among cases and controls is given in Table 5.5a. Since it appeared that the cases tended to have more education, on the average, than the controls, a chi-square test of homogeneity was calculated with the data grouped into categories of low education level (no high school) and high education level (all other categories). The results of this test are given in Table 5.5b, which shows that the association was not statistically significant.

TABLE 5.2

AGE DISTRIBUTION FOR LEUKEMIA CASES AND CONTROLS IN GROUP-MATCHED STUDY DESIGN, RUBBER WORKERS 1964-1973

	Cases		Controls	
Age at January 1, 1964	Number	(Per Cent)	Number	(Per Cent)
40-44	1	(4.3)	161	(10.7)
45-49	1	(4.3)	167	(11.1)
50-54	4	(17.4)	168	(11.2)
55-59	4	(17.4)	242	(16.1)
60-64	4	(17.4)	256	(17.1)
65-69	5	(21.7)	243	(16.2)
70-74	3	(13.0)	153	(10.2)
75-79	1	(4.3)	109	(7.3)
Totals	23	(100.0)	1499	(100.0)

Source: OHSG Files.

TABLE 5.3

RACIAL DISTRIBUTIONS FOR LEUKEMIA CASES AND CONTROLS IN GROUP-MATCHED STUDY DESIGN, RUBBER WORKERS 1964-1973

	Cases		Controls	
Race	Number	(Per Cent)	Number	(Per Cent)
White	20	(87.0)	1284	(85.7)
Black	3	(13.0)	209	(13.9)
Missing/Unknown	0	(0.0)	6	(0.4)
Totals	23	(100.0)	1499	(100.0)

Source: OHSG Files.

TABLE 5.4

MARITAL STATUS DISTRIBUTIONS FOR LEUKEMIA CASES AND CONTROLS IN GROUP-MATCHED STUDY DESIGN, RUBBER WORKERS 1964-1973

	Cases		Controls	
Marital Status	Number	(Per Cent)	Number	(Per Cent)
Married	21	(91.3)	1361	(90.8)
Single	0	(0.0)	70	(4.7)
Divorced/Separated	2	(8.7)	43	(2.9)
Widowed	0	(0.0)	25	(1.7)
Totals	23	(100.0)	1499	(100.0)

Source: OHSG Files.

TABLE 5.5a

EDUCATION DISTRIBUTIONS FOR CASES AND CONTROLS IN GROUP-MATCHED STUDY DESIGN (PHASE II)

	Cases		Controls	
Education Level	Number	(Per Cent)	Number	(Per Cent)
No High School	10	(43.5)	792	(52.8)
1-3 Years High School	7	(30.4)	379	(25.3)
4 Years High School	3	(13.0)	235	(15.7)
1-3 Years College	1	(4.3)	40	(2.7)
4+ Years College	0	(0.0)	5	(0.3)
Missing/Unknown	2	(8.7)	48	(3.2)
Totals	23	(100.0)	1499	(100.0)

Source: OHSG Files.

TABLE 5.5b

CHI-SQUARE TEST OF ASSOCIATION ON EDUCATION

	Cases	Controls	
Low Education Level (< High School)	10	786	796
High Education Level (All Other)	11	658	669
Totals	21	1444	1465

$$\chi_1^2 = \sum_{i=1}^{4} \frac{(O_i - E_i)^2}{E_i} = .387$$

Tabulated $_{.05}\chi_1^2 = 3.84 > .387$

Source: Table 5.5a.

The distributions of place-of-birth for cases and controls are given in Table 5.6. Here, as in Chapter IV, a higher percentage of cases, relative to controls, was born in the United States. The difference was not statistically significant.

Since marital status and place-of-birth were so invariant, these variables were not included in the ensuing statistical analyses. Age had been shown to have little effect on job patterns in a preliminary OHSG report so it was also omitted from the initial set of analyses. This left the variables education and race (along with case-control designation) to be utilized as potential predictors of time in solvent-exposed OT's.

Statistical Analysis

Analogous to Chapter IV, this exercise begins with the following Analysis of Covariance model:

$$\underset{\sim}{Y}_{1516\times1} = X_{1516\times4}\,\underset{\sim}{\beta}_{4\times1} + \underset{\sim}{\varepsilon}_{1516\times1} \qquad (5.1)$$

where $\underset{\sim}{Y}_{1516\times1}$ is the vector of total time spent in solvent-exposed OTG's at any time in their careers by the twenty-three cases of leukemia and the 1493 members of the PAR sample with known race.

$X_{1516\times4}$ is the design matrix defined by columns as follows:

$X_1 = 1$ to determine the overall mean

$$X_2^{\dagger} = \begin{cases} .20693, \text{ if case} \\ -.00319, \text{ if control} \end{cases}$$

TABLE 5.6

PLACE-OF-BIRTH DISTRIBUTIONS FOR LEUKEMIA CASES AND CONTROLS IN GROUP-MATCHED STUDY DESIGN, RUBBER WORKERS 1964-1973

	Cases		Controls	
Place of Birth	Number	(Per Cent)	Number	(Per Cent)
United States of America	22	(95.7)	1315	(87.7)
Canada	0	(0.0)	1	(0.1)
Central America	0	(0.0)	1	(0.1)
South America	0	(0.0)	1	(0.1)
Europe	1	(4.3)	174	(11.6)
Other	0	(0.0)	4	(0.3)
Missing/Unknown	0	(0.0)	3	(0.2)
Totals	23	(100.0)	1499	(100.0)

Source: OHSG Files.

$$X_3^{\dagger} = \begin{cases} -.02436, & \text{if education} < \text{High School} \\ .02708, & \text{otherwise} \end{cases}$$

$$X_4^{\dagger} = \begin{cases} -.06370, & \text{if race is black} \\ .01036, & \text{if race is white} \end{cases}$$

Since the majority of workers in the PAR sample are still alive, it was not practical to use age-at-death as one of the covariables. However, age-at-death was found to be generally unimportant in Chapter IV (its regression coefficient was not statistically significant). Therefore, the inherent discrepancies in the models for Phases I and II were felt to be not great.

The Analysis of Variance for model (5.1) is given in Table 5.7a. As in model (4.1) the effect of education is not significant and the effect of race is highly significant (see Table 5.7b). However, the effect of the case-control code is not statistically significant, with controls spending almost as much time, as a group, in solvent-exposed OT's as cases. The value of $R^2(.08)$ is low.

Following the pattern of Phase I, the next step was to restrict the analysis to cases of lymphatic leukemia only and again use the PAR sample as the control group. Thus, model (5.2) was of the form

$$\underset{\sim}{Y}_{1507\times1} = X_{1507\times4}\ \underset{\sim}{\beta}_{4\times1} + \underset{\sim}{\varepsilon}_{1507\times1} \qquad (5.2)$$

where the variables are defined analogously to model (5.1) with slight changes in the transformed variables to account for the reduction in cases.

---

†See footnote to model (4.1) for explanation of codes.

TABLE 5.7a

ANOVA FOR MODEL (5.1)

Source	D.F.	M.S.	F	P
Case-Control Code	1	.22	< .01	0.967
Education Code	1	10.70	.08	0.776
Race Code	1	332.13	3.37	<0.001
Error	1512	132.38		

$R^2 = 0.08$

Source: OHSG Files.

TABLE 5.7b

REGRESSION COEFFICIENTS FOR MODEL (5.1)

Term	β Value	T	P (Two-sided)
Intercept	9.77	33.05	< 0.001
Case-Control Code	.48	.04	0.484*
Education Code	-3.29	-.28	0.776
Race Code	128.44	11.11	< 0.001

*One-sided test of the hypothesis that cases spend more time in solvent-exposed OT's rather than controls.

Source: OHSG Files.

The Analysis of Variance for model (5.2) is given in Table 5.8a and the β regression coefficients are given in Table 5.8b. As expected, there is little difference compared to the results of model (5.1) with the exception of the case-control effect which appears to be relatively more important in model (5.2), although still not achieving statistical significance ($p = .099$ for the one-sided test). Thus, while the direction of the difference is the same as in Phase I the magnitude of the change has decreased.

Further attempts were made to refine the Analysis of Covariance model in Phase II; Year of Birth was added as a variable; Whites were examined separately in an attempt to eliminate the effect of race and interaction terms were added to the model. However, in each case the value of R^2 was disappointingly low (less than 0.11) and the value of the Mean Square Error was much higher than in Phase I (in the range of 130). The modeling exercise was terminated at this point because of the lack of fit.

The method of estimating relative risk for Phase II (Kupper, McMichael and Spirtas, 1975) has been discussed in Chapter II, which defined the notation. However, for purposes of convenience these definitions are repeated below:

$P(D|E)$ = Probability of disease in the subset of exposed individuals.

$P(D|\bar{E})$ = Probability of disease in the subset of nonexposed individuals.

$P(E|D)$ = Probability of exposure in the subset of diseased individuals.

TABLE 5.8a

ANOVA FOR MODEL (5.2)

Source	D.F.	M.S.	F	P
Case-Control Code	1	220.56	1.66	0.197
Education Code	1	10.14	.08	0.782
Race Code	1	16,370.18	123.43	< 0.001
Error	1503	132.63		

$R^2 = 0.08$

Source: OHSG Files.

TABLE 5.8b

REGRESSION COEFFICIENTS FOR MODEL (5.2)

Term	β Value	T	P (Two-sided)
Intercept	9.80	33.04	< 0.001
Case-Control Code	14.85	1.29	0.099*
Education Code	-3.20	-.28	0.782
Race Code	128.63	11.11	< 0.001

*One-sided test of the hypothesis that cases spend more time in solvent-exposed OT's than controls.

Source: OHSG Files.

P(E) = Unconditional probability of exposure in the overall population.

Assumptions:

(1) $P(E|D)$ is known a priori.

(2) $P(E)$ can be estimated from the PAR sample as $\frac{x}{n}$, where x is the number of exposed workers and n is the size of the simple random sample.

Exposure was defined as any positive time spent in one of the three solvent OT's. Death from lymphatic leukemia was chosen as the outcome variable because of the results of the modeling exercise.

First, the overall relative risk was estimated using data from all fourteen cases of lymphatic leukemia and from all 1493 controls (with known race). The data and computations for this estimate are given in Table 5.9. The results yield a relative risk estimate of 1.16 with a ninety-five per cent confidence interval (1.05,1.28).

By means of comparison, the Cornfield estimate of relative risk was next computed (Cornfield, 1961). Since this procedure assumes that the population under study is divided into two mutually exclusive groups--cases and noncases--it was necessary to remove the two overlapping cases from the PAR sample before making the calculations (see Table 5.10). The point estimate of ψ is 1.16 and the ninety-five per cent confidence interval is (.36,3.72).

