560590006

Pattern Recognition Analysis of VA/EPA PCDD and PCDF Data

Final Report

by

Karin M. Bauer and John S. Stanley

Midwest Research Institute
425 Volker Boulevard
Kansas City, MO 64110

Janet Remmers, Work Assignment Manager
John Schwemberger, Work Assignment Manager
Dr. Joseph J. Breen, Project Officer

U.S. Environmental Protection Agency
Office of Toxic Substances
Exposure Evaluation Division
401 M Street, SW
Washington, D.C. 20460

EPA Contract No. 68-02-4252
MRI Project No. 8863-A
Work Assignment No. 25

September 17, 1990

U.S. Environmental Protection Agency
Region 5, Library (PL-12J)
77 West Jackson Boulevard, 12th Floor
Chicago, IL 60604-3590

DISCLAIMER

This document has been reviewed and approved for publication by the Office of Environmental Epidemiology, Department of Veterans Affairs (VA), and the Office of Pesticides and Toxic Substances, U.S. Environmental Protection Agency (EPA). The use of trade names or commercial products does not constitute VA or EPA endorsement or recommendation for use.

PREFACE

This report details the results from a multivariate statistical analysis of the PCDD and PCDF data from the collaborative effort between the Department of Veterans Affairs (VA), Office of Environmental Epidemiology, and the U.S. Environmental Protection Agency, Office of Toxic Substances. The data base was provided by Midwest Research Institute (MRI) through EPA Contract No. 68-02-4252, Work Assignment No. 24. The chemical analyses of the human adipose tissue samples were conducted by MRI under the VA and EPA intergovernmental agency agreement (VA Reference No. V101(91)-P86009, EPA Reference No. RW36930735-01-3).

The statistical analyses were performed by MRI through the above-mentioned intergovernmental agency agreement under the direction of Ms. Karin M. Bauer, Principal Statistician, for EPA's Office of Toxic Substances, Exposure Evaluation Division (Contract No. 68-02-4252, Work Assignment No. 25). Throughout the project, Dr. John S. Stanley, Head, Analytical Chemistry Section, provided guidance and consulting on chemical matters. Ms. Jean E. Pelkey, Associate Statistician, assisted in generating figures and tables for this report.

The assistance provided by the EPA Work Assignment Managers, Ms. Janet Remmers, Mr. Bradley Schultz, and Mr. John Schwemberger, as well as the additional demographic and military/civilian status information on the individuals provided by Dr. Han Kang and Mr. Kevin Watanabe of the VA, are gratefully acknowledged.

MIDWEST RESEARCH INSTITUTE

Karin M. Bauer
Principal Statistician
Work Assignment Leader

Approved:

Paul C. Constant
Program Manager

September 17, 1990

EXECUTIVE SUMMARY

The objective of this analysis was to evaluate, by pattern recognition techniques, relationships of chlorinated dibenzo-*p*-dioxin and dibenzofuran levels in human adipose tissue among the analytes themselves and with respect to demographic variables (military/civilian status, age, collection year, race, geographic region). The conclusions of this study provide a multivariate context for the data and strengthen the conclusions obtained in the univariate study (on 2,3,7,8-TCDD) conducted by the Department of Veterans Affairs (VA) and the U.S. Environmental Protection Agency (EPA). The result of most interest, based on a set of matched NHATS specimens (36 Vietnam-veterans, 79 non-Vietnam veterans, and 80 civilians), is that Vietnam veterans in this study do not exhibit higher levels of 2,3,7,8-TCDD, PCDDs, and PCDFs than the other two groups. Also, cluster analysis of the PCDDs and PCDFs demonstrates that 2,3,7,8-TCDD is not strongly correlated with any of the other compounds, indicating possible differences in exposure routes to these chemicals.

The primary focus of the VA/EPA collaborative study was on 2,3,7,8-tetrachlorodibenzo-*p*-dioxin (TCDD), a contaminant of the herbicide Agent Orange used in Vietnam. However, the analytical protocol used in that study provided data on all 17 of the 2,3,7,8-substituted chlorinated dioxin and furan congeners. These chemical data, together with the available demographic and military service status information, were the basis of the present study. Pattern recognition techniques (cluster and principal component analyses) were used to evaluate relationships and their significance within the data. The intent of this study was not to duplicate the previous analysis efforts (VA and USEPA, 1989), but rather to complement them by considering all PCDD and PCDF residue levels and demographic information simultaneously.

The tissue samples considered in this study were obtained from a large, archived set of specimens obtained through the National Human Adipose Tissue Survey (NHATS). The availability of military service status on males born between 1936 and 1954 was the basis for the selection of specimens in the three groups--Vietnam veterans, non-Vietnam veterans, and civilians--considered in this study. To the extent that the NHATS population is not representative of Vietnam veterans, the results of this study are not representative of the population of Vietnam veterans. The relatively small sample size (36) of Vietnam veterans would limit the significance of results relating to TCDD levels in the population of Vietnam veterans in any case. However, within the restricted NHATS population, the comparative results are valid.

The most important outcome from the analyses performed in the present study is that there is no visible stratification of the NHATS specimens in groups as defined by their military service or civilian status with respect to either the PCDD and PCDF levels or the demographic variables.

From the univariate analysis results obtained during the current study, which parallel those obtained by the VA, the following general remarks can be highlighted.

- The levels of 2,3,7,8-TCDD are, for all practical purposes, independent of the levels of all other analytes quantified in this set of NHATS specimens. As subsequently confirmed by the multivariate analyses results, 2,3,7,8-TCDD behaves unlike any of the other analytes; that is, its level in adipose tissue cannot be predicted accurately by the levels of any other analyte or analyte combinations.

- The levels of 2,3,7,8-TCDD decrease with successive collection years, at the rate of 0.63 pg/g per year of collection. Due to the very low predictive power of that variable, however, the general downward trend of 2,3,7,8-TCDD levels cannot be predicted with any degree of certainty by the collection year.

- Based on year of birth, the levels of 2,3,7,8-TCDD decrease with successive year, at the rate of 0.04 pg/g per year. Again, due to the very low power of this variable in predicting the concentration levels in adipose tissue, no strong conclusion can be drawn as to the effect of this variable on 2,3,7,8-TCDD levels.

The most interesting chemical results were obtained through multivariate statistical analyses, which were the focus of the present study. The objective of cluster analysis, performed on the log-transformed concentration levels of the analytes, was to investigate interrelationships of the PCDD and PCDF analytes. The following points highlight the cluster analysis results:

- 2,3,7,8-TCDD does not group with any of the other analytes--its behavior is unlike the behavior of the others.

- Apart from 2,3,7,8-TCDD and those analytes with relatively high frequencies of nondetects, the levels of the higher chlorinated PCDDs and PCDFs are strongly interrelated. All correlations are high and positive, that is, the levels of these analytes are proportional to each other.

Principal component analysis focuses on representing data in a space with preferably no more than three dimensions. These dimensions are defined by linear combinations (principal components) of the variables included in the analysis. The objective of principal component analysis is to detect a structure in the data base, if it exists, that could possibly distinguish between meaningful groups of specimens. This is accomplished by using all data related to the chemical (the concentration levels were log-transformed), the demographic variables, and the military service/civilian status. After applying this technique to all the data available in this data base, the following conclusions can be drawn:

- None of the analytes, demographic variables, or military service/civilian status information provided any means of classifying the NHATS specimens into more than one large group. That is, no pattern-distinguishing--for example, between Vietnam veterans, non-Vietnam veterans, and civilians--was visible on a plot of the specimen scores along the most important principal components.

- A large proportion (77%) of the variability in the total data set, as determined by the values of the NHATS specimens on all the variables, is explained by the combination of two principal components. The first and predominant principal component (explaining 64% of the total variability) is a combination of the higher chlorinated PCDDs and PCDFs, each analyte having about equal contribution to this principal component. This is consistent with the results of the cluster analysis. The second, and by far less important principal component (13% of the total variability explained), consists mainly of age of specimen at time of death or surgery and specimen collection year.

The results of these additional statistical analysis efforts provide evidence of interrelationships of PCDDs and PCDFs that could be possible indicators of sources of exposure. That is, the source of 2,3,7,8-TCDD may be completely different from those of the other PCDDs and PCDFs. Further statistical evaluations of body burden data may require more comprehensive data bases that identify body burden half-life of the higher chlorinated PCDDs and PCDFs vs. 2,3,7,8-TCDD, as well as the potential contributions arising from commercial products and environmental matrices (e.g., incinerator emissions, ambient air, water, fish).

In summary, the plots of the two most important principal components, the "Higher Chlorinated PCDDs and PCDFs" and the "Age Plus Collection Year" components, explain about 80% of the total variation in the data set. However, these plots do not reveal any clear stratification of the specimens into groups, whether defined by military service/civilian status of the individual, geographic region, race, or age. This study indicates, based on the set of matched NHATS specimens, that the group of 36 Vietnam veterans in this study does not exhibit higher levels of PCDDs and PCDFs, including 2,3,7,8-TCDD, than either the group of 79 non-Vietnam veterans or the control group of 80 civilians.