The fact that the point estimates obtained via the two formulas are so close is no accident. The assumption made in the Cornfield method that $P(E|\bar{D}) \doteq P(E)$ is an equality in the KMS formulation (if $P[E|\text{PAR Sample}]$ is substituted for $P[E|\bar{D}]$). Thus, the results are mathematically equivalent except for the minor difference due to

TABLE 5.9

OVERALL ESTIMATE OF RELATIVE RISK OF LEUKEMIA DUE TO SOLVENT EXPOSURE USING KMS PROCEDURE

	Cases	PAR Sample
E	10	1019
$\bar{E}$	4	474
Totals	14	1493

$$P(E|D) = \frac{10}{14} = .71429$$

$$P(\bar{E}|D) = \frac{4}{14} = .28571$$

$$P(E) \hat{=} \frac{1019}{1493} = .68252 = \hat{P}$$

$$\hat{RR} = \frac{P(E|D)}{P(\bar{E}|D)} \left[\frac{1}{P(E)} - 1\right] = \frac{10}{4}\left[\frac{1}{.68252} - 1\right] = 1.16290$$

Let $\hat{P}(E) = \hat{P}$

Then $$\hat{L} = \hat{P} - Z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}(1 - \frac{n}{N})}$$

$$L = .68252 - 1.96\sqrt{\frac{(.68252)(1 - .68252)}{1493}(1 - \frac{1493}{6678})}$$

$$\hat{L} = .66171$$

$$\hat{U} = p + Z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}(1 - \frac{n}{N})} = .70333$$

$$\text{Confidence Interval} = \left[\frac{P(E|D)}{P(\bar{E}|D)}\left(\frac{1}{\hat{U}} - 1\right), \frac{P(E|D)}{P(\bar{E}|D)}\left(\frac{1}{\hat{L}} - 1\right)\right]$$

$$C.I. = \left[\frac{10}{4}\left(\frac{1}{.70333} - 1\right), \frac{10}{4}\left(\frac{1}{.66171} - 1\right)\right.$$

$$C.I. = [1.05, 1.28]$$

could change to reflect age stratification.

TABLE 5.10

OVERALL ESTIMATE OF RELATIVE RISK OF LEUKEMIA DUE TO SOLVENT EXPOSURE USING CORNFIELD PROCEDURE

	Cases	Controls	
E	10	1018	
$\bar{E}$	4	473	
Totals	14	1491	1505

$$RR = \frac{ad}{bc} = \frac{10(473)}{4(1018)} = 1.16159 = \psi$$

$$\ln \psi = .14979$$

$$\text{Var } (\ln \psi) \approx \frac{1}{a} + \frac{1}{b} + \frac{1}{c} = \frac{1}{d} = .35310$$

$$\text{S.D. } (\ln \psi) = .59422$$

$$95\% \text{ C.I. on } \ln \psi = .14979 \pm 1.96(.59422)$$

$$= [-1.01488, 1.31446]$$

$$95\% \text{ C.I. on } \psi = [e^{-1.01488}, e^{1.31446}]$$

$$\text{C.I.} = [.36, 3.72]$$

subtracting the appropriate values to account for the two overlapping cases.

The two sets of confidence intervals are quite different. This is not surprising since the KMS assumption that $P(E|D)$ is not stochastic should intuitively result in a tighter confidence interval. In other words, by statistically inferring only from the sample to the finite population of the PAR, rather than to some larger population, it becomes possible to create much tighter confidence bounds on the estimate of relative risk.

However, inferences from the KMS estimate are severely limited. It is only possible to interpret the findings from Table 5.9 to imply that the relative risk would probably be greater than 1.05 if the cases were compared with the entire PAR. It would not be proper to use the statistical findings to infer statistically to any other population or time period. However, wider inference by informed judgment would be appropriate.

Since the previous statistical analyses had shown the importance of race as a confounding variable, it was decided to stratify the data on the basis of race. It is then possible to remove the component of risk due to the confounding factor to get an estimate of "residual relative risk." The computations in Table 5.11 show that the estimate of residual relative risk = $\hat{R}_r = 1.18$ with ninety-five per cent confidence interval (1.07,1.29). Therefore, the effect of ignoring race was to slightly dampen the measure of relative risk in Table 5.9.

As a final use of the relative risk formula the estimates of relative risk for each of the three solvent OTG's is given in Table 5.12 along with the respective ninety-five per cent confidence interval (for

TABLE 5.11

ESTIMATE OF RELATIVE RISK$_{KMS}$ OF LEUKEMIA DUE TO SOLVENT EXPOSURE, STRATIFIED BY RACE

	Whites			Blacks	
	C	PAR		C	PAR
E	10	957	E	0	62
$\bar{E}$	2	327	$\bar{E}$	2	147
Totals	12	1284	Totals	2	209

$$P_1(E|D) = \frac{10}{12} \qquad P_2(E|D) = 0$$

$$n_1 = 1284 \qquad n_2 = 209$$

$$P^*(E) = \sum_{j=1}^{2} P_j(E|D)\,\frac{n_j}{n} = \frac{1}{1493}\left[\frac{10}{12}(1284) + 0(209)\right] = .71668$$

$$\hat{R}_r = \frac{P^*(E)}{1-P^*(E)}\left[\frac{1}{\hat{P}(E)} - 1\right] = \frac{.71668}{.28332}\left[\frac{1}{.68252} - 1\right] = 1.18$$

$$\hat{L} = \hat{P}(E) - Z_{\alpha/2}\sqrt{\frac{\hat{P}(E)[1-\hat{P}(E)]}{n}\left(1 - \frac{n}{N}\right)}$$

$$= .68252 - 1.96\sqrt{\frac{(.68252)(1 - .68252)}{1493}\left(1 - \frac{1493}{6678}\right)}, \text{ where } \alpha = .05$$

$$= .68252 - .02082$$

$$\hat{L} = .66170$$

$$\hat{U} = .68252 + .02082$$

$$\hat{U} = .70334$$

$$\text{C.I.} = \left[\frac{P^*(E)}{1 - P^*(E)}\left(\frac{1}{\hat{U}} - 1\right), \frac{P^*(E)}{1 - P^*(E)}\left(\frac{1}{\hat{L}} - 1\right)\right]$$

$$= \left[\frac{.71668}{.28332}\left(\frac{1}{.70334} - 1\right), \frac{.71668}{.28332}\left(\frac{1}{.66170} - 1\right)\right]$$

$$\text{C.I.} = [1.07, 1.29]$$

TABLE 5.12

ESTIMATES OF RELATIVE RISK$_{KMS}$ OF LEUKEMIA
FOR INDIVIDUAL OTG'S FOR WHITES

	OTG	$\hat{RR}$	Lower 95% C.I.	Upper 95% C.I.
High Solvent	(1)	4.68	4.01	5.57
Medium Solvent	(2)	1.38	1.25	1.52
Low Solvent	(3)	1.36	1.23	1.50
Total Solvent		1.71	1.53	1.90

whites only to avoid the confounding effect of race). As can be seen by examining this table, the risk is much greater for the High Solvent OTG than for the other two OTG's. The presence of a gradient (although the estimates for OTG's 2 and 3 are quite similar) parallels the results from Table 4.16b. This tends to confirm the belief that the jobs in OTG 1 are the most hazardous with respect to the risk of getting lymphatic leukemia. (Note that for whites only, the overall $\hat{RR} = 1.71$.)

Finally, in an effort to determine if any individual Occupational Title within the three solvent OTG's was contributing an extraordinary amount toward the discrepancy between cases and controls, it was decided to examine the proportion of OT's worked in by cases and controls. Table 5.12 presents the results of this exploratory investigation. For this analysis, each individual job has been treated as the unit of observation rather than each person. Thus, the results are not directly comparable to the previous analyses which utilized a measure of total time spent in groups of OT's. Because of the limitations imposed on cell sizes in breaking down OTG's to the individual OT level, all twenty-three cases of leukemia were used as the case group in this analysis.

As can be seen from examining Table 5.13 several individual OT's were relatively more prevalent among cases than among controls, namely: Cementing Treads (CEMD), Cement Mixing (CEMG), Maintenance-Painting (MNPT), Finishing and Inspection-Tires (FIRA), Repair Tires (FIRB), Plystock Handling (PLHA), Tuber Service-Tread Tuber (STOD) and Tire Builder (TRBL). Since FIRA and FIRB are closely related jobs on the assembly line, it seems natural to single out these two OT's for comparison. By combining the two it can be seen that 14.4% of all jobs

TABLE 5.13

PROPORTION OF OT's HELD IN OTG's 1, 2 AND 3 BY CASES OF LEUKEMIA AND PAR SAMPLE

			Cases		Controls	
OTG	OT	Description	Per Cent	Number	Per Cent	Number
1	CALT	Calendar Tending	0.3	1	0.5	144
1	CEMD	Cementing Treads	2.2	7	0.4	143
1	CEMG	Cement Mixing	1.9	6	0.3	106
1	MNPT	Maintenance - Painting	0.3	1	0.1	44
2	CALO	Calendar Operation	0.0	0	0.3	100
2	FIRA	Finish/Inspect - Tires	9.4	30	4.8	1534
2	FIRB	Repair - Tires	5.0	16	1.3	410
2	FITU	Finish/Repair - Tubes, Airbags	0.9	3	1.2	380
2	INRE	Inspect/Repair - Green Tires	0.3	1	1.0	330
2	MNEC	Mechanic	0.0	0	0.1	40
2	MNEL	Maintenance - Electrical	0.0	0	0.4	124
2	MNMA	Machinist	0.0	0	0.9	279
2	MNWR	Maintenance - Millwright	0.0	0	1.3	401
2	PALI	Paint/Line - Green Tires	0.0	0	0.9	294
3	BEBL	Bead Building	0.0	0	0.3	93
3	PLHA	Plystock Handling	4.7	15	2.7	873
3	STOD	Tuber Service - Tread Tuber	2.8	9	2.2	714
3	TRBL	Tire Building	8.8	28	4.8	1523
3	VVMA	Valve Preparation	0.0	0	0.0	16

Source: OHSG Files.

held by cases compared with 6.1 per cent of all jobs held by controls were in the same OT's. This suggested that these two OT's, especially, be tested for solvent exposures by industrial hygiene analytical sampling--which is presently being done by OHSG personnel.

Summary

In summary, the analyses done in this chapter imply that the association between solvent-exposed OT's and leukemia is not statistically significant (for the cohort under study) when adjustment is made for the effects of education and race. However, the continued existence of a gradient for the relative risk estimate confirms the belief that the High Solvent OT's have been the most hazardous with respect to lymphatic leukemia. The comparison of these results with the results of Chapter IV will be given in Chapter VI.

CHAPTER VI

CONCLUSIONS AND DISCUSSION

Summary

The primary goals of this study, i.e. to "analyze statistically the association between work experience and occurrence of leukemia using two retrospective study designs" has been accomplished. The ETA computer program has been written and tested. The utility of the ETA is such that it has already been employed in several occupational health studies in addition to the present study. Multivariate statistical methods have been used with varying degrees of success. Race has been found to be the most important covariable among those studied, reflecting past differential job placement.

With regard to the hypothesis that the excess of leukemia deaths is associated with solvent exposure, it appears that this association is limited to lymphatic leukemia. Further, it appears that the association is strongest for the High Solvent OTG. The interactions depicted in Figure 7, as well as the fact that neither of the two black cases had any solvent exposure, are not consistent with the hypothesis under study. Hopefully, the collection of additional data by the OHSG, will be helpful in giving more precise answers.

Interpretation of Results

Since the SMR for lymphatic leukemia (Table 1.3) was 2.2, it was interesting to find the Relative Risk estimates for exposure to solvents to be 2.1 (Table 4.15) and 2.8 (Table 4.19) respectively, for the group-matched and pair-matched analyses in Phase I and 1.2 (Table 5.9) for the group-matched analysis in Phase II. An interpretation of this is that there is an overall increase in the risk of lymphatic leukemia owing to general exposure within the rubber factory and that, super-imposed on this risk increment, is a component of risk, due to specific exposures to solvents, which is approximately of the same order of magnitude.

The great difference between the SMR's for the age ranges forty to sixty-four and sixty-five to eighty-four (7.08 and 0.77, respectively) may indicate a secular trend ("cohort effect") of an increasing leukemogenic hazard, or it may be explained by a process of biologic susceptibility by which those susceptibles are soon afflicted and only the hardy non-susceptible individuals live to age sixty-five.

The results of model (4.5) and Figure 7, which show the effects of race and age-at-death on time spent in solvent-exposed jobs, are at odds with the above-mentioned difference in SMR's. The plot of predicted values for whites implies that the exposure differential between cases and controls increases with age-at-death. So the younger cases, who show the increase in relative risk for lymphatic leukemia, have less of a difference in solvent exposure when compared with their matched controls than do the older cases. This finding is inconsistent with the hypothesis that increased time in solvent-exposed jobs is associated with increased risk of lymphatic leukemia.

Two possible explanations for the gradient of risk exhibited in this study are as follows:

(1) There are at least two separate chemical hazards within the plant; first, a solvent exposure and second, an as-yet-undetermined exposure (or exposures) which is (are) pervasive throughout the plant.

(2) Solvents evaporate quickly and the ventilation system quickly spreads the pollutant across the plant, even to workers who do not directly work with these chemicals. In addition solvents are trucked from point to point in the plant allowing for exposures along the avenue of transit. Also, solvents used in the manufacturing process do not become a part of the finished tire--rather they evaporate somewhere within the plant, possibly at some distance from the initial point of use.

The second explanation seems more plausible, especially since the list of chemicals known to cause blood dyscrasias (Table 1.4) points to benzene as the most likely etiologic agent known to be used in the rubber industry.

The relative advantages and disadvantages of pair-matching versus group-matching have been discussed for over twenty years. Recent work by McKinlay (1975) discusses the utility of pair-matching in the design stage coupled with covariance adjustment in the analysis stage of a study. However, in the present study a precise comparison is difficult because different assumptions were made for each study design. Also, the sample sizes were quite different for the respective control groups, and different covariates were used. In spite of all these differences

it is interesting to note that both approaches yielded similar results, i.e. that exposure to solvents is associated with a greater than expected number of deaths due to lymphatic leukemia. This association increases with level of exposure. Under both study designs, race has been found to be a significant confounding factor, having a high association with job patterns. From the review of literature it is uncertain whether race, per se, has any effect on the risk of developing leukemia. Under both study designs it has been difficult to develop a parametric model which satisfies the underlying statistical assumption of normality. Use of univariate nonparametric statistical procedures has yielded results essentially the same as with parametric models.

However, the differences in results for the two phases of this study are equally enlightening. The pair-matched phase showed a much stronger association of solvent-exposure and lymphatic leukemia than did the group-matched phase. As a general principle, one would expect a larger sample size (for the controls) to more sharply define any true underlying differences in exposure for the cases and controls. That the opposite is true (see Table 6.1) is somewhat puzzling.

The most likely explanation for the disparity of results in the two phases of this study is that, due to random fluctuation, the average solvent exposure for the control group in Phase I was abnormally low. This argument would lead to discounting the strength of the association in Phase I (Chapter IV). Thus, even though great care was taken to make the control selection as random as possible (within the constraints of the study design), chance variation may have produced a control group with unusually low solvent exposure.

TABLE 6.1

AVERAGE TIME (IN YEARS) SPENT IN SOLVENT OT's
BY WHITE MALES

	Phase I	Phase II
Cases*	11.3	11.3
Controls	4.1	9.8

*Fourteen male cases of lymphatic leukemia.

Source: OHSG Files.

Another possible factor in this discrepancy of results involves the restrictions underlying the choice of respective control groups. In Phase I, the controls were constrained not only by age-at-death, race and sex, but also by cause of death. Thus, in order to be eligible for control selection in this phase, the former worker had to have died without any evidence of a neoplasm on the death certificate. This latter constraint was not a requirement for eligibility in Phase II.

Now if solvent-exposure (or more accurately any exposure in jobs where solvents are used) were causally related to neoplasms other than lymphatic leukemia, then the restrictions in Phase I would magnify the difference in time spent in solvent-exposed OT's for the cases and controls. This would be done by reducing the exposure time, on the average, for the controls. In fact, a recent study by McMichael, et al. (in press) on this same cohort of workers shows associations between solvent-exposed jobs and cancers of the bladder, colorectal system, lymphatic and hematopoietic system, prostate, respiratory system and stomach. Appearing repeatedly with respect to exposures in cement mixing and cementing treads are two OT's which, in addition to being in the High Solvent OTG, were also found to be individually suspect in Table 5.12. Thus, while the results of both phases of this study are valid, given the underlying study designs, care should be exercised in discussing the single disease outcome of lymphatic leukemia exclusive of other neoplasms.

Since only a minority of the PAR sample died during the period of study and, of those who died, only a fraction died of cancer (approximately twenty per cent), it is unlikely that the restrictions on control selection in Phase I could have accounted for very much of the

discrepancy between the two phases of this study. It is still conceivable that the solvent hazard may be much broader and more serious than indicated by the results of the present study. Thus, it is possible that exposure to solvents promotes the carcinogenic etiology or acts as a cocarcinogen requiring genetic or enzymatic susceptibility to particular types of malignancies in order to initiate the carcinogenic outcome.

The capability for temporal slicing of work histories with the ETA exists and has been utilized in other OHSG studies (McMichael et al., 1975, McMichael et al., in press). However, this option was not used in the present study for two reasons:

(1) The latency period for chemically induced leukemia is not well defined.

(2) With the emphasis on model fitting in this study it was felt inappropriate to throw away any data, especially when the study group of greatest interest (the lymphatic leukemias) contained only fourteen cases.

When data becomes available for larger case groups (possibly for diseases with well-defined latency periods), it is recommended that the work history analysis be restricted to the etiologic period of interest.

Ideas for Further Study

In any study involved with a new field or with new data sources, there are likely to be as many questions raised and new ideas generated as answers found. The present study which utilizes work history data in a statistical analysis linking exposure to qualitatively different environments with epidemiologic findings on an excess of leukemia deaths is no exception. In order to help in conceptualizing these

ideas, they have been categorized into statistical extensions and environmental-epidemiological extensions although there is considerable overlap involved.

Statistical Extensions

(1) Nonparametric Multivariate Analysis. The statistical models used in Chapters IV and V which were multivariate in nature involved the assumption of normality. The nature of work history data, especially for some of the rarer jobs, is quite skewed with many workers spending no time in the particular job but a few spending many years at that particular job. It would be worthwhile to explore the utility of procedures like Rank Analysis of Covariance (Quade, 1967) as a distribution-free way of examining the data.

(2) Estimation of Latency Period. In the present study the Experience Transformation Algorithm has been used as a tabulating mechanism to keep track of total experience in several work areas for cases and controls. It seems reasonable to extend the computer program to measure time backward from date of death in order to focus on the etiological period of interest for diseases with known latency periods. Next, by examining a large group of cases of a disease for which there was a fairly strong association with certain jobs, the ETA program could be used to explore different ways of defining the latency period.

(3) Quantitative Exposures. Analytical data regarding severity and nature of solvent (as well as most other) exposures is expensive to obtain in a prospective sense and almost nonexistent retrospectively.

Efforts are now being made by OHSG industrial hygienists to reconstruct solvent usage patterns from purchasing and accounting records of the company under study. It then becomes possible to think in terms of weighting various exposures by their severity to come closer to a time-weighted cumulative dose. The ETA could be easily modified to handle this situation.

(4) Stochastic Processes. The current study ignores the sequence of jobs held, simply totaling time spent in prespecified categories. However, it may be that a certain sequence of jobs is necessary to first initiate and then promote the carcinogenic etiology of lymphatic leukemia as well as other diseases. Even more likely, certain jobs may not play a role in causing a particular disease but may show up in the statistical analysis as being associated with occurrence of the disease because of the tendency of workers in a truly hazardous job to transfer into another job. Discussions with company and union officials as well as the subjective evaluation of OHSG personnel confirm the intuitive suspicion that workers tend to follow certain job patterns, which may be influenced by social status as well as by economic incentives and seniority. One might attempt to investigate techniques such as semi-Markov or higher order Markov Processes which are capable of constructing a transition matrix incorporating the facts that the length of time in a particular job is a random variable and that there is an element of memory involved, i.e. a worker who transfers from job A to job B may be much more likely to return to job A than a worker who transfers from job C to job B.

Environmental-Epidemiological Extensions

(1) Environmental Controls. Good industrial hygiene practice is called for to minimize exposure to solvents. This is particularly the case for solvents containing benzene or other volatile components known or suspect as agents effecting blood forming organs. Until there is strong evidence that they do not represent a hazard, it should be assumed that solvents (especially aromatic solvents) are potential carcinogens and proper steps should be taken to protect the workers involved in direct handling of solvents as well as the general plant population.

(2) Toxocologic Studies. Given the present uncertainty as to the carcinogenicity of solvents used in the rubber industry, it seems reasonable to test all solvents in current use in an attempt to identify specific carcinogens. This should be done in addition to continuing and expanding the mortality studies. Because of the latency period for malignancies and the dynamic nature of chemical usage in the rubber industry, it is likely that different solvents are being used today than were used during the etiologic period of interest. However, in order to be prudent, animal (or bacteria) experimentation is suggested.