TABLE OF CONTENTS

		Page
Preface		iii
Executive Summary		v
List of Figures		x
List of Tables		xi
I.	Introduction	1
	A. Project Background	1
	B. Study Objectives	2
II.	Data Base Description	3
	A. Chemical Data Base	3
	B. Demographic Data Base	4
	C. Military/Civilian Status Data Base	4
III.	Descriptive Statistics	5
	A. Descriptive Statistics of the Chemical Data	5
	B. Descriptive Statistics of the Demographic Data by Military/Civilian Status	10
	C. Relationships Between 2,3,7,8-TCDD and Selected Demographic Variables	13
IV.	Multivariate Statistical Methods	17
	A. Cluster Analysis of Variables	17
	B. Principal Component Analysis	18
V.	Multivariate Analysis Results and Discussion	20
	A. Cluster Analysis of Analytes	20
	B. Principal Component Analysis	25
VI.	References	43
Appendix - Histograms of Concentration Levels of All 16 Analytes		A-1

LIST OF FIGURES

Number	Title	Page
1	Plot of 2,3,7,8-TCDD vs. age of individual at time of collection	15
2	Plot of 2,3,7,8-TCDD concentrations vs. collection year	15
3	Plot of 2,3,7,8-TCDD concentrations vs. birth year of individual	16
4	Plot of 2,3,7,8-TCDD concentrations vs. body mass index	16
5	Tree diagram showing clustering process of all 16 analytes based on the correlations of the log-transformed concentrations	22
6	Cluster analysis results of all 16 analytes based on the correlations of the log-transformed concentrations	26
7	Plot of principal component 2 vs. 1--all 195 specimens coded by military/civilian status	33
8	Plot of principal component 3 vs. 1--all 195 specimens coded by military/civilian status	34
9	Plot of principal component 3 vs. 2--all 195 specimens coded by military/civilian status	35
10	Plot of principal component 2 vs. 1--all 195 specimens coded by geographic region	36
11	Plot of principal component 3 vs. 1--all 195 specimens coded by geographic region	37
12	Plot of principal component 3 vs. 2--all 195 specimens coded by geographic region	38
13	Plot of principal component 2 vs. 1--all 195 specimens coded by race	39
14	Plot of principal component 3 vs. 1--all 195 specimens coded by race	40
15	Plot of principal component 3 vs. 2--all 195 specimens coded by race	41

LIST OF TABLES

Number	Title	Page
1	Number and Percent of Nondetects, Traces, and Positive Quantifiable Levels per Analyte in 197 NHATS Specimens....	6
2	Descriptive Concentration (pg/g) Statistics per Analyte Based on 197 NHATS Specimens.....	8
3	Correlation Coefficients Between Pairs of Analyte Concentrations Based on 197 NHATS Specimens.....	9
4	Distribution of 195 NHATS Specimens Across Military/ Civilian Status, Geographic Regions, and Race Groups.....	11
5	Descriptive Statistics of Selected Demographic Variables....	12
6	Simple Regression Results of 2,3,7,8-TCDD Concentrations vs. Selected Demographic Variables.....	14
7	Summary Table of Analyte Clusters Formed and Their Similarity Values.....	23
8	Principal Component Analysis Results.....	28

I. INTRODUCTION

This report presents the results and interpretation of a series of multivariate statistical analyses performed on a data base of levels of polychlorinated dibenzo-p-dioxins (PCDDs) and dibenzofurans (PCDFs) in human adipose tissue. The objective of this work assignment effort is to provide a means of evaluating the relationships between, and significance of, chemical and demographic variables and military/civilian status of the individuals represented in this data base. This work is to provide a complementary examination of the previous univariate analyses performed by the VA and USEPA (1989).

A. Project Background

Under a joint Department of Veterans Affairs/U.S. Environmental Protection Agency agreement, Midwest Research Institute (MRI) has analyzed approximately 200 human adipose tissue specimens for PCDDs and PCDFs. These tissue samples were obtained from a large, archived, set of specimens obtained through the National Human Adipose Tissue Survey (NHATS). The availability of military service status on males born between 1936 and 1954 was the basis for the selection of specimens in the three groups--Vietnam veterans, non-Vietnam veterans, and civilians--considered in this study (refer to the VA and USEPA 1989 report for selection procedure details). To the extent that the NHATS population is not representative of Vietnam veterans, the results are not representative of TCDD levels in the population of Vietnam veterans. The relatively small sample size (36) of Vietnam veterans would limit the significance of results relating to TCDD levels in the population of Vietnam veterans in any case. However, within the restricted NHATS population, the comparative results are valid.

The analytical results, expressed in parts per trillion (ppt), have been reported on a total of 17 selected compounds. These are the 2,3,7,8-substituted tetra- through octachloro-PCDD and PCDF congeners, which are considered to be the most toxic. Because two analytes, 1,2,3,4,7,8- and 1,2,3,6,7,8-HxCDD, coeluted, data were available on a total of 16 analytes. The complete list of analytes follows.

1. 2,3,7,8-TCDF
2. 2,3,7,8-TCDD
3. 1,2,3,7,8-PeCDF
4. 2,3,4,7,8-PeCDF
5. 1,2,3,7,8-PeCDD
6. 1,2,3,4,7,8-HxCDF
7. 1,2,3,6,7,8-HxCDF
8. 2,3,4,6,7,8-HxCDF
9. 1,2,3,7,8,9-HxCDD
10. 1,2,3,4,7,8/1,2,3,6,7,8-HxCDD, referred to as the pair-HxCDD
11. 1,2,3,7,8,9-HxCDD
12. 1,2,3,4,6,7,8-HpCDF
13. 1,2,3,4,7,8,9-HpCDF
14. 1,2,3,4,6,7,8-HpCDD
15. OCDF
16. OCDD

The data base was generated by MRI under Work Assignment 24 (EPA Contract No. 68-02-4252) using an analytical method (USEPA. 1986) that had been previously evaluated specifically for the VA/EPA program. All data were reported to VA/EPA in the form of 20 batch reports. The VA was responsible for analyzing the 2,3,7,8-TCDD data, the key compound with respect to Vietnam veteran exposure (VA and USEPA. 1989).

Since the data had been reported for all compounds, some effort was needed for the assessment of relationships between compounds and with respect to demographic data. Pattern recognition (multivariate statistical analysis) was proposed as an exploratory analysis tool. This approach, investigation of potential multidimensional relationships among the analyte results, was to be followed by classification of the subjects into groups, if groups could be identified. Thus the VA provided demographic information on these subjects. After all statistical analyses were completed, the military/civilian status (whether Vietnam veterans, non-Vietnam veterans, or civilians) information on the donors was made available so that possible groupings of the results could be identified that would parallel the three groups of donors.

B. Study Objectives

The objectives of this study were to examine the chemical and demographic data obtained on approximately 200 subjects using multivariate statistical techniques or pattern recognition. The relationships so elucidated among chemical and demographic variables, or among chemical variables alone, may be useful in answering pertinent questions such as, "Is there a difference among the exposure groups in adipose tissue levels of 2,3,7,8-TCDD?" and "If there is a difference, is there evidence that it is related to exposure other than Agent Orange?" The intent of this study was not to duplicate the VA's efforts, but rather to complement them by looking at all chemical data (PCDDs and PCDFs) and demographic variables simultaneously.

Since military information as well as the classification of the some 200 subjects into the three groups--Vietnam veterans, non-Vietnam veterans, and civilians--was kept confidential to the EPA and the VA until the end of the statistical analysis, MRI proposed performing first a "blind" statistical analysis of the available data and providing the EPA and VA with a possible classification of the subjects determined solely on the basis of chemical and demographic information. Groups identified in the pattern recognition approach would be identified by listing specimen identification numbers. The VA/EPA would then decode the results and determine whether any grouping of subjects would match exposure groups or any other groups defined by military-related data.

The approach taken to meet these study objectives was as follows. First, to elucidate relationships among the 16 analytes based on the sample concentrations, the approach focused on reducing the number of analytes by analyzing the correlation among them. Next, the chemical data, possibly considering a reduced set of analytes, and the demographic data were analyzed to identify groups among the individuals in the data set. Cluster analysis and principal component analysis were the two most appropriate techniques for this approach. Finally, the military/civilian status information was used in

the coding of the multivariate analysis results with respect to the three groups.

This report presents first a description of the chemical, demographic, and military/civilian status data bases (Section II). Next, basic descriptive statistics for each measured variable are presented and discussed (Section III). Section IV is a synopsis of each statistical procedure used for the analyses. In Section V, the results of each analysis are presented and discussed. The Appendix contains histograms of the concentration levels of each of the 16 analytes for which data were available.

II. DATA BASE DESCRIPTION

Three data bases were made available for this work assignment effort. The first and largest data base contained chemical information (PCDD and PCDF data) only, the second data base contained demographic information on the subjects, and the third, which was provided after all other analyses were completed, provided the military/civilian status information on each individual.

A. Chemical Data Base

The data base used had been generated previously by MRI and was transferred on a floppy diskette for subsequent statistical analysis. Thus no manual data entry was required. Accuracy of the data base was checked by means of various frequency checks, random record comparisons with the batch reports provided to the EPA and VA, and other logical checks to determine consistency throughout the data base.

The data base contained information on a total of 274 samples. The total consisted of 197 NHATS samples, 20 method blanks, and 57 quality assurance samples. For each sample, the following information was available:

1. WID, an encoded sample identifier;
2. MRI lab number;
3. Sample type;
4. Batch number;
5. Low resolution mass spec run date (PCDDs and PCDFs);
6. High resolution mass spec run date (2,3,7,8-TCDD and 2,3,7,8-TCDF);
7. Specimen wet weight (g); and
8. Percent lipid of specimen.

For the 16 analytes, the following three variables were listed:

9. Data qualifier (ND = not detected, TR = trace, PQ = positive quantifiable);

10. Limit of detection (pg/g) if the data qualifier was either "ND" or "TR"; and

11. Concentration if the data qualifier was either "TR" or "PQ." All concentrations were percent lipid based and expressed in picograms per gram (pg/g) or equivalently in parts per trillion (ppt).

B. Demographic Data Base

The data base containing selected demographic information on the subjects was provided by the VA. These data were transferred to MRI in the form of an ASCII file on a floppy diskette. For each subject, the following information was given:

1. WID, to match the records across data bases;

2. Geographic region (North Central, North East, South, or West) determined by the location of the individual's hospital;

3. Race of individual (Caucasian, Black, Oriental, American Indian, or Other);

4. Age of individual (years) at time of specimen collection (i.e., at time of death or surgery)

5. Specimen collection year;

6. Year of birth; and

7. Body mass index (kg/m^2).

C. Military/Civilian Status Data Base

After all statistical analysis activities were concluded, a data base including the military/civilian status of each individual in the study was provided by the VA and EPA. This data base, which was provided to MRI in the form of an ASCII file on floppy diskette, included two variables:

1. WID to match the three data bases; and

2. Military/civilian status code (Vietnam veteran, non-Vietnam veteran, civilian).

The data base contained information on 195 of the 197 NHATS specimens (36 Vietnam veterans, 79 non-Vietnam veterans, and 80 civilians). In some

cases, demographic and military/civilian status information was not available for all specimens analyzed. These differences in sample sizes are reflected in subsequent tables.

III. DESCRIPTIVE STATISTICS

The analysis results presented and discussed in this report are based solely on the 197 NHATS specimens. Analysis of the quality assurance sample results (method blanks, duplicates, performance audit samples, split samples, and spiked and unspiked QC samples) is not reported. A statistical analysis of these sample results could be undertaken in a future task and incorporated with the results reported herein as a means of: (a) validating relationships, if they exist, among the 197 samples and the chemical variables; and (b) analyzing the QA/QC results in a multivariate context. Since the data analysis techniques used here are variance based, it would be useful to explicitly examine the fraction of the total variance that may be attributed to the analytical/sampling procedures. Knowledge of these variance sources would add considerable strength to the inferences drawn about distinctions or lack thereof among the sample and control populations.