(3) Chromosomal and Enzymatic Studies. From the literature review in Chapter I as well as the early part of this chapter, it appears that individual susceptibility may play an important role in the etiology of lymphatic leukemia. If a screening test could be developed to determine such individuals, it would be of great utility in worker placement.

(4) Prescription Drug Usage. From Table 1.4 it can be seen that many drugs, in addition to chemicals, are suspect as possible etiologic agents. It would be helpful to know something about the extent to which

rubber workers use these drugs, in order to determine whether drug usage was a confounding factor. Such information is readily available from the company's drug payment program wherein the company pays all but the first dollar cost of all prescription drugs. If it were discovered that workers were using drugs suspected of causing blood dyscrasias, it might be possible to substitute other drugs not exhibiting this side effect.

LIST OF REFERENCES

Aksoy, M. et al.: "Haematological Effects of Chronic Benzene Poisoning in 217 Workers." British Journal of Industrial Medicine, 28: 296-302, 1971.

Aksoy, M. et al.: "Chronic Exposure to Benzene as a Possible Contributory Etiologic Factor in Hodgkin's Disease." Blut, 28: 293-298, 1974 (Translated).

Anderson, T. W.: An Introduction to Multivariate Statistical Analysis, New York, John Wiley and Sons, 1958, pp. 126-152.

Armitage, P.: Statistical Methods in Medical Research, New York, John Wiley and Sons, 1971, pp. 332-343.

Bean, R. H. D.: "Phenylbutazone and Leukemia: A Possible Association." British Medical Journal, I: 1552-1555, 1960.

Bizzozero, O. J., Johnson, K. G., and Ciocco, A.: "Radiation-Related Leukemia in Hiroshima and Nagasaki 1946-64." The New England Journal of Medicine, 274: 1095-1100, 1966.

Boyland, E.: "Occupational Carcinogenesis." The Practitioner, 199: 277-284, 1967.

Bradley, J. V.: Distribution-Free Statistical Tests, Englewood Cliffs, New Jersey, 1968, pp. 96-117.

Brauer, M. J. and Dameshek, W.: "Hypoplastic Anemia and Myeloblastic Leukemia Following Chloramphenicol Therapy." The New England Journal of Medicine, 227: 1003-1005, 1967.

Brown, S. M. "Leukemia and Potential Benzene Exposure." Letter to the editor Journal of Occupational Medicine, 17: 5-6, 1975.

Bruce, D. L. et al.: "Causes of Death Among Anesthesiologists: A 20-Year Survey." Anesthesiology, 29: 555-569, 1968.

Burbank, F.: "Patterns in Cancer Mortality in the United States: 1950-1967." National Cancer Institute Monograph 33, U. S. Government Printing Office, 1971.

Case, R. A. M. and Hosker, M. E.: "Tumour of the Urinary Bladder as an Occupational Disease in the Rubber Industry in England and Wales." British Journal of Preventative Social Medicine, 8: 39-50, 1954.

Case, R. A. M. et al.: "Tumours in the Urinary Bladder in Workmen Engaged in the Manufacture and Use of Certain Dyestuff Intermediates in the British Chemical Industry." British Journal of Industrial Medicine, 11: 75-104, 1954.

Clarkson, B. D. and Burchenal, J. H.: "Progress in Leukemias." Progress in Clinical Cancer, 1: 625-663, 1965.

Cobb, S., Miller, M. and Wald, N.: "On the Estimation of the Incubation Period in Malignant Disease, the Brief Exposure Case, Leukemia." Journal of Chronic Diseases, 9: 385-393, 1959.

Cochran, W. G.: Sampling Techniques, 2d ed., New York, John Wiley and Sons, 1967.

Connecticut State Department of Health: Cancer in Connecticut, 1935-62, Hartford, Connecticut, 1966.

Cornfield, J.: "A Method of Estimating Comparative Rates from Clinical Data. Applications to Cancer of Lung, Breast, Cervix." Journal of the National Cancer Institute, 11: 1269-1275, 1951.

Court Brown, W. M. and Doll, R.: Leukemia and Aplastic Anemia in Patients Irradiated for Ankylosing Spondylitis, Medical Research Council, Special Report Series, No. 295, London, Her Majesty's Stationery Office, 1957.

_____.: "Adult Leukemia--Trends in Mortality in Relation to Aetiology." British Medical Journal I: 1063-1069, 1959.

Cronkite, E. P.: "Evidence for Radiation and Chemicals as Leukemogenic Agents." Archives of Environmental Health, 3: 297-303, 1961.

Cutler, S. J.: "Report on the Third National Cancer Survery." Seventh National Cancer Conference Proceedings, American Cancer Society, 1973, pp. 639-652.

Dameshek, W. and Gunz, F.: Leukemia, New York, Grune and Stratton, 1964.

DeGowin, R.: "Benzene Exposure and Aplastic Anemia Followed by Leukemia 15 Years Later." Journal of the American Medical Association, 185: 748-751, 1963.

Dinman, B. D.: The Nature of Occupational Cancer, Springfield, Illinois, Charles C. Thomas, 1974.

Doll, R.: "The Epidemiologic Picture." Current Research in Leukemia, ed. F. G. J. Hayhoe, Cambridge at the University Press, England; 1965, pp. 280-299.

Dorland's Illustrated Medical Dictionary, 24th ed., Philadelphia, W. B. Saunders Company, 1965.

Dorn, H. F. and Cutler, S. J.: Morbidity from Cancer in the United States, Part I and Part II combined. U. S. Department of Health, Education, and Welfare, Public Health Monograph 56, Washington, D. C., Government Printing Office, 1959.

Draper, N. R. and Smith, H.: Applied Regression Analysis, New York, John Wiley and Sons, 1968, pp. 134-136.

Erdogan, G. and Aksoy, M.: "Cytogenetic Studies in Thirteen Patients with Pancytopenia and Leukemia Associated with Long-Term Exposure to Benzene." New Istanbul Contribution to Clinical Science, 10: 230-247, 1973.

Erf, L. A. and Rhoads, C. P.: "The Hematological Effects of Benzene (Benzol) Poisoning." Journal of Industrial Hygiene and Toxicology, 21: 421-435, 1939.

Erslev, A. J. and Wintrobe, M. M.: "Detection and Prevention of Drug-Induced Blood Dyscrasias." Journal of the American Medical Association, 181: 134-139, 1962.

Falconer, E. H.: "An Instance of Lymphatic Leukemia Following Benzol Poisoning." American Journal of Medical Science, 186: 353-361, 1933.

Fisher, R. A.: "The Use of Multiple Measurements in Taxonomic Problems." Annals of Eugenics, 7 (Part II): 179-188, 1936.

Forni, A. M. and Luciano, M.: "Cytogenetic Studies in a Case of Benzene Leukemia." European Journal of Cancer, 3: 251-255, 1967.

Forni, A. M. et al.: "Chromosome Changes and Their Evolution in Subjects with Past Exposure to Benzene." Archives of Environmental Health, 23: 385-391, 1971.

Fox, A. J., Lindars, D. C., and Owen, R.: "A Survey of Occupational Cancer in the Rubber and Cablemaking Industries: Results of Five-Year Analysis, 1967-71." British Journal of Industrial Medicine, 31: 140-151, 1974.

Fraumeni, J. A. and Miller, R. W.: "Epidemiology of Human Leukemia: Recent Observations." Journal of the National Cancer Institute, 38: 593-605, 1967.

Fraumeni, J. F. and Wagoner, J. K.: "Changing Sex Differentials in Leukemia." Public Health Service Reports, 79: 1093-1100, 1964.

Gamble, J. F. and Spirtas, R.: "Job Classification and Use of Complete Work Histories in Epidemiological Studies." Journal of Occupational Medicine, in press.

Greenburg, L. et al.: "Benzene (Benzol) Poisoning in the Rotogravure Printing Industry in New York City." Journal of Industrial Hygiene and Toxicology, 31: 395-420, 1939.

Gunz, F. W., Fitzgerald, P. H. and Adams, A.: "An Abnormal Chromosome in Chronic Lymphocytic Leukemia." British Medical Journal, I: 1097-1099, 1962.

Guralnick, L.: "Mortality by Occupational Level and Causes of Death Among Men 20-64 Years of Age, U.S.A.1950." Vital Statistics Special Reports, 53: 480-481, 1963.

Haenszel, W., Marcus, S. C. and Zimmerer, E. G.: Cancer Morbidity in Urban and Rural Iowa, U. S. Department of Health, Education, and Welfare, Public Health Monograph 37, Washington, D. C., Government Printing Office, 1956.

Hamilton, A.: "General Review: Benzene (Benzol) Poisoning." Archives of Pathology, 11: 434-454, 1931.

Hewitt, D.: "Geographical Pathology of Leukaemia in England and Wales." Acta, Unio Internationalis Contra Cancrum, 16: 1643-1647, 1960.

Hewitt, D.: "Some Features of Leukaemia Mortality." British Journal of Preventive Social Medicine, 9: 81-88, 1955.

Hotelling, H.: "The Generalization of Student's Ratio." Annals of Mathematical Statistics, 2: 360-378, 1931.

Hunter, D.: The Diseases of Occupations, 4th ed., Boston, Little, Brown and Company, 1969.

Ishimaru, T. et al.: "Occupational Factors in the Epidemiology of Leukemia in Hiroshima and Nagasaki." American Journal of Epidemiology, 93: 157-165, 1974.

Jedlicka, V. et al.: "Paramyeloblastic Leukaemia Appearing Simultaneously in Two Blood Cousins After Simultaneous Contact with Gammexane (Hexachorcyclohexane)." Acta Medica Scandinavica, 161: 447-451, 1958.

Jensen, M. K. and Roll, K.: "Phenylbutazone and Leukaemia." Acta Medica Scandinavica, 178: 505-513, 1965.

Kahn, Harold A.: "The Dorn Study of Smoking and Mortality Among U. S. Veterans: Report on Eight and One-Half Years of Observation." Epidemiological Approaches to the Study of Cancer and Other Chronic Diseases, ed. W. Haenszel, National Cancer Institute Monograph 19, U. S. Department of Health, Education, and Welfare, Bethesda, Maryland, 1966.

Kendall, M. G.: A Course in Multivariate Analysis, London, Charles Griffin and Company, 1961, pp. 144-170.

Kessler, I. I. and Lilienfeld, A. M.: "Perspectives in the Epidemiology of Leukemia." Advances in Cancer Research, 12: 225-302, 1969.

Kieć, E. and Kuński, H.: "A Case of Chronic Lymphocytic Leukemia Due to the Long Lasting Exposure to Benzene." Journal of the Institute of Industrial Medicine (Medycyna Pracy), 16: 362-365, 1965 (in Polish with English Summary).

Kitagawa, E. M. and Hauser, P. M.: Differential Mortality in the United States, Cambridge, Harvard University Press, 1973.

Kjeldsberg, C. R. and Ward, H. P.: "Leukemia in Arsenic Poisoning." Annals of Internal Medicine, 77: 935-937, 1972.

Kramer, C. G. and Mutchler, J. E.: "The Correlation of Clinical and Environmental Measurements for Workers Exposed to Vinyl Chloride." American Industrial Hygiene Association Journal, 33: 19-30, 1972.

Kupper, L. L., McMichael, A. J. and Spirtas, R.: "A Hybrid Epidemiologic Study Design Useful in Estimating Relative Risk." Journal of the American Statistical Association, 70: 524-528, 1975.

Lee, A. M. and Fraumeni, J. F.: "Arsenic and Respiratory Cancer in Man: An Occupational Study." Journal of the National Cancer Institute, 42: 1045-1052, 1969.

Levin, D. L. et al.: Cancer Rates and Risks, 2d ed., U. S. Department of Health, Education, and Welfare, Publication Number (NIH) 75-691, 1974.

Lilienfeld, A. M., Levin, M. L. and Kessler, I. I.: Cancer in the United States, (American Public Health Association Vital and Health Statistics Monograph), Cambridge, Harvard University Press, 1972.

Lilienfeld, A. M., Pederson, E. and Dowd, J. E.: Cancer Epidemiology: Methods of Study, Baltimore, The Johns Hopkins Press, 1967.

Lloyd, J. W. et al.: "Long-Term Mortality Study of Steelworkers: IV. Mortality by Work Area." Journal of Occupational Medicine, 12: 151-157, 1970.

Lubin, A.: "Linear and Nonlinear Discriminating Functions." British Journal of Psychology (Statistical Section), 3: 90-104, 1950.

MacMahon, B.: "Prenatal X-Ray Exposure and Childhood Cancer." Journal of the National Cancer Institute, 28: 1173-1191, 1962.

MacMahon, B. and Koller, E. K.: "Ethnic Differences in the Incidence of Leukemia." Blood, 12: 1-10, 1957.

McMichael, A. J., Andjelkovic, D. A. and Tyroler, H. A.: "Cancer Mortality Among Rubber Workers." Annals of the New York Academy of Sciences, in press.

McMichael, A. J., Spirtas, R. and Kupper, L. L.: "An Epidemiologic Study of Mortality Within a Cohort of Rubber Workers, 1964-72." Journal of Occupational Medicine, 16: 458-464, 1974.

McMichael, A. J. et al.: "Solvent Exposure and Leukemia Among Rubber Workers: An Epidemiologic Study." Journal of Occupational Medicine, 17: 234-239, 1975.

McMichael, A. J. et al.: "Mortality Among Rubber Workers: Relationship to Specific Jobs." Journal of Occupational Medicine, in press.

Mahalanobis, P. C.: "On Tests and Measures of Group Divergence." Journal of the Proceedings of the Asiatic Society of Bengal (New Series), 26: 541-588, 1930.

Mallory, T. B., Gall, E. A. and Brickley, W. J.: "Chronic Exposure to Benzene (Benzol). III. The Pathologic Results." Journal of Industrial Hygiene and Toxicology, 21: 355-377, 1939.

Mancuso, T. F., Ciocco, A. and El-Attar, A. A.: "An Epidemiological Approach to the Rubber Industry." Journal of Occupational Medicine, 10: 213-232, 1968.

Mancuso, T. F. and Brennan, M. J.: "Epidemiological Considerations of Cancer of the Gallbladder, Bile Ducts, and Salivary Glands in the Rubber Industry." Journal of Occupational Medicine, 12: 333-341, 1970.

Mantel, N.: "The Detection of Disease Clustering and a Generalized Regression Approach." Cancer Research, 27: 209-220, 1967.

March, H. C.: "Leukemia in Radiologists in a 20 Year Period." American Journal of Medical Science, 220: 282-286, 1950.

Matanoski, G. M. et al.: "The Current Mortality Rates of Radiologists and Other Physician Specialists: Specific Causes of Death." American Journal of Epidemiology, 101: 199-209, 1975.

Mather, F.: Biometrical Genetics, London, Methuen and Company, Ltd., 1949, pp. 31-37.

McKinley, S. M.: "The Design and Analysis of the Observational Study--A Review." Journal of the American Statistical Association, 70: 503-520, 1975

Miettinen, O. S.: "Standardization of Risk Ratios." American Journal of Epidemiology, 96: 383-388, 1972

_____.: "Matching and Design Efficiency in Retrospective Studies." American Journal of Epidemiology, 91: 111-118, 1970a.

_____.: "Estimation of Relative Risk from Individually Matched Series." Biometrics, 26: 75-86, 1970b.

Milham, S., Jr.: "Leukemia and Multiple Myeloma in Farmers." American Journal of Epidemiology, 94: 307-310, 1971.

Miller, R. W.: "Radiation, Chromosomes and Viruses in the Etiology of Leukemia." The New England Journal of Medicine, 271: 30-36, 1964.

_____.: "Relation Between Cancer and Congenital Defects in Man." The New England Journal of Medicine, 275: 87-93, 1966.

Monson, R. R. and Nakano, K. K.: "Mortality Among Rubber Workers." Unpublished paper presented to the Society of Epidemiologic Research, Berkeley, California, June, 1974.

Moore, C. V.: "The Leukemias." Cecil-Loeb Textbook of Medicine, 13th ed., ed. P. Beeson and W. McDermott, Philadelphia, W. B. Saunders Co., 1971, pp. 1534-1546.

Pagnotto, L. D. et al.: "Industrial Benzene Exposure from Petroleum Naptha: I. Rubber Coating Industry." American Industrial Hygiene Association Journal, 22: 417-421, 1961.

Parkes, H. G.: Health in the Rubber Industry, Manchester, England, A. Mason and Son, Ltd., 1966.

Perkins, C. G.: A Guide to the Supplementary Procedure Library for the Statistical Analysis System, Department of Statistics, North Carolina State University, 1974.

Peterson, R. D. A., Cooper, M. D. and Good, R. A.: "Disorders of Thymus and Other Lymphoid Tissues." Progress in Medical Genetics, 4: 1-31, 1965.

Polednak, A. P.: "Latency Periods in Neoplastic Diseases." American Journal of Epidemiology, 100: 354-356, 1974.

Quade, D.: "Rank Analysis of Covariance." Journal of the American Statistical Association, 62: 1187-1200, 1967.

Radhakrishna, S.: "Discrimination Analysis in Medicine." Statistician, 14: 147-167, 1964.

Redmond, C. K., Gustin, J. and Kaman, E.: "Long-Term Mortality Experience of Steelworkers. VIII. Mortality Patterns of Open Hearth Workers (A Preliminary Report)." Journal of Occupational Medicine, 17: 40-43, 1975.

Rejsek, K. and Rejsková, M.: "Long Term Observation of Chronic Benzene Poisoning." Acta Medica Scandinavica, 152: 71-78, 1955.

Sacks, M. S. and Seeman, I.: "A Statistical Study of Mortality from Leukemia." Blood, 2: 1-14, 1947.

Schrenk, H. H. et al.: "Comparative Physiological Effects of Pure, Commercial and Crude Benzenes." Journal of Industrial Hygiene and Toxicology, 22: 53-63, 1940.

Segi, M. and Kurihara, M.: Cancer Mortality for Selected Sites in 24 Countries, No. 4 (1962-1963), Sendai, Japan, Department of Public Health, Tohoku University School of Medicine, 1966.

Service, J.: A User's Guide to the Statistical Analysis System, Raleigh, North Carolina, Student Supply Stores, North Carolina State University, 1972, pp. 94-118.

Siegel, S.: Nonparametric Statistics for the Behavioral Sciences, New York, McGraw-Hill, 1956, pp. 116-127.

Silverberg, E. and Holleb, A. I.: "Cancer Statistics, 1973." Ca--A Cancer Journal for Clinicians, 23: 10, 1973.

Starmer, F. and Grizzle, J.: "A Computer Program for Analysis of Data by General Linear Models." Institute of Statistics Mimeo Series, No. 560, 1968.

Sugiyama, T., Kurita, Y. and Nishizuka, Y.: "Chromosome Abnormality in Rat Leukemia Induced by 7, 12-Dimethylbenz [a] anthracene." Science, 158: 1058-1059, 1967.

Tareeff, E. M., Kontchalovskaya, N. M. and Zorina, L. A.: "Benzene Leukemias." Acta, Unio Internat. Contra Cancrum, 19: 751-755, 1963.

Thorpe, J. J.: "Epidemiologic Survey of Leukemia in Persons Potentially Exposed to Benzene." Journal of Occupational Medicine, 16: 375-382, 1974.

Tough, I. M. and Court Brown, W. M.: "Chromosome Aberrations and Exposure to Ambient Benzene." Lancet, 288: 684, 1965.

U. S. Department of Health, Education, and Welfare: Smoking and Health (Report of the Advisory Committee to the Surgeon General of the Public Health Service), Washington, D. C., Public Health Service Publication No. 1103, 1964.

_____.: Eighth Revision: International Classification of Diseases, Adapted, Volume 1 Tabular List, Washington, D. C., Public Health Service Publication No. 1693, 1969.

_____.: Vital Statistics of the United States, 1968: Volume II. Mortality, Part A, Rockville, Md., (HSM) 72-1101, 1972a.

_____.: End Results in Cancer: Report No. 4, ed. L. M. Axtell, S. J. Cutler and M. H. Meyers, Bethesda, Md., DHEW Publication No. (NIH) 73-272, 1972b.

_____.: Progress Against Leukemia (National Cancer Institute Research Report), Bethesda, Md., DHEW Publication No. (NIH) 73-367, 1972c.

_____.: Criteria for a Recommended Standard: Occupational Exposure to Benzene, HEW Publication No. (NIOSH) 74-137, 1974.

Viadana, E. and Bross, I. D. J.: "Leukemia and Occupations." Preventive Medicine, 1: 513-521, 1972.

Vigliani, E. C. and Saita, G.: "Benzene and Leukemia." The New England Journal of Medicine, 271: 872-876, 1964.