The following subsections present basic descriptive, i.e., univariate and bivariate, statistics for the chemical and demographic data.

A. Descriptive Statistics of the Chemical Data

Chemical information on a total of 197 samples was available for analysis. For each of the samples, lipid-based concentrations (pg/g) on each of the selected 16 analytes was given if the analyte was detected. When the analyte was not detected, the limit of detection was reported (also equivalent to lipid concentration). Throughout all analyses, half the limit of detection value was used for those samples with levels below detection limits. This approach was taken to conform with the approach chosen by VA statisticians. When the analyte was detected at a trace level, that is, between the limit of detection and the limit of quantitation, the reported concentration was used in the analyses.

1. Univariate Statistics

As a first step, percentages of nondetects (ND), traces (TR), and positive quantifiable (PQ) levels were summarized for each analyte. The results, both in frequency counts and in percentages, are shown in Table 1. An upper limit of 25% for the total number of specimens with levels below detection limit for a given analyte (i.e., at least 75% PQ+TR) was selected as the cutoff point for that analyte to be included in subsequent multivariate data analyses. By that rule, a total of five compounds were excluded from analyses. They were: 2,3,7,8-TCDF (35.0% NDs); 1,2,3,7,8-PeCDF (79.2% NDs); 1,2,3,7,8,9-HxCDF (95.4% NDs); 1,2,3,4,7,8,9-HpCDF (53.8% NDs); and OCDF (27.9% NDs).

Table 1. Number and Percent of Nondetects, Traces, and Positive Quantifiable Levels per Analyte in 197 NHATS Specimens

	Number			Percent		
Analyte Name/No.	ND	TR	PQ	ND	TR	PQ
2378-TCDF 1*	69	20	108	35.0	10.2	54.8
2378-TCDD 2	1	2	194	0.5	1.0	98.5
12378-PeCDF 3*	156	23	18	79.2	11.7	9.1
23478-PeCDF 4	2		195	1.0		99.0
12378-PeCDD 5	1	1	195	0.5	0.5	99.0
123478-HxCDF 6	2	1	194	1.0	0.5	98.5
123678-HxCDF 7	5	1	191	2.5	0.5	97.0
234678-HxCDF 8	19	19	159	9.6	9.6	80.7
123789-HxCDF 9*	188	7	2	95.4	3.6	1.0
123478/123678-HxCDD 10			197			100
123789-HxCDD 11	2	2	193	1.0	1.0	98.0
1234678-HpCDF 12			197			100
1234789-HpCDF 13*	106	49	42	53.8	24.9	21.3
1234678-HpCDD 14			197			100
OCDF 15*	55	64	78	27.9	32.5	39.6
OCDD 16			197			100

Note: Analytes marked by an * were excluded from some subsequent statistical analyses due to high occurrences of nondetects

ND = Nondetect--level below detection limit
TR = Trace--level between detection and quantitation limits
PQ = Positive Quantifiable--level above quantitation limit

The analyte with the largest percentage of nondetects (9.6%) included in the analyses is 2,3,4,6,7,8-HxCDF. All other analytes retained for subsequent analyses have percentages of nondetects below 3%.

The concentration data are summarized in Figures 1 through 16 in the Appendix. These plots show histograms of the concentration levels found in the 197 NHATS specimens for each of the 16 analytes.

The histograms show evidence of relatively skewed distributions of analyte levels in the adipose tissue specimens of the study subjects. For each analyte, most of the levels are near the lower end of the scale, with a few relatively large values in each case. These distributions could well be represented by log-normal distributions, and this fact is considered in the discussion of cluster analysis in Section V.A. This approach has also been discussed by Sielken (1987) for TCDDs in human adipose tissue in particular.

Basic descriptive statistics of the analytical results are summarized in Table 2. For each analyte, mean, standard deviation, minimum, 25th percentile, median or 50th percentile, 75th percentile, and maximum of the analyte levels across the 197 specimens have been computed. All statistics are in picograms per gram (or ppt). For levels below detection limits, half the detection limit was used for the computations.

For each analyte, the median is lower than the mean, which confirms the positive skewness of the distributions mentioned earlier. This can also be observed in the values of the 25th and 75th percentiles for each analyte. These values are such that 50% of the specimens have concentration levels between these two numbers; 25% of the specimens have concentrations above the 75th percentile; and 25% of the specimens have concentrations below the 25th percentile.

For example, 50% of the specimens have 2,3,7,8-TCDD levels between 7.8 and 16.5 pg/g. The median level is 11.5 pg/g. The average concentration of 14.1 pg/g, however, is influenced by one high value of 106 pg/g. The second highest 2,3,7,8-TCDD value is 55.8 pg/g and is from a control specimen. (It should be noted that the extreme 2,3,7,8-TCDD level of 106 pg/g also is from a control subject. Most of this subject's analyte levels are in the upper 25th percentile.) At a more refined level, 10% of the samples have 2,3,7,8-TCDD levels above 25 pg/g, and 5% of the sample concentrations exceed 33.4 pg/g (not shown in Table 2). On the low side of the spectrum, 10% of the specimens have levels below 6.1 pg/g, and 5% of the sample concentrations fall below 5 pg/g (not shown in Table 2). The histograms in the Appendix depict this situation best for each analyte.

2. Bivariate Statistics

Correlation coefficients between each pair of analytes were computed. All correlations are positive, and with a few exceptions, most are statistically significantly different from zero at the 99% confidence level. Due to the large number of specimens (197), even moderately small correlations will be statistically significant. For example, a correlation coefficient of 0.18 is significant at the 99% level for a sample size of 200. Table 3 lists

Table 2. Descriptive Concentration (pg/g) Statistics per Analyte Based on 197 NHATS Specimens

Analyte Name/No.		Mean	Standard Deviation	Minimum	25th Percentile	Median	75th Percentile	Maximum
2378-TCDF	1	2.10	2.48	0.13	0.77	1.46	2.62	18.1
2378-TCDD	2	14.1	10.9	2.5	7.8	11.5	16.5	106.0
12378-PeCDF	3	0.60	0.87	0.08	0.20	0.33	0.60	7.01
23478-PeCDF	4	23.0	15.0	2.0	14.7	20.1	26.1	112.6
12378-PeCDD	5	18.9	12.1	2.8	12.6	16.3	20.8	89.8
123478-HxCDF	6	21.4	14.1	2.7	12.7	18.0	25.1	102.0
123678-HxCDF	7	10.9	8.2	1.4	6.7	9.2	11.8	65.0
234678-HxCDF	8	3.3	2.7	0.3	1.9	2.7	3.8	26.1
123789-HxCDF	9	0.33	0.33	0.04	0.15	0.23	0.36	2.8
123478/123678-HxCDD	10	163.9	106.7	31.8	114.0	136.0	171.5	831.5
123789-HxCDD	11	18.0	11.8	2.8	11.8	15.1	19.5	102.0
1234678-HpCDF	12	36.5	25.6	5.5	20.7	29.0	43.7	179.4
1234789-HpCDF	13	1.4	1.2	0.20	0.65	1.02	1.79	9.07
1234678-HpCDD	14	273.4	248.4	59.1	147.9	214.9	304.7	2,677
OCDF	15	3.1	4.1	0.31	1.09	1.97	3.23	28.2
OCDD	16	1,268	1,008	289.0	732.2	998.0	1,412	8,804

Note: Statistics are calculated using one half the limit of detection for all nondetects

Table 3. Correlation Coefficients Between Pairs of Analyte Concentrations Based on 197 NHATS Specimens

Analyte Name/No.		1*	2	3*	4	5	6	7	8	9*	10	11	12	13*	14	15*
2378-TCDF	1*															
2378-TCDD	2															
12378-PeCDF	3*	0.67														
23478-PeCDF	4		0.50													
12378-PeCDD	5		0.59		0.76											
123478-HxCDF	6		0.62		0.84	0.79										
123678-HxCDF	7		0.55		0.89	0.79	0.95									
234678-HxCDF	8				0.72	0.66	0.82	0.83								
123789-HxCDF	9*															
123478/123678-HxCDD	10		0.58		0.82	0.92	0.88	0.89	0.76							
123789-HxCDD	11		0.57		0.76	0.87		0.88	0.85		0.91					
1234678-HpCDF	12		0.54		0.67	0.65	0.87	0.81	0.76		0.75	0.81				
1234789-HpCDF	13*				0.61	0.59	0.77	0.75	0.74		0.67	0.74	0.83			
1234678-HpCDD	14	0.51	0.50		0.63	0.65	0.81	0.79	0.93		0.76	0.87	0.80	0.76		
OCDF	15*															
OCDD	16		0.61		0.73	0.73	0.88	0.87	0.82		0.85	0.89	0.84	0.76		0.89

Note: Analytes marked by an * had levels below detection limit in more than 25% of the specimens (see Table 1)

Only coefficients greater than or equal to 0.5 are shown

all correlation coefficients between pairs of analyte concentrations that are greater than or equal to 0.50. The cutoff value of 0.50 for the correlation coefficients was selected for the following reason. The correlation coefficient, R, between two analytes, X and Y, is a measure of the linear dependency between X and Y. In the context of a linear regression of Y on X, R^2, the square of the correlation coefficient, will be the proportion of total variation about the mean of Y explained by the regression of Y on X. Thus, setting the cutoff for R at 0.50 is equivalent to setting the cutoff for R^2 at 25%. For all practical purposes, anything below that figure can be considered negligible in terms of linear association and predictive power.

Of the 120 possible pairs of analyte concentrations, slightly more than half (66) yield correlation coefficients above 0.50. Of these, about half (29) are above 0.80, with 1,2,3,4,7,8-HxCDF and 1,2,3,6,7,8-HxCDF being the most highly correlated pair of analytes with a coefficient of 0.95.

Of the correlations above 0.50, 2,3,7,8-TCDD shows overall the lowest correlation with most other analytes. Using 0.80 as a cutoff value, one could say that 2,3,7,8-TCDD is not strongly correlated with any of the other analytes.

B. Descriptive Statistics of the Demographic Data by Military/Civilian Status

A set of demographic variables was made available by the VA to be incorporated into the analysis. The variables were geographic region, race, age, collection year, year of birth, and body mass index.