Walter, W. A. and Gilliam, A. G.: "Leukemia Mortality. Geographic Distribution in the United States for 1949-1951." Journal of the National Cancer Institute, 17: 475-480, 1956.

```
STMT LEVEL NEST
  1                  TRANS: PROC OPTIONS(MAIN);                                                    00001000
                                                                                                   00002000
  2    1             DCL IA PICTURE'999';                                                          00003000
  3    1              DCL (Y1,Y2) PICTURE'9999';                                                   00004000
  4    1             DCL (ID1,ID2) CHAR(20) VARYING;                                               00005000
  5    1             DCL HOPE CHAR(6);                                                             00006000
  6    1             DCL REC CHAR(100) VARYING;                                                    00007000
  7    1              DCL (THIS,THAT) PICTURE'999999';                                             00008000
  8    1             DCL A(200) CHAR(100) VARYING;                                                 00009000
  9    1             DCL YEARS PICTURE'9999999', (SEV1,SEV2) PICTURE'9999999';                     00010000
 10    1             DCL IN FILE INPUT, OUT FILE OUTPUT;                                           00011000
 11    1             DCL BLANK CHAR(3) INITIAL('   ');                                             00012000
 12    1              DCL                                                                          00013000
                     (DATE1(10),DATE2(10),GOOD1,GOOD2)                                             00014000
                     PICTURE'9999999',                                                             00015000
                     (INDT(200),OUTDT(200)) PICTURE'999999',(IST(5),ILN(5)) PICTURE'99',           00016000
                     CTS CHAR(10) VARYING,                                                         00017000
                     ACC(88,10) PIC'99V999';                                                       00018000
 13    1             DCL ITOT PICTURE'999V999';                                                    00019000
 14    1             DCL                                                                           00020000
                     (FRAC1(10),FRAC2(10)) PICTURE '9V.99',                                        00021000
                     (LOW(10),HIGH(10)) PICTURE'9999',                                             00022000
                     (BIRTH_ST,ID_ST,ID_LN,YEAR1(10),YEAR2(10),DATEIN_ST,                          00023000
                     DATEOUT_ST,WAY,METHOD,CAT_ST,CAT_LN,TOT_CAT,SUB_CAT(20))                      00024000
                     PICTURE'9999',GROUP(20,10) CHAR(10) VARYING,                                  00025000
                      LAST(20) PICTURE'999V999', ID_NUM PICTURE'99';                               00026000
 15    1             DCL (IIND,ITOTAL)            DEC FIXED(7);                                    00027000
                                                                                                   00028000
 16    1             ON ENDFILE(SYSIN) BEGIN;                                                      00029000
 18    2             PUT SKIP EDIT('ERROR IN PARAMETER CARDS')(A);  GO TO FINISH2;  END;           00030000
 21    1                     ON ERROR BEGIN;                                                       00031000
 23    2                       PUT SKIP EDIT ('ERROR ',INDT,OUTDT) (A);                            00032000
 24    2                       END;                                                                00033000
 25    1             IND=0;                                                                        00034000
 26    1             TEN: GET LIST(METHOD);  OP=0;                                                 00035000
 28    1             PUT DATA(METHOD);                                                             00036000
 29    1             IF METHOD >= 1 & METHOD <= 100 THEN DO;                                       00037000
 31    1     1       IF METHOD=1 THEN CALL OT(OP); IF OP=1 THEN GO TO FINISH2;                     00038000
 35    1     1       GO TO TEN;  END;                                                              00039000
 37    1                     PUT LIST('METHOD INVALID; PROGRAM TERMINATED')SKIP;                   00040000
 38    1                     GO TO FINISH2;                                                        00041000
```

```
STMT LEVEL NEST
                                                                                00042000
                                                                                00043000
 39    1          OT: PROC(OP);                                                  00044000
                                                                                00045000
 40    2          GET LIST(KKLL);                                                00046000
 41    2          BEGIN; DCL ARRAY(KKLL) CHAR(10) VARYING; GET LIST(ARRAY);      00047000
                                                                                00048000
                  /*     ID_ST IS THE STARTING LOCATION OF THE IDENTIFYING FIELD. 00049000
                         ID_LN IS THE LENGTH OF THE IDENTIFYING FIELD.           00050000
                         CAT_ST IS THE STARTING LOCATION OF THE METHOD OF EXPOSURE00051000
                            ON THE DATA RECORDS.                                00052000
                         CAT_LN IS THE LENGTH OF THE FIELD CONTAINING THE METHOD OF00053000
                            EXPOSURE.                                           00054000
                         TOT_CAT IS THE TOTAL NUMBER OF GROUPS OF CATEGORIES OF INTEREST.00055000
                         SUB_CAT(I) IS THE NUMBER OF CATEGORIES IN GROUP(I).    00056000
                         GROUP(I,J) IS THE JTH CATEGORY IN TH ITH GROUP.   */  00057000
                                                                                00058000
 44    3          GET LIST(ID_NUM,(IST(IT),ILN(IT) DO IT=1 TO ID_NUM));          00059000
 45    3          GET LIST(CAT_ST,CAT_LN,TOT_CAT);                               00060000
 46    3          DO I=1 TO TOT_CAT;                                             00061000
 47    3    1           GET LIST(SUB_CAT(I));                                   00062000
 48    3    1     PUT DATA(SUB_CAT(I));                                         00063000
 49    3    1           DO J=1 TO SUB_CAT(I);                                   00064000
 50    3    2           GET LIST(GROUP(I,J));                                   00065000
 51    3    2     PUT DATA(GROUP(I,J));                                         00066000
 52    3    2           END; END;                                               00067000
 54    3                CALL CHECK;                                             00068000
 55    3          IF OP=1 THEN RETURN;                                          00069000
 57    3                GET LIST(DATEIN_ST,DATEOUT_ST,WAY)  ;                   00070000
 58    3           PUT DATA(DATEIN_ST,DATEOUT_ST,WAY);                          00071000
                                                                                00072000
                  /*   WAY IS THE WAY OF STRATIFYING THE DATA                   00073000
                       THERE ARE 4 WAYS TO STRATIFY THE DATA: 1) CALENDAR TIME  00074000
                       2)CHRONOLOGICAL FROM YEARS SPECIFIED 3) FRACTION OF WORK 00075000
                       HISTORIE 4) AGE CATEGORIES              */               00076000
                                                                                00077000
 59    3          IF WAY=1 THEN CALL CAL;                                       00078000
 61    3                IF WAY=2 THEN CALL CHR;  IF WAY=3 THEN CALL FRA;        00079000
 65    3          IF WAY=4 THEN CALL AGE;                                       00080000
 67    3          IF WAY=0 THEN DO; NUM_CAL=1; FLAG=0; END;                     00081000
 72    3          IF WAY=0 ! WAY=1! WAY=2! WAY=3! WAY=4 THEN GO TO TTEN;        00082000
 74    3                PUT LIST('INVALID WAY OF STRATIFYING DATA.PROGRAM       00083000
 75    3                WILL TERMINATE')SKIP;  GO TO FINISH;
```

```
STMT LEVEL NEST

 76     3          TTEN:  READ FILE(IN) INTO(REC);                                    00084000
 77     3            IIND=0;        /*  TOTAL # OF INDIVIDUALS  */                   00085000
 78     3          ITOTAL=1;  /* TOTAL # OF RECORDS */                               00086000
 79     3          TWEN:  ACC=0;                                                      00087000
                                                                                     00088000
                   /*  ACC IS WHERE ACCUMULATED TIME IS STORED.                       00089000
                   */                                                                00090000
                                                                                     00091000
 80     3          ID1=''; DO IT=1 TO ID_NUM; ID1=ID1||SUBSTR(REC,IST(IT),ILN(IT)); END; 00092000
 84     3                 A(1)=REC;                                                   00093000
 85     3                 IK=1;                                                       00094000
 86     3          IIND=IIND+1;                                                       00095000
                                                                                     00096000
                   /*   A WILL STORE ALL THE WORK HISTORIES FOR 1 PERSON AT A TIME.  00097000
                        IK WILL COUNT THE NUMBER OF WORK HISTORIES                   00098000
                   */                                                                00099000
                                                                                     00100000
 87     3               ON ENDFILE(IN) GO TO TONE;                                    00101000
 89     3            SOOP: READ FILE(IN) INTO(REC);                                   00102000
 90     3          ID2=''; DO IT=1 TO ID_NUM; ID2=ID2||SUBSTR(REC,IST(IT),ILN(IT)); END; 00103000
 94     3          ITOTAL=ITOTAL+1;                                                   00104000
 95     3                 IF ID1 ¬=ID2 THEN GO TO THIRTY;                             00105000
 97     3                 IK=IK+1;                                                    00106000
 98     3                 A(IK)=REC;                                                  00107000
 99     3                 GO TO SOOP;                                                 00108000
100     3            TONE: IND=1;                                                     00109000
                                                                                     00110000
                   /*   IND IS AN INDICATOR FOR END OF FILE                          00111000
                   */                                                                00112000
                   /*   BEGIN PROCESSING ONE INDIVIDUALS SET OF WORK HISTORIES.      00113000
                   */                                                                00114000
                                                                                     00115000
                                                                                     00116000
```

```
STMT LEVEL NEST

                                                                                00117000
                        /*  BEGIN LOOP1  */                                     00118000
                                                                                00119000
101     3             THIRTY:  DO LB=1 TO NUM_CAL;                              00120000
                                                                                00121000
                        /*  BEGIN LOOP2  */                                     00122000
102     3     1       DO LA=1 TO IK;                                            00123000
103     3     2       JKL=0;                                                    00124000
104     3     2       CALL NUMBER(SUBSTR(A(LA),DATEIN_ST,6));                   00125000
105     3     2       IF JKL=1 THEN GO TO SIXTYONE;                             00126000
107     3     2       INDT(LA)=SUBSTR(A(LA),DATEIN_ST,6);                       00127000
108     3     2       CALL CHANGE(INDT(LA));                                    00128000
109     3     2       JKL=0;                                                    00129000
110     3     2       CALL NUMBER(SUBSTR(A(LA),DATEOUT_ST,6));                  00130000
111     3     2       IF JKL=1 THEN GO TO SIXTYONE;                             00131000
113     3     2        OUTDT(LA)=SUBSTR(A(LA),DATEOUT_ST,6);                    00132000
114     3     2        CALL CHANGE(OUTDT(LA));                                  00133000
                                                                                00134000
115     3     2        IF FLAG=0 THEN DO;                                       00135000