Demographic data were not available for all 197 NHATS samples. Most data were available for 196 subjects; age, race, and geographic region were missing for two of the 197 subjects; and body mass index was available for only 169 subjects. Military/civilian status was reported on 195 of the 197 specimens in the study.

Table 4 provides a cross-tabulation of the 195 specimens across military/civilian status, geographic regions, and race groups. Eighty percent of the individuals were Caucasian, and the remaining 20% were distributed across four race groups. Consequently, the classification of individuals as to race was simplified in subsequent analyses to two groups--Caucasian and non-Caucasian.

Table 5 summarizes basic descriptive statistics for the demographic variables--number of specimens, mean, standard deviation, minimum and maximum values, and coefficient of variation. The statistics in Table 5 are shown across all specimens (varying sample sizes) as well as within each category determined by the military/civilian status of the individual. The careful choice of matched NHATS specimens across the three groups can be clearly seen in this table by noting the small differences in these statistics.

Table 4. Distribution of 195 NHATS Specimens across Military/Civilian Status, Geographic Region, and Race Groups

Military/Civilian Status	Geographic Region	Caucasian	Black	Oriental	American Indian	Other	Total *1	
							Number	% of 195
Vietnam Veteran	North Central	6	1	0	0	0	7	3.6
	North East	4	2	0	0	0	6	3.1
	South	10	5	0	0	0	15	7.7
	West	7	1	0	0	0	8	4.1
	All	27	9	0	0	0	36	18.5
Non-Vietnam Veteran	North Central	19	2	0	0	0	21	10.8
	North East	16	0	0	0	1	17	8.7
	South	29	6	0	0	0	35	17.9
	West	5	0	0	1	0	6	3.1
	All	69	8	0	1	1	79	40.5
Civilian	North Central	9	2	0	0	0	11	5.6
	North East	22	5	0	0	0	27	13.8
	South	23	9	1	1	0	34	17.4
	West	6	1	0	0	1	8	4.1
	All	60	17	1	1	1	80	41.0
All Specimens	Number	156	34	1	2	2	195	
	Percent	80.0	17.4	0.5	1.0	1.0		

Note *1: Demographics were unavailable for 2 of the 197 NHATS specimens

Table 5. Descriptive Statistics of Selected Demographic Variables

Demographic Variables	Number of Specimens	Mean	Standard Deviation	Minimum	Maximum	Coefficient of Variation (%)
Age (years)	195	32.1	5.3	20	45	16.5
Vietnam Veterans	36	31.2	4.9	22	41	15.8
Non-Vietnam Veterans	79	32.6	5.6	20	43	17.1
Civilians	80	32.0	5.2	21	45	16.3
Collection Year	195	1978 *1	3.0	1971	1982	3.8
Vietnam Veterans	36	1978	3.1	1972	1982	4.0
Non-Vietnam Veterans	79	1979	3.0	1971	1982	3.9
Civilians	80	1978	3.0	1972	1982	3.8
Birth Year	195	1946 *1	4.1	1936	1954	9.0
Vietnam Veterans	36	1947	3.1	1939	1952	6.7
Non-Vietnam Veterans	79	1945	4.8	1936	1954	10.7
Civilians	80	1946	3.7	1937	1951	8.1
Body Mass Index (kg/m2)**	169	26.6	5.7	17.7	67.8	21.6
Vietnam Veterans	34	26.7	5.3	19.1	40.7	19.9
Non-Vietnam Veterans	69	26.0	3.7	20.4	39.1	14.2
Civilians	66	27.3	7.5	17.7	67.8	27.4

Note *1: Statistic is median rather than mean for these variables.

C. Relationships Between 2,3,7,8-TCDD and Selected Demographic Variables

The levels of 2,3,7,8-TCDD, the most toxic dioxin and the predominant contaminant of Agent Orange, were examined more closely than other congeners in relationship to a few demographic variables. Relationships were analyzed by means of simple regression analysis, that is, with one independent variable at a time. Both the concentration levels and the natural logarithm of the concentrations were regressed against age, collection year, birth year, and body mass index. These regression analyses were also performed after excluding the one specimen with the highest 2,3,7,8-TCDD level (106 pg/g). The second highest level was 55.8 pg/g. All regression results, with and without log-transformation and with and without the 2,3,7,8-TCDD extreme value, are shown in Table 6.

No strong linear relationships of 2,3,7,8-TCDD levels emerged with any of these variables. Some regression models were significant at the 5% significance level, that is, the slopes were significantly different from 0 at the 5% level; however, the extremely low values of R-squared, a measure of the percent variance in the data explained by the model, do not warrant drawing any conclusions as to the ability of these variables to predict levels of 2,3,7,8-TCDD. These results are in general agreement with those obtained from a stepwise regression analysis performed by the VA (VA and USEPA. 1989) of 2,3,7,8-TCDD level vs. demographic variables. The regression results are summarized in Table 6 below. Scatterplots of 2,3,7,8-TCDD levels vs. demographic variables are displayed in Figures 1 through 4 to illustrate the random distribution of 2,3,7,8-TCDD levels across these variables.

Figures 1 and 4 depict 2,3,7,8-TCDD levels vs. the individuals' age (at time of death or surgery) and body mass index, respectively. These plots underline the nonsignificant regression results summarized in Table 6.

Figure 2 shows 2,3,7,8-TCDD levels against specimen collection year. Due to the significant regression results mentioned above, the regression line with a slope of -0.63 has been drawn here. There is a slight decrease of 2,3,7,8-TCDD levels with successive collection year, but the predictive power of that variable in estimating the analyte level is very low (R^2 of 3%). The significance of the small, negative slope is mostly due to the large sample size in this specific situation.

Figure 3 shows 2,3,7,8-TCDD levels vs. the individuals' year of birth. The significant regression line with a slope of -0.60 has been drawn to show the slight decrease of 2,3,7,8-TCDD levels with successive birth year. It should be noted, however, that although this slope is significantly different from 0 at the 5% level, the predictive power of birth year with respect to that analyte is very low (R^2 of 5%). Again, the significance of the small, negative slope is mostly due to the large sample size.

Table 6. Simple Regression Results of 2,3,7,8-TCDD Concentrations vs. Selected Demographic Variables

Variable Used	2,3,7,8-TCDD vs:	Sample Size	Significant*1 at 5% Level?	R-Squared (%)	Regression Slope	Standard Error of Slope
Concentrations	Age	195	NO			
	Specimen Collection Year	196	YES	3	-0.63	0.26
	Birth Year	196	YES	5	-0.60	0.18
	Body Mass Index	169	NO			
Log Concentrations	Age	195	NO			
	Specimen Collection Year	196	YES	10	-0.06	0.01
	Birth Year	196	YES	7	-0.04	0.01
	Body Mass Index	169	NO			
Concentrations (Without extreme value of 106 pg/g)	Age	194	NO			
	Specimen Collection Year	195	YES	7	-0.75	0.20
	Birth Year	195	YES	5	-0.48	0.15
	Body Mass Index	168	NO			
Log Concentrations (Without extreme value of 106 pg/g)	Age	194	NO			
	Specimen Collection Year	195	YES	12	-0.06	0.01
	Birth Year	195	YES	7	-0.03	0.01
	Body Mass Index	168	NO			

Note *1: Test of whether the slope is statistically significantly different from zero.

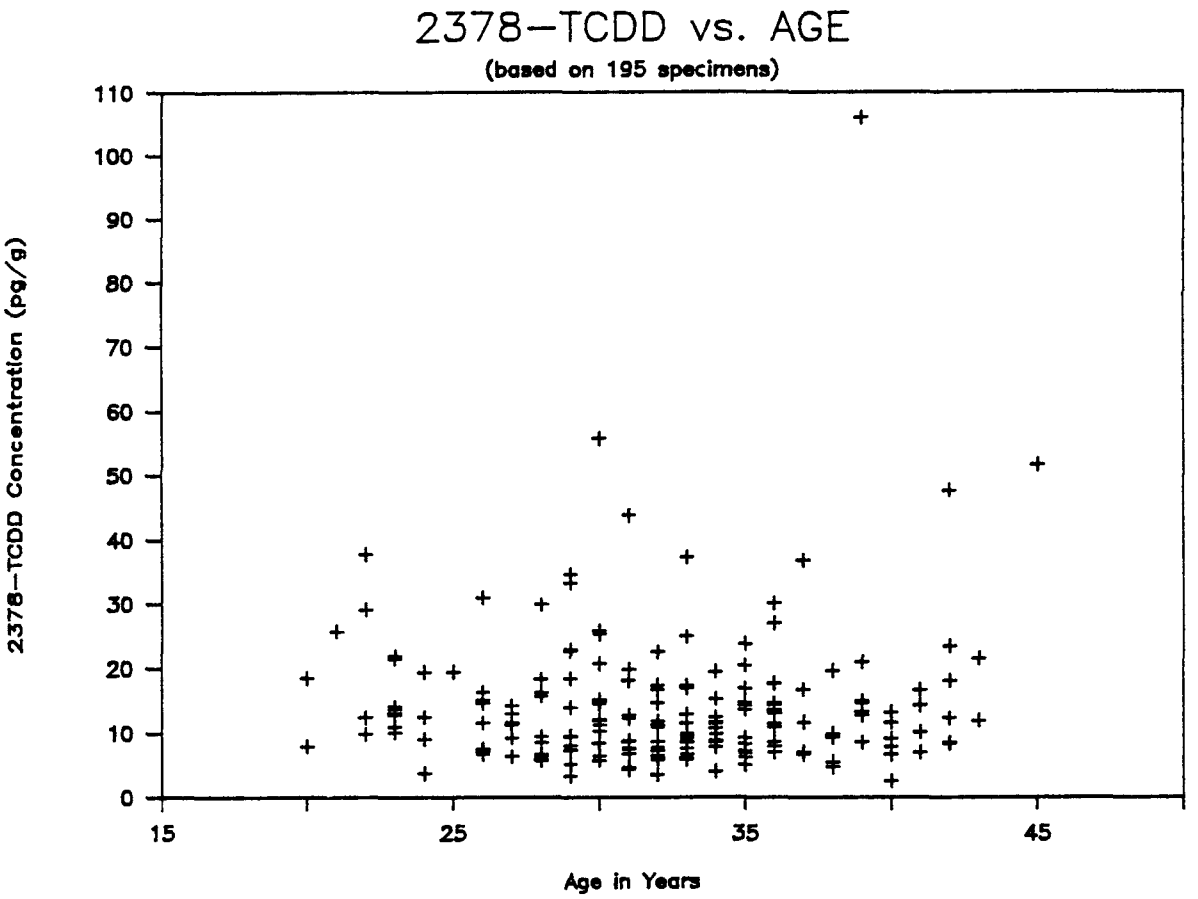


Figure 1: Plot of 2,3,7,8-TCDD vs. age of individual at time of collection.

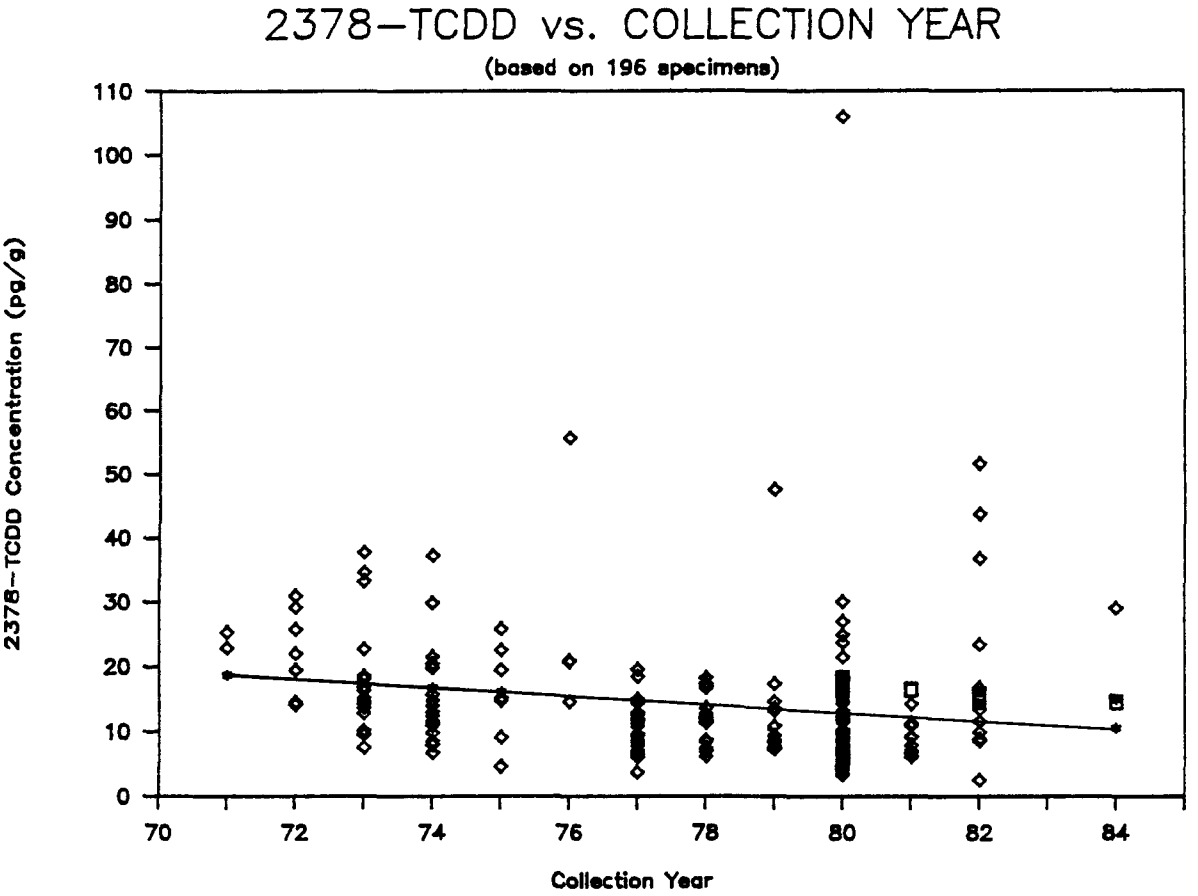


Figure 2: Plot of 2,3,7,8-TCDD vs. collection year.

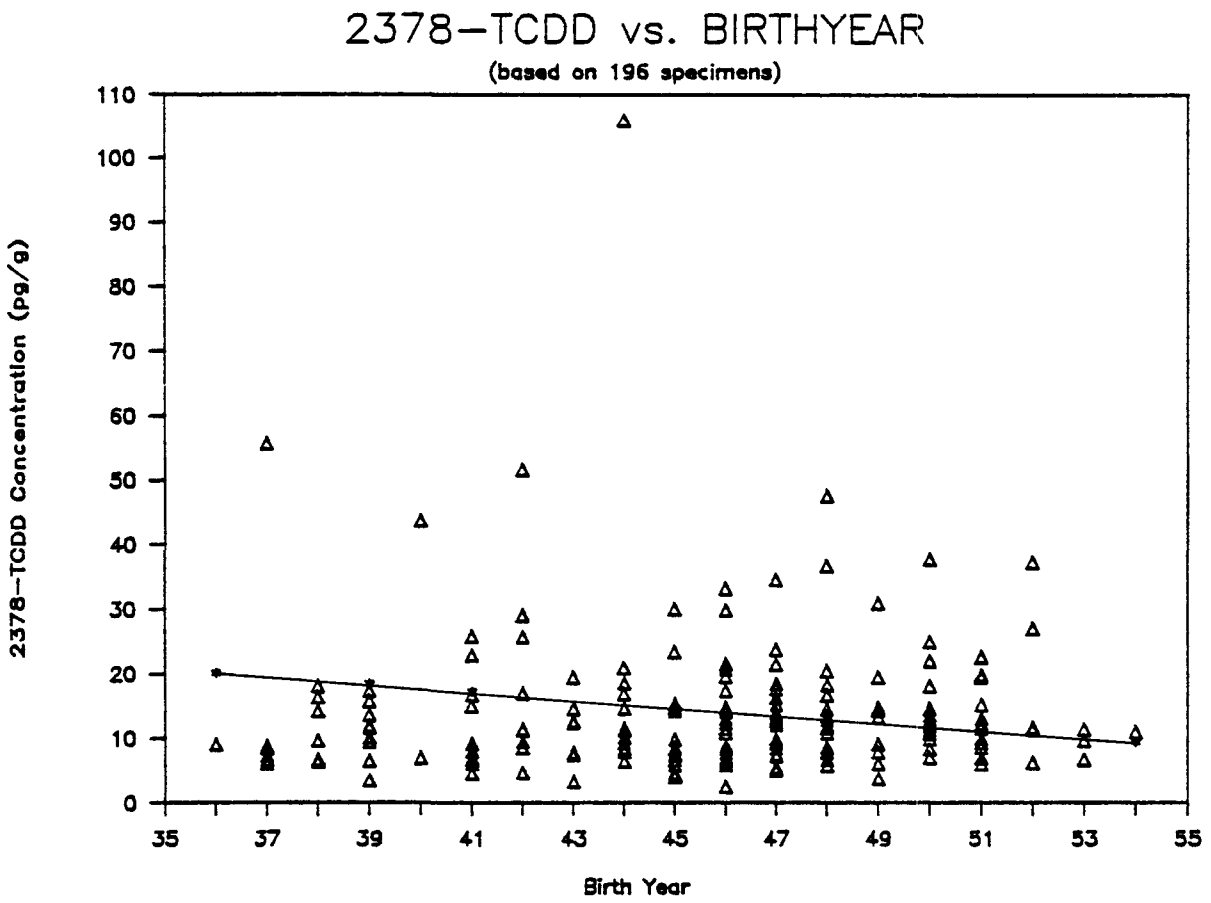


Figure 3: Plot of 2,3,7,8-TCDD vs. birth year of individual.

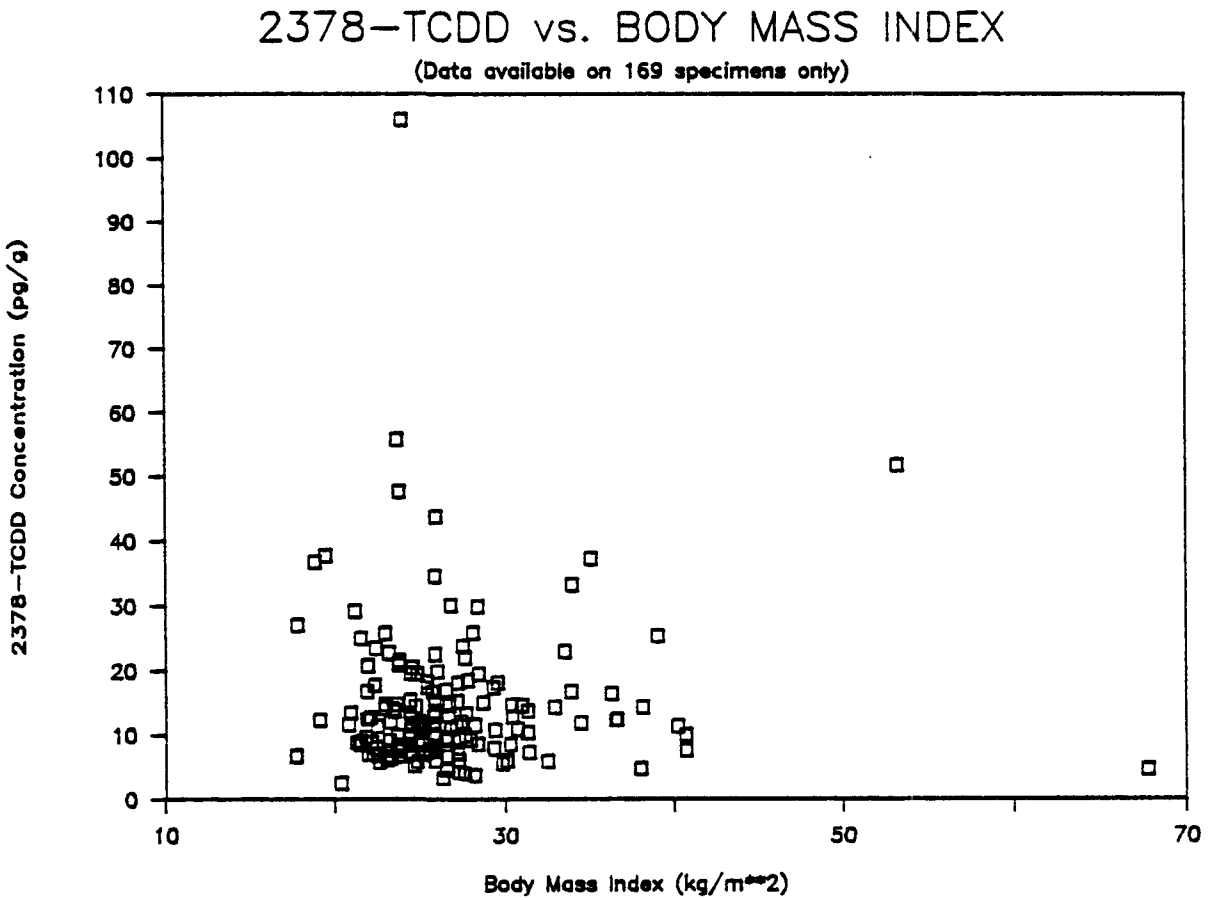


Figure 4: Plot of 2,3,7,8-TCDD vs. body mass index.