117     3     3            DATE1(LB)=INDT(LA); DATE2(LB)=OUTDT(LA);             00136000
119     3     3            GO TO FORTY;  END;                                   00137000

                                                                                00138000
121     3     2        IF FLAG=2 THEN DO;                                       00139000
123     3     3        THIS=SUBSTR(A(1),DATEIN_ST,6);  CALL CHANGE(THIS);       00140000
125     3     3        HOPE=THIS;                                               00141000
126     3     3        DATE1(LB)=ADDYEAR(HOPE,YEAR1(LB));                       00142000
127     3     3        DATE2(LB)=ADDYEAR(HOPE,YEAR2(LB));                       00143000
128     3     3        IF DATE1(LB) > 1000000 THEN DATE1(LB)=DATE1(LB)-1000000; 00144000
130     3     3        IF DATE2(LB) > 1000000 THEN DATE2(LB)=DATE2(LB)-1000000; 00145000
132     3     3        IF YEAR2(LB)=99 THEN DO;                                 00146000
134     3     4        JKL=0;                                                   00147000
135     3     4        CALL NUMBER(SUBSTR(A(IK),DATEOUT_ST,6));                 00148000
136     3     4        IF JKL=1 THEN GO TO SIXTYONE;                            00149000
138     3     4        THIS=SUBSTR(A(IK),DATEOUT_ST,6);  CALL CHANGE(THIS);     00150000
140     3     4        HOPE=THIS;                                               00151000
141     3     4        DATE2(LB)=HOPE;                                          00152000
142     3     4        GO TO FORTY;  END;                                       00153000
144     3     3        GO TO FORTY;  END;                                       00154000
```

```
STMT LEVEL NEST·

                                                                                00155000
146   3    2      IF FLAG=3 THEN DO;                                            00156000
148   3    3      JKL=0;                                                        00157000
149   3    3      CALL NUMBER(SUBSTR(A(1),DATEIN_ST,6));                        00158000
150   3    3      IF JKL=1 THEN GO TO SIXTYONE;                                 00159000
152   3    3      JKL=0;                                                        00160000
153   3    3      CALL NUMBER(SUBSTR(A(IK),DATEOUT_ST,6));                      00161000
154   3    3      IF JKL=1 THEN GO TO SIXTYONE;                                 00162000
156   3    3      THIS=SUBSTR(A(1),DATEIN_ST,6);  CALL CHANGE(THIS);  SEV1=THIS; 00163000
159   3    3      THIS=SUBSTR(A(IK),DATEOUT_ST,6); CALL CHANGE(THIS);  SEV2=THIS; 00164000
162   3    3      CALL TIME(SEV1,SEV2);  YEARS=1DAY/365;                        00165000
164   3    3      Y1=YEARS*FRAC1(LB);                                           00166000
165   3    3      Y2=YEARS*FRAC2(LB);                                           00167000
166   3    3      THIS=SUBSTR(A(1),DATEIN_ST,6);  CALL CHANGE(THIS);            00168000
168   3    3      HOPE=THIS;                                                    00169000
169   3    3      DATE1(LB)=ADDYEAR(HOPE,Y1);                                   00170000
170   3    3      DATE2(LB)=ADDYEAR(HOPE,Y2);                                   00171000
171   3    3      IF DATE1(LB) > 1000000 THEN DATE1(LB)=DATE1(LB)-1000000;      00172000
173   3    3      IF DATE2(LB) > 1000000 THEN DATE2(LB)=DATE2(LB)-1000000;      00173000
175   3    3      GO TO FORTY;  END;                                            00174000
                                                                                00175000
177   3    2      IF FLAG=4 THEN DO;                                            00176000
179   3    3      THIS=SUBSTR(A(LA),BIRTH_ST,6);  CALL CHANGE(THIS);            00177000
181   3    3      HOPE=THIS;                                                    00178000
182   3    3      DATE1(LB)=ADDYEAR(HOPE,LOW(LB));                              00179000
183   3    3      DATE2(LB)=ADDYEAR(HOPE,HIGH(LB));                             00180000
184   3    3      IF DATE1(LB) > 1000000 THEN DATE1(LB)=DATE1(LB)-1000000;      00181000
186   3    3      IF DATE2(LB) > 1000000 THEN DATE2(LB)=DATE2(LB)-1000000;      00182000
188   3    3      GO TO FORTY;  END;                                            00183000
```

```
STMT LEVEL NEST                                                                 00184000
                        /*   THIS SECTION CALCULATES THE TWO DATES BETWEEN WHICH TIME   00185000
                             SHOULD BE ACCUMULATED  */                          00186000
                                                                                00187000
   190      3    2      FORTY:                                                  00188000
   191      3    2              IF (DATE1(LB)<=INDT(LA) & DATE2(LB)>=OUTDT(LA)) THEN DO;  00189000
   192      3    3      GOOD1=INDT(LA);  GOOD2=OUTDT(LA);                       00190000
   194      3    3      GO TO SIXTY;  END;                                      00191000
   196      3    2      IF (DATE1(LB)>=OUTDT(LA) &DATE2(LB)>OUTDT(LA)) THEN GO TO SIXTYONE;  00192000
   198      3    2      IF (INDT(LA)<=DATE1(LB)&DATE1(LB)<=OUTDT(LA)&OUTDT(LA)<=DATE2(LB))  00193000
   199      3    2       THEN DO;                                               00194000
   200      3    3      GO TO SIXTY;  END;                                      00195000
   202      3    2      GOOD1=DATE1(LB);  GOOD2=OUTDT(LA);                      00196000
   204      3    2      IF (DATE1(LB)<=INDT(LA)&INDT(LA)<DATE2(LB)&DATE2(LB)<=OUTDT(LA))  00197000
   205      3    2       THEN DO;                                               00198000
   206      3    3      GOOD1=INDT(LA);  GOOD2=DATE2(LB);                       00199000
   208      3    3      GO TO SIXTY;  END;                                      00200000
   210      3    2      IF (DATE1(LB)<DATE2(LB)&DATE2(LB)<=INDT(LA)&INDT(LA)<OUTDT(LA))  00201000
   211      3    2      THEN GO TO SIXTYTWO;                                    00202000
   212      3    2      IF(INDT(LA)<=DATE1(LB)&DATE1(LB)<DATE2(LB)&DATE2(LB)<=OUTDT(LA))  00203000
   213      3    2       THEN DO;                                               00204000
   214      3    3      GOOD1=DATE1(LB);  GOOD2=DATE2(LB);                      00205000
   216      3    3    GO TO SIXTY;  END;                                        00206000
   218      3    2      GO TO SIXTYONE;                                         00207000
                                                                                00208000
                        /*  BEGIN ACCUMULATION                                  00209000
                         */                                                     00210000
                                                                                00211000
   219      3    2      SIXTY:                                                  00212000
                              CALL TIME(GOOD1,GOOD2);                           00213000
   220      3    2      OTS=SUBSTR(A(LA),CAT_ST,CAT_LN);                        00214000
   221      3    2      DO KIJ=1 TO KKLL;                                       00215000
   222      3    3      IF (OTS=ARRAY(KIJ)) THEN DO;                            00216000
   224      3    4      ACC(KIJ,LB)=IDAY/365+ACC(KIJ,LB);                       00217000
   225      3    4      GO TO SIXTYONE;  END;                                   00218000
   227      3    3      END;                                                    00219000
   228      3    2      PUT SKIP LIST('THE FOLLOWING RECORD HAS AN INVALID OT');  00220000
   229      3    2      PUT SKIP LIST(A(LA));                                   00221000
                         /*   END OF LOOP2, LOOP OVER WORK HISTORIES  */       00222000
   230      3    2      SIXTYONE:END;                                           00223000
                         /*   END OF LOOP1, LOOP FOR DIFFERENT # OF SLICES  */ 00224000
   231      3    1      SIXTYTWO:  END;                                         00225000
                        /*                                                      
```

```
STMT LEVEL NEST

                                                                                00226000
                        READY TO PRINT                                          00227000
                                                                                00228000
                       */                                                       00229000
 232     3               CALL PRINT;                                            00230000
 233     3                IF IND = 1 THEN GOTO FINISH;                          00231000
 235     3                GO TO TWEN;                                           00232000

                                                                                00233000
                        /*  PROC CHECK CHECKS FOR INVALID OT'S ENTERED BY USER */ 00234000
 236     3              CHECK:  PROC;                                            00235000
 237     4              DO I=1 TO TOT_CAT; DO J=1 TO SUB_CAT(I); DO K=1 TO KKLL;  00236000
 240     4    3         IF GROUP(I,J)=ARRAY(K) THEN GO TO ENDJ; ENDK; END;       00237000
 243     4    2         PUT SKIP LIST('USER HAS ENTERED AN INVALID CATEGORY',GROUP(I,J),  00238000
 244     4    2         'PROGRAM WILL TERMINATE');  OP=1; RETURN;                00239000
 246     4    2         ENDJ: END;  END;  RETURN;  END CHECK;                    00240000

                                                                                00241000
                        /*  PROC CAL IS CALLED IF WAY=1  */                     00242000
 250     3               CAL: PROC;                                              00243000
 251     4                    GET LIST(NUM_CAL);                                 00244000
 252     4               PUT DATA(NUM_CAL);                                      00245000
 253     4                    DO I=1 TO NUM_CAL;                                 00246000
 254     4    1         GET LIST(THIS,THAT);                                    00247000
 255     4    1         CALL CHANGE(THIS);  CALL CHANGE(THAT);                  00248000
 257     4    1         DATE1(I)=THIS;  DATE2(I)=THAT;                          00249000
 259     4    1         PUT SKIP LIST(DATE1(I),DATE2(I));                       00250000
 260     4    1              END;                                               00251000
 261     4                   FLAG=1;                                            00252000
 262     4                   RETURN;                                            00253000
 263     4                   END CAL;                                           00254000

                                                                                00255000
                        /*   PROC CHR IS CALLED IF WAY=2  */                    00256000
 264     3               CHR: PROCEDURE;                                         00257000
 265     4                    GET LIST(NUM_CAL);                                 00258000
 266     4               PUT SKIP LIST(NUM_CAL);                                 00259000
 267     4                    DO I=1 TO NUM_CAL;                                 00260000
 268     4    1               GET LIST(YEAR1(I),YEAR2(I));                       00261000
 269     4    1         PUT DATA(YEAR1(I),YEAR2(I));                            00262000
 270     4    1              END;                                               00263000
 271     4                   FLAG=2;                                            00264000
 272     4                   RETURN;                                            00265000
 273     4                   END CHR;                                           00266000
```

```
STMT LEVEL NEST

                                                                                00267000
                    /*  PROC FRA IS CALLED IF WAY=3  */                         00268000
274     3             FRA: PROC;                                                00269000
275     4           GET LIST(NUM_CAL);                                          00270000
276     4           PUT DATA(NUM_CAL);                                          00271000
277     4           DO I=1 TO NUM_CAL;                                          00272000
278     4   1             GET LIST(FRAC1(I),FRAC2(I));                          00273000
279     4   1       PUT DATA(FRAC1(I),FRAC2(I));                                00274000
280     4   1             IF FRAC1(I)>1 | FRAC2(I)>1 THEN DO;                   00275000
282     4   2             PUT LIST('INVALID INPUT FOR FRACTION SUBROUTINE')SKIP; 00276000
283     4   2             GO TO FINISH; END;                                    00277000
285     4   1             END;                                                  00278000
286     4                 FLAG=3;  RETURN; END FRA;                             00279000

                                                                                00280000
                    /*   PROC AGE IS CALLED IF WAY=4  */                        00281000