It should be noted that the three variables, age, year of birth, and collection year, are interrelated as follows: age = collection year-year of birth. Therefore, considering any two of the three variables is sufficient in subsequent analyses, since the third variable provides no additional information.

IV. MULTIVARIATE STATISTICAL METHODS

This section presents a brief overview of the multivariate techniques described in Section V. To this point, simple statistics such as means, medians, standard deviations, and frequencies of all the available variables have been discussed. These statistics provide an efficient summary of the data space, but only on an individual variable basis. To show relationships between any pair of variables, if they exist, sample correlation coefficients are needed. The limitations of these two-way relationships are immediately obvious, however. If one were to look at all two-way relationships between the 16 analytes on the one hand and the 6 demographic variables on the other, plotting and analyzing would require 96 bivariate graphs. And these relationships would not include the 120 two-way plots between pairs of analytes. Not only would it be a large analytical task, but most important, relationships between more than two variables would never be considered. Thus the objective is to be able to consider sets of variables simultaneously. The problem then becomes one of dimensionality; namely, with more than three variables, we are practically unable to visualize any relationship in space.

Multivariate techniques, or pattern recognition techniques, have been developed and provide rules for combining variables in an optimal way (the criterion of optimality varies from one technique to another) so as to reduce the dimensionality of the set of variables with as little loss of information as possible. In this context, two statistical techniques have been used to complement each other: cluster analysis and principal component analysis.

There are three major objectives in pattern recognition: data reduction, feature extraction, and classification. In cluster analysis, in which little a priori knowledge of possible grouping of variables is required, natural groupings and structures among variables are studied without prejudice. The same technique can be applied to the observations (specimens) in the study. In principal component analysis, the goal is to define linear combinations of the original variables and thus to reduce the dimension of the data space.

The Biomedical Programs (BMDP) and the Statistical Analysis System (SAS) software packages have been used exclusively in this study.

A. Cluster Analysis of Variables

Given a large set of variables, multivariate relationships are sought instead of the simple bivariate relationships found with correlation analysis techniques. Cluster analysis is applied to identify similarities between analytes, not between observations (specimens in this case). Analytes

contributing nearly the same information to the data structure are clustered; one can then select one or more analytes from a cluster, discard others if desired, and proceed. Cluster analysis thus shows whether there are natural groups of analytes with similar behavior.

The first step in cluster analysis, then, is to define a measure of association for clustering variables. In this study, the correlation coefficient is chosen. The single linkage option in the BMDP P1M procedure is selected for the clustering. This linkage process continues until all variables have been included in a single large cluster. The clustering results are then printed in a dendrogram or tree diagram. The determination of the final number of clusters is, however, left to the user and depends on a chosen minimum similarity value (the correlation coefficient in this case). If the value chosen is too high (i.e., close to 1), then many clusters will be formed, and most of them might be single-point clusters. Since outputs provided by the classical statistical software packages show results after complete clustering, a decision as to an optimum similarity value can be made a posteriori. This is the cluster analysis technique that was applied exclusively to the 16 analytes.

B. Principal Component Analysis

Principal component analysis (PCA) is an effective tool for data reduction; and at the same time, it provides a means for data representation in two- or three-dimensional space. In PCA, the goal is to assess the structure of the variables within a given set of variables, independently of any relationship they may have to variables outside this set.

The first step in PCA is to choose the chemical and demographic variables to be included in the set and a measure of association between them. In order to treat all variables as equally important, we have chosen the correlation matrix as the measure of association (i.e., the data were autoscaled). Also, all concentration data have been log-transformed for PCA to improve on the skewness of the distributions of analyte results.

The next step is to combine the original variables into linear combinations in a manner that accounts for the maximum amount of variance present in the data. If the starting set of variables contains **m** variables, then **m** principal components will be generated. The first principal component chosen is that which represents the single best linear combination of variables, in other words, accounts for the highest percentage of the total variance in the data. The second principal component is chosen to account for as much of the remainder of variance as possible. Hence, in PCA, the principal components are orthogonal to each other, which means they are uncorrelated. This process continues until **m** linear combinations are found.

In general, a few principal components are sufficient to explain a large amount of the variance in the data. The importance of a principal component is mostly judged by its associated eigenvalue in the correlation matrix. Eigenvalues decrease with decreasing importance of the principal components. Past experience indicates that principal components with eigenvalues below 1.0 explain little correlated behavior among the measured

variables and that they often reflect random variance contributions due to measurement and sampling error. Therefore, principal components with eigenvalues below 1.0 are generally discarded. A second criteria in the selection of the final number of principal components is the final communality estimates associated with each variable. Final communality estimates are a measure of how well the variables are accounted for by the number of principal components selected. Thus, an additional component, although having a low eigenvalue, might be included in the final set.

The coefficients of the variables (or transformed variables) in the linear combinations are called loadings. They are a measure of the relative contribution of a variable within a principal component to the variance explained. In fact, the principal component loadings are the correlation coefficients between the original (or transformed) variables and the principal components, and thus provide the key information about the nature of the principal component. These loadings can be used to order the variables within a given principal component and help affix a "label" to this principal component. This step in the PCA is the most difficult one because it is largely a matter of judgment. This is where the interpretation of the results begins.

Once the principal components have been derived and a final selection of significant principal components made, principal component scores for each data point (specimen) are calculated from the standardized data. This is done by using the linear combinations of variables with their loadings on each principal component. Thus, if three principal components are retained, three scores will be computed for each specimen, and three two-dimensional plots of the scores can be generated: principal component 2 vs. principal component 1, principal component 3 vs. principal component 1, and principal component 3 vs. principal component 2, or one three-dimensional plot. Since the principal components are linear combinations of the original (transformed) variables, it is expected that the plot of principal component 1 vs. principal component 2 should contain more information about the data structure than any simple two-variable plot. The second most difficult step is then to give practical meaning to any pattern seen on the principal component plots. On the one hand, care should be taken not to oversimplify the relationships between variables; on the other hand, one should not read too much into the patterns found.

When plotting the two- or three-dimensional principal component plots, the data points can be labeled using available codes. These codes could be those representing selected categorical demographic variables, for example, race, geographic region, or a combination of both, or military/civilian status. An additional approach is to plot only those scores lying between -3 and +3, thus including all but a few large values. This achieves a higher resolution of the plots so that possible patterns in the figures can be made more visible.

V. MULTIVARIATE ANALYSIS RESULTS AND DISCUSSION

Following the initial univariate and bivariate (descriptive statistics, regressions and correlations) analyses of the data, a series of multivariate procedures, as described, were applied. The same techniques were used to analyze various subsets of the data, the results of one technique leading to the use of another. Cluster analysis results for the chemical information are discussed first, followed by a discussion of the results of principal component analysis.

A. Cluster Analysis of Analytes

1. Approach

The objective of the cluster analysis, the first analysis step, was to uncover possible significant interrelationships among analytes or groups of analytes. The input for that analysis was the data matrix of all analytical results from the 197 NHATS subjects. Levels below detection limits were replaced by one-half the limit of detection. Since the detection limit for most chemicals is well below typical values, the effect of this procedure for handling nondetects is minimal.

Four sets of cluster analyses were performed. They differed with respect to the number of analytes included and whether the concentrations or the natural logarithms of the concentrations were used as input. The log-transform approach was selected to resolve two potential problems: (1) non-normality of the PCDD and PCDF concentration levels and (2) sensitivity of the cluster analysis to extreme concentration values. The cluster analyses were performed both with and without the five PCDFs with more than 25% nondetects.

Absolute correlations (in this case, simple correlations since all correlations are positive) were used as the measure of similarity between any two analytes. When a cluster contained more than one analyte, the similarity between any two clusters was computed as the maximum similarity over all possible pairings of analytes between the two clusters. At each step in the clustering process, the two clusters with the maximum similarity were combined. The process continued until all 16 analytes were included into a single cluster.

The BMDP P1M procedure, cluster analysis of variables, was used to perform these analyses.

2. Results

Of the cluster analyses mentioned above, only those results obtained when using the log-transformed concentrations for all 16 analytes are presented here. The results of the other three cluster analyses were very similar. Excluding those five analytes with high occurrences of nondetects from the set of analytes did not influence the grouping of the remaining analytes. As shown below, these five analytes remained "individuals" in the clustering process, that is, one-variable clusters.

The log-transformation had some effect on the correlations shown between pairs or groups of analytes. Overall, the correlation coefficients between pairs of log-transformed analyte concentrations dropped slightly as compared to the correlation coefficients presented in Table 3 (based on untransformed concentrations). However, as noted earlier, due to the skewness of the concentration distributions (see histograms in the Appendix), emphasis was placed on the results obtained using the log concentrations. This approach is also in line with that taken by the VA statisticians in their analysis of 2,3,7,8-TCDD data (VA and USEPA. 1989).

The clustering results are shown in Figure 5 in the form of a tree diagram. A companion table (Table 7) has been developed to show the similarity between two clusters at the time they are joined in the final clustering process. The analyte names and numbers are listed on the left side of the diagram in Figure 5. The tree diagram is printed over the similarity matrix and shows the clusters formed during the linkage steps. Solid lines indicate boundaries of clusters such that the similarity between the last two clusters joined exceeds 0.80. Broken lines are used in all other cases.

The first two columns in Table 7 list the two analytes that form the cluster boundaries when a given cluster is considered. The third column shows the number of analytes in that specific cluster. The last column indicates the value of the similarity at the time when the cluster is formed. For example, take analyte 10. A two-analyte cluster consisting of analyte 10 (first column) and analyte 5 (second column) is formed early in the process with a similarity of 0.92 (last column). Since this cluster contains only two analytes, the value of 0.92 is simply the correlation between these two analytes. A similar situation occurs with analyte 7 and analyte 6 (first and second column in Table 7) with a correlation of 0.94.

Next, take analyte 12 (first column in Table 7). A four-analyte cluster (third column) bound by analytes 12 and 4 (second column) is formed with an overall similarity of 0.85 (last column). This cluster also contains the two-analyte cluster formed by analytes 6 and 7 (see Figure 5).

An example of a larger cluster would be the cluster delineated by analyte 4 (first column in Table 7) and analyte 8 (second column) and containing 10 (third column) of the 16 analytes with an overall similarity of 0.82. The 10 analytes contained in that cluster can be read from Figure 5 where it is shown by a solid line as the largest cluster.

The last column in Table 7 shows a drop in similarity values between the 3rd and the 4th rows (0.59 and 0.82, respectively) and between the 12th and 13th rows (0.83 and 0.72, respectively). These relatively sharp drops in similarity values were the basis for the cutoff point selection. Also, with a sample size of 197, using a correlation of 0.80 as a cutoff provides a conservative approach to the interpretation of the clustering process.

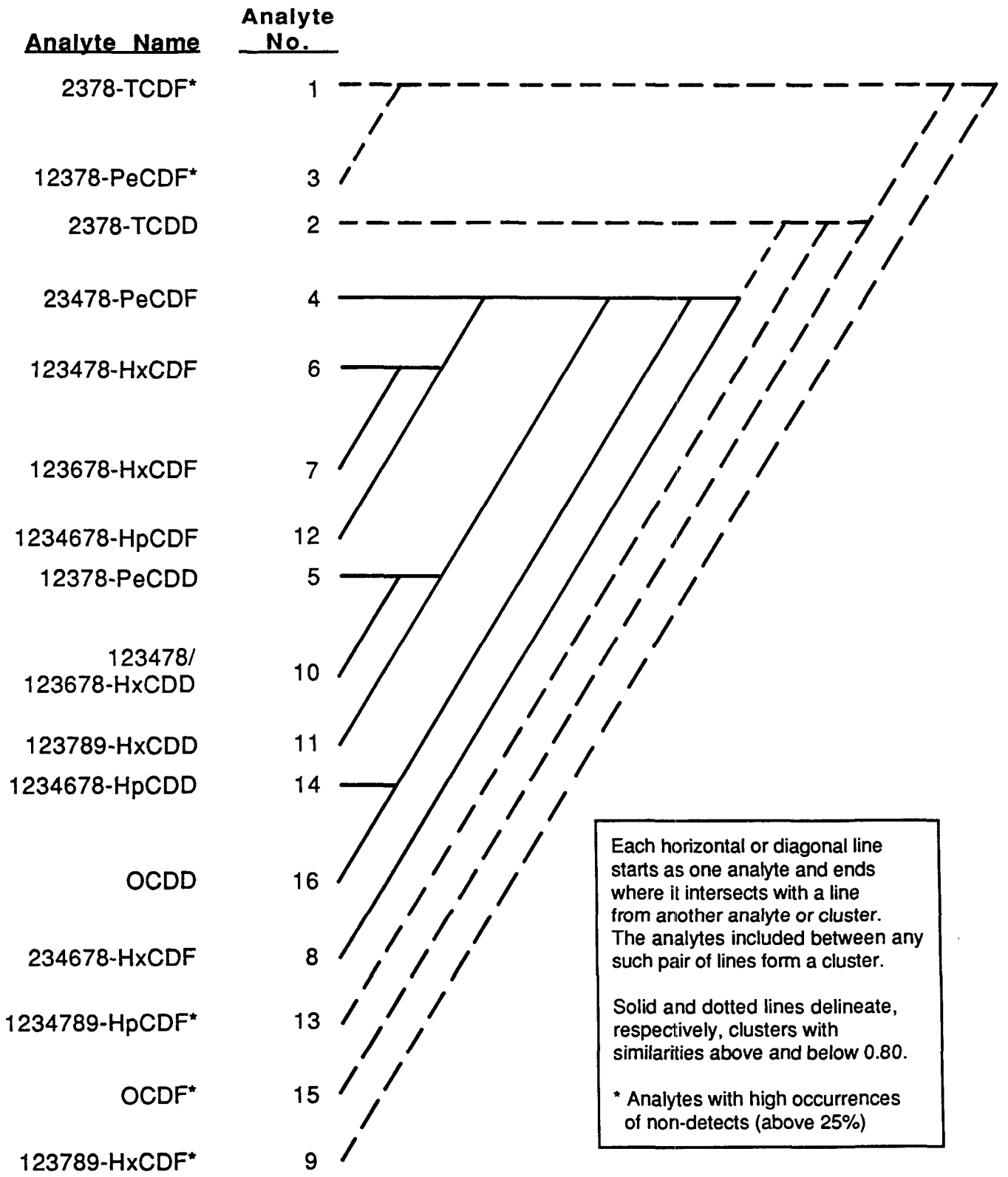


Figure 5: Tree diagram showing clustering process of all 16 analytes based on the correlations of the log-transformed concentrations.

Table 7. Summary Table of Analyte Clusters Formed and their Similarity Values

Analyte Name/No.		Other Boundary Of Cluster (Analyte No.)	Number of Analytes in Cluster	Similarity Value When Cluster Formed
2378-TCDF	1*	9	16	0.37
12378-PeCDF	3*	1	2	0.45
2378-TCDD	2	15	13	0.59
23478-PeCDF	4	8	10	0.82
123478-HxCDF	6	12	3	0.86
123678-HxCDF	7	6	2	0.94
1234678-HpCDF	12	4	4	0.85
12378-PeCDD	5	11	3	0.88
123478/123678-HxCDD	10	5	2	0.92
123789-HxCDD	11	4	7	0.84
1234678-HpCDD	14	16	2	0.86
OCDD	16	4	9	0.83
234678-HxCDF	8	2	11	0.72
1234789-HpCDF	13*	2	12	0.70
OCDF	15*	1	15	0.44
123789-HxCDF	9*	1	16	0.37

Notes: Similarity values above .80 are between the two solid lines

Analytes marked by an * had levels below detection limit in over 25% of the specimens (see Table 1)

Results are based on correlations between log-transformed concentration levels

(Results in Table 7 correspond to those shown in Figure 5)

A joint interpretation of the tree diagram (Figure 5) and the similarity table (Table 7) reveals the following.

- Three two-analyte clusters with high correlations were formed. These are:

  1,2,3,4,7,8- and 1,2,3,6,7,8-HxCDF;

  1,2,3,7,8-PeCDD and 1,2,3,4,7,8-/1,2,3,6,7,8-HxCDD pair; and

  1,2,3,4,6,7,8-HpCDD and OCDD.

  These strong groupings were also observed from the bivariate statistical analysis (see Section III) based on simple correlations. These high positive correlations between analytes within a pair indicate that the two analytes occur simultaneously and are either both high or both low in concentration levels. Thus, tracking the levels of one analyte of a pair is practically equivalent to tracking the other.

Further clusters that are noteworthy will consist of at least three analytes.

- Two three-analyte clusters with a high similarity value each emerged from the process. These are:

  1,2,3,7,8-PeCDD, 1,2,3,4,7,8-/1,2,3,6,7,8-HxCDD pair, and 1,2,3,7,8,9-HxCDD (similarity value of 0.88); and
  1,2,3,4,7,8- and 1,2,3,6,7,8-HxCDF and 1,2,3,4,6,7,8-HpCDF (similarity value of 0.86).

- Next, one cluster consisting of four analytes is formed when 2,3,4,7,8-PeCDF joins the three-analyte cluster of 1,2,3,4,7,8- and 1,2,3,6,7,8-HxCDF and 1,2,3,4,6,7,8-HpCDF with a similarity value of 0.85.

- The four-analyte cluster then is joined with the other three-analyte cluster and with a similarity value of 0.84 between them.

- The clustering process continues until all but the five PCDFs with high occurrences of nondetects and 2,3,7,8-TCDD are joined. That cluster (denoted by the largest triangle of solid lines in Figure 5) contains 10 analytes with a similarity value of 0.82 when it was formed.

- 2,3,7,8-TCDD and the remaining five PCDFs are kept as single analyte clusters since they do not correlate highly with each other (see also Table 3) or with any other cluster formed earlier in the process. The presence of the five PCDFs in this category can be explained by the large number of specimens with levels below the detection limit. In this case, only small variations are seen in concentration levels across specimens which contribute little information to the data base. 2,3,7,8-TCDD, detected in 98.5% of the specimens, exhibits a behavior that is altogether unlike that of the other analytes.

If one were to select an a priori number of clusters by drawing a vertical line through the upper right-hand corner of the large triangle (solid line) in Figure 5, then one would obtain six disjoint clusters. The final results of this cluster analysis are depicted in Figure 6. This figure is another way of presenting the information contained in Table 7 and Figure 5.

3. Interpretation

The results of the multivariate cluster analysis clearly demonstrate that although 2,3,7,8-TCDD was measured at quantifiable levels in practically all samples of the study design, there is no definitive correlation between its incidence of detection and concentration and the other dioxin and furan congeners that were predominant in the human adipose tissue specimens. The lack of correlation of 2,3,7,8-TCDD with other PCDDs and PCDFs may result from several factors which include the sources of exposure and the half-life of the specific dioxin and furan congeners in the human body. Data to support either or both of these factors are limited. However, it is recognized that major sources of exposure include commercial products (such as Agent Orange, trichlorophenol, pentachlorophenol, etc.) and incineration emissions. The levels of specific PCDDs and PCDFs vary within specific chlorinated commercial products. For example, 2,3,7,8-TCDD was the major dioxin contaminant in Agent Orange, whereas the hexa- through octachlorodioxins and furans are the predominant contaminants for a product such as pentachlorophenol.

Incineration emissions containing the 2,3,7,8-substituted compounds have been implicated as another potential source of exposure. However, before correlations between emission and body burden levels can be either determined or disproved, further work will be necessary. As a minimum, such efforts would consist of compiling a number of discrete but detailed data bases from the available literature and investigating relationships among them.