289     3             AGE: PROC;                                                00282000
290     4                 GET LIST(NUM_CAL);                                    00283000
291     4           PUT DATA(NUM_CAL);                                          00284000
292     4                 DO I= 1 TO NUM_CAL;                                   00285000
293     4   1             GET LIST(LOW(I),HIGH(I));                             00286000
294     4   1       PUT DATA(LOW(I),HIGH(I));                                   00287000
295     4   1             END;                                                  00288000
296     4                 GET LIST(BIRTH_ST);                                   00289000
297     4           PUT DATA(BIRTH_ST);                                         00290000
298     4                 FLAG=4;                                               00291000
299     4                 RETURN;                                               00292000
300     4                 END AGE;                                              00293000
                    /*   PROC CHANGE IS CALLED TO CONVERT DATES FROM MONTH,DAY,YEAR 00294000
                         TO YEAR, MONTH, DAY  */                                00295000
301     3           CHANGE:  PROC(A);                                           00296000
302     4           DCL AA PICTURE'999999';                                     00297000
303     4           DCL A PICTURE'999999';                                      00298000
304     4           DCL YR PICTURE'999',(MO,DA) PICTURE'99',R PICTURE'9999';    00299000
305     4            MO=A/10000; R=MOD(A,10000);  DA=R/100;  YR=MOD(R,100);     00300000
309     4           A=YR*10000+MO*100+DA;                                       00301000
310     4           RETURN;  END CHANGE;                                        00302000
                                                                                00303000
```

```
STMT LEVEL NEST

                        /*   PROC PRINT ADDS UP TIME SPENT IN EACH OF GROUP AND PRINTS      00304000
                             RESULTS.  */                                                  00305000
312     3           PRINT: PROC;                                                          00306000
313     4           DO IA=1 TO NUM_CAL;                                                   00307000
314     4     1     LAST=0;                                                               00308000
315     4     1     ITOT=0;                                                               00309000
316     4     1     DO IB=1 TO KKLL;                                                      00310000
317     4     2     ITOT=ITOT+ACC(IB,IA);                                                 00311000
318     4     2     END;                                                                  00312000
319     4     1       DO IC=1 TO TOT_CAT;                                                 00313000
320     4     2     DO ID=1 TO SUB_CAT(IC);                                               00314000
321     4     3     DO IB=1 TO KKLL;                                                      00315000
322     4     4     IF (ARRAY(IB)=GROUP(IC,ID)) THEN DO;                                  00316000
324     4     5     LAST(IC)=LAST(IC)+ACC(IB,IA);                                         00317000
325     4     5     GO TO LOOPD;                                                          00318000
326     4     5      END;                                                                 00319000
327     4     4     END;                                                                  00320000
328     4     3     LOOPD:   END;                                                         00321000
329     4     2     LOOPC:  END;                                                          00322000
330     4     1     IF GROUP(TOT_CAT,SUB_CAT(TOT_CAT))='ALL ' THEN DO;                    00323000
332     4     2     LAST(TOT_CAT)=ITOT;                                                   00324000
333     4     2     TOT_CATT=TOT_CAT-1;                                                   00325000
334     4     2     DO IC=1 TO TOT_CATT;                                                  00326000
335     4     3     LAST(TOT_CAT)=LAST(TOT_CAT)-LAST(IC);                                 00327000
336     4     3     END;  END;                                                            00328000
338     4     1     PUT FILE(OUT) EDIT(IA,ID1,(LAST(IC) DO IC=1 TO TOT_CAT))              00329000
                    (SKIP,F(2),X(1),A(20),20(F(6,3),X(1)));                             00330000
339     4     1   LOOPA:  END;                                                          00331000
340     4           RETURN;                                                             00332000

341     4           END PRINT;                                                          00333000
```

```
START LEVEL TEST

                                                                                00334000
                        /*   PROC NUMBER CHECKS EACH DATE TO BE SURE IT IS A NUMBER.    00335000
                             A MESSAGE IS PRINTED AND THE LINE OF WORK HISTORY SKIPPED   00336000
                             IF A DATE IS ALL NINES OR BLANKS OR A MONTH OR DAY IS NINES 00337000
                             OR IF INVALID CHARACTER APPEARS.  */                       00338000
                                                                                00339000
342     3           NUMBER:  PROC(VAR);                                         00340000
343     4           DCL VAR CHAR(6);                                            00341000
344     4           IF SUBSTR(VAR,1,2)='99' | SUBSTR(VAR,3,2)='99' THEN DO;     00342000
346     4     1       PUT SKIP LIST(' A RECORD HAS 9S IN A DAY OR MONTH',A(LA)); 00343000
347     4     1         JRL=1;  RETURN;  END;                                     00344000
350     4           NUM=VERIFY(VAR,'9 ');  IF NUM=0 THEN DO;                    00345000
353     4     1     PUT SKIP LIST('A DATE HAS BLANKSOR 9S IN THE FOLLOWING RECORD',A(LA));00346000
354     4     1     JRL=1;  RETURN;  END;                                       00347000
357     4           NUMB=VERIFY(VAR,'0123456789');                              00348000
358     4           IF NUMB¬=0 THEN DO;                                         00349000
360     4     1     PUT SKIP LIST(' THE FOLLOWING RECORD HAS AN INVALID CHARACTER IN A 00350000
                    DATE FIELD ',A(LA));                                        00351000
361     4     1     JRL=1;                                                      00352000
362     4     1       END;                                                      00353000
363     4           RETURN;  END NUMBER;                                        00354000
                                                                                00355000
                        /*   PROC ADDYEAR WILL ADD A NUMBER OF YEARS TO A DATE  */ 00356000
365     3         ADDYEAR:  PROC(X,Y);                                          00357000
366     4           DCL X CHAR(6),XX PICTURE'999999' DEFINED X, VALUE PICTURE'9999999', 00358000
                    Y  PICTURE'9999',                                           00359000
                    YY PICTURE'99';                                             00360000
367     4             VALUE=XX+Y*10000;                                         00361000
368     4           RETURN(VALUE);  END ADDYEAR;
```

```
STMT LEVEL NEST

                        /*   PROC TIME CALCULATES THE NUMBER OF DAYS BETWEEN TWO GIVEN      00362000
                             DATES  */                                                      00363000
370     3           TIME:  PROC(D1,D2);                                                     00364000
371     4          DCL (D1,D2) PICTURE'9999999',R1 PICTURE'9999',                           00365000
                   (MO1,DA1,DA2,MO2) PICTURE'99',                                          00366000
                   (YR1,YR2) PICTURE'999',                                                 00367000
                   COUNT(12) FIXED(2) INIT(31,28,31,30,31,30,31,31,30,31,30,31);           00368000
372     4           YR1=D1/10000; R1=MOD(D1,10000);                                        00369000
374     4          MO1=R1/100;  DA1=MOD(R1,100);                                           00370000
376     4          YR2=D2/10000;  R1=MOD(D2,10000);                                        00371000
378     4          MO2=R1/100;  DA2=MOD(R1,100);                                           00372000
380     4          IDAY=0;                                                                 00373000
381     4          LOOP:  DO I=YR1 TO YR2;                                                 00374000
382     4   1      ITOP=MOD(I,100);                                                        00375000
383     4   1     IREMAIN=MOD(ITOP,10);                                                    00376000
384     4   1      IYR=ITOP/10;                                                           00377000
385     4   1      IZ=MOD(IYR,2);                                                         00378000
386     4   1      IF (IZ=0&(IREMAIN=0|IREMAIN=4|IREMAIN=8))|                             00379000
387     4   1      (IZ¬=0&(IREMAIN=2|IREMAIN=6)) THEN IADD=1;                             00380000
388     4   1      ELSE IADD=0;                                                           00381000
389     4   1      IF YR1=YR2 THEN DO;                                                    00382000
391     4   2      IF MO1=MO2 THEN DO;                                                    00383000
393     4   3      IDAY=IDAY+(DA2-DA1);                                                   00384000
394     4   3      GO TO BET;  END;                                                       00385000
396     4   2      DO J=MO1 TO MO2;                                                       00386000
397     4   3      IF J=2 THEN COUNT(J)=COUNT(J)+IADD;                                    00387000
399     4   3      IF J=MO1 THEN IDAY=IDAY+COUNT(J)-DA1;                                  00388000
401     4   3      ELSE IF J=MO2 THEN IDAY=IDAY+DA2;                                      00389000
403     4   3      ELSE IDAY=IDAY+COUNT(J);                                               00390000
404     4   3      END;                                                                   00391000
405     4   2      GO TO BET;                                                             00392000
406     4   2      END;                                                                   00393000
                                                                                          00394000
```

```
STMT LEVEL NEST                                                                  00395000
407     4     1     IF I=YR1 THEN DO;                                            00396000
409     4     2     DO J=MO1 TO 12;                                              00397000
411     4     3     IF J=2 THEN COUNT(J)=COUNT(J)+IADD;                          00398000
413     4     3     IF J=MO1 THEN IDAY=IDAY+COUNT(J)-DA1;                        00399000
414     4     3     ELSE IDAY=IDAY+COUNT(J);                                     00400000
415     4     3     END;                                                         00401000
416     4     2     END;                                                         00402000
417     4     1     ELSE IF I=YR2 THEN DO;                                       00403000
419     4     2     DO J=1 TO MO2;                                               00404000
420     4     3     IF J=2 THEN COUNT(J)=COUNT(J)+IADD;                          00405000
422     4     3     IF J=MO2 THEN IDAY=IDAY+DA2;                                 00406000
424     4     3     ELSE IDAY=IDAY+COUNT(J);                                     00407000
425     4     3     END;                                                         00408000
426     4     2     END;                                                         00409000
427     4     1     ELSE IDAY=IDAY+365+IADD;                                     00410000
428     4     1     BET:  COUNT(2)=28;                                           00411000
429     4     1     END LOOP;                                                    00412000
430     4           RETURN; END TIME;                                            00413000
432     3             END;                                                       00414000
433     2           FINISH:  PUT SKIP LIST(' ALL THE RECORDS HAVE BEEN PROCESSED');00415000
434     2           PUT EDIT('TOTAL # OF INDIVIDUALS',IIND)(SKIP,A,F(5));        00416000
435     2            PUT EDIT('TOTAL # OF RECORDS',ITOTAL)(SKIP,A,F(10));        00417000
436     2             OP=1;                                                      00418000
437     2           RETURN;  END OT;                                             00419000
                       /*  END OF MAIN PROGRAM    */                           00420000
439     1           FINISH2:  END TRANS;                                         00421000
```