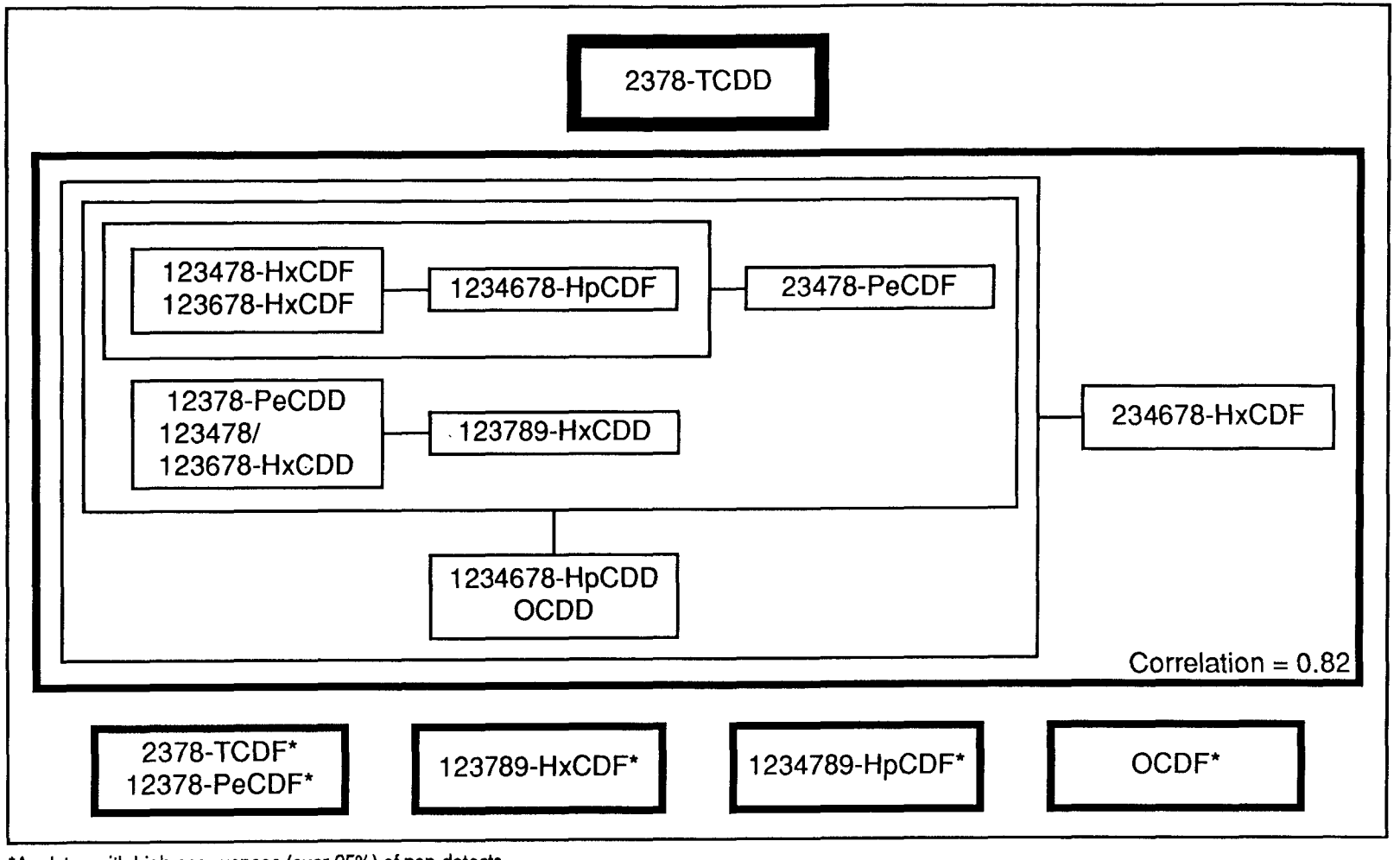
The half-life estimates in the human body for dioxins and furans have been limited solely to the 2,3,7,8-TCDD. The Centers for Disease Control Ranch Hand study (Pirkle et al. 1989) has resulted in an estimate of half-life for this compound at approximately 5 to 10 years. Again, since no additional half-life information is available for the other compounds, correlations with the observed body burden data are not possible.

B. Principal Component Analysis

Principal component analysis (PCA) was performed on the global data set consisting of analytical and demographic data. The following subsections present a brief description of the approach taken, the results, and the interpretation of these results.

1. Approach

In PCA, as in cluster analysis, one concentrates on relationships within a set of variables. Whereas cluster analysis is a tool for determining whether groups of analytes with similar behavior exist, PCA is a technique



*Analytes with high occurrences (over 25%) of non-detects

Notes: The relative position of the analytes in this figure is for illustrative purposes only.

Each rectangle outlines a cluster. One cluster may consist of sub-clusters and individual analytes.

The 6 clusters outlined by heavy dark lines are those formed when a 0.80 cutoff is selected for the similarity measure between clusters.

The 10 analytes within the large rectangle are all highly inter-correlated.

The 4 rectangles at the bottom contain those analytes detected in less than 75% of the specimens.

The cluster on the top shows 2378-TCDD, which does not significantly correlate with any other analyte or group of analytes.

Figure 6. Cluster analysis results of all 16 analytes based on the correlations of the log-transformed concentrations.

used to reduce the dimensionality of the original set of variables in a different manner. The objective of PCA is to derive a few new components as linear combinations of the original variables which will provide a description of the structure of the data with as little loss of information as possible.

Various subsets of the data were subjected to PCA. All analyses were performed using the principal component analysis option of the FACTOR procedure followed by the SCORE procedure from the Statistical Analysis System (SAS) package. Also, a subroutine was added to enhance the output by ranking the variables by their percent contributed variance within each principal component. The analyses were performed on the correlation matrix; i.e., the variables were autoscaled (standardized) so as to be treated with equal importance. All concentration data were log-transformed prior to PCA.

2. Results

Of the 16 analytes in the data base, those with less than 25% of nondetects were excluded; a total of 11 analytes remained. The list of the five PCDFs excluded is given in Section III.A. The demographic variables considered at first were individual's age at sampling, specimen collection year, and body mass index. The inclusion of body mass index reduced the number of specimens to 169, since body mass index was available for only that many.

The results, in terms of significant principal components, showed that body mass index, for all practical purposes, did not contribute to explaining any of the variation in the data set. Thus, the decision was made to eliminate that variable from subsequent analyses. A total of 11 analytes (log-transformed concentrations) and 2 demographic variables, age at sampling and specimen collection year, and a total of 195 samples were kept for the final analyses described below.

Three principal components were kept for further examination. The first and most predominant principal component explains 64.4% of the total variance in the data set. The second principal component explains 12.7% of the total variance and the third principal component contributes 6.39% of the variance. Thus, the three principal components account for 83.5% of the total variance in the data, with the first two alone accounting for a total of 77.1%.

The profiles of the three principal components are presented in Table 8 along with the variable loadings and their individual contributions (percent variance) to the variance explained by the given principal component. Within each principal component, the variables are listed in order of decreasing absolute loadings; a minus sign preceding a loading denotes a negative correlation of that variable with the principal component. The fourth column lists cumulative percentages and thus adds up to 100% within each principal component. The last part of Table 8, entitled "Final Communality Estimates" is explained next.

The eigenvalues associated with these three principal components are 8.37, 1.65, and 0.83, respectively. Of a possible total of 13 (number of variables included in the PCA), these correspond to the proportions of 64.4%,

Table 8. Principal Component Analysis Results (cont.)

Principal Component 1 (64.4% of total variance explained)
"Higher Chlorinated Furans and Dioxins"

Variable (Analyte No.)*1 (ordered by % variance)		Variable Loading	Variance Explained (%)	Cumulative Variance (%)
123478-HxCDF	(6)	0.94	10.55	10.55
123789-HxCDD	(11)	0.92	10.02	20.58
123678-HxCDF	(7)	0.91	9.99	30.56
123478/123678-HxCDD	(10)	0.90	9.71	40.28
OCDD	(16)	0.88	9.20	49.48
1234678-HpCDF	(12)	0.87	8.98	58.46
1234678-HpCDD	(14)	0.86	8.83	67.29
12378-PeCDD	(5)	0.85	8.71	76.00
23478-PeCDF	(4)	0.83	8.28	84.28
234678-HxCDF	(8)	0.78	7.23	91.50
2378-TCDD	(2)	0.78	7.20	98.70
Collection Year		-0.33	1.29	100.00
Age at Sampling		-0.02	0.00	100.00

Principal Component 2 (12.7% of total variance explained)
"Age and Collection Year"

Variable (Analyte No.)*1 (ordered by % variance)		Variable Loading	Variance Explained (%)	Cumulative Variance (%)
Age at Sampling		0.92	51.68	51.68
Collection Year		0.84	42.47	94.15
1234678-HpCDF	(12)	-0.16	1.57	95.72
123478/123678-HxCDD	(10)	0.14	1.19	96.91
23478-PeCDF	(4)	0.14	1.10	98.01
123678-HxCDF	(7)	0.12	0.93	98.95
12378-PeCDD	(5)	0.10	0.65	99.60
123789-HxCDD	(11)	0.06	0.19	99.79
1234678-HpCDD	(14)	-0.04	0.10	99.89
2378-TCDD	(2)	-0.03	0.06	99.95
OCDD	(16)	0.02	0.04	99.98
123478-HxCDF	(6)	-0.01	0.01	99.99
234678-HxCDF	(8)	-0.01	0.01	100.00

Note 1*: All concentrations were log-transformed
Note 2: Shaded area within each set highlights variables with a variance contribution above average (1/13 or 7.7%)

Table 8. Principal Component Analysis Results (concl.)

Principal Component 3 (6.39% of total variance explained)
"Lower Chlorinated Furans and Dioxins"

Variable (Analyte No.)*1 (ordered by % variance)		Variable Loading	Variance Explained (%)	Cumulative Variance (%)
234678-HxCDF	(8)	0.43	22.09	22.09
1234678-HpCDD	(14)	0.37	16.85	38.94
12378-PeCDD	(5)	-0.37	16.73	55.67
2378-TCDD	(2)	-0.37	16.62	72.30
1234678-HpCDF	(12)	0.22	5.62	77.92
OCDD	(16)	0.21	5.56	83.48
Collection Year		0.21	5.51	89.00
23478-PeCDF	(4)	-0.21	5.32	94.32
123478/123678-HxCDD	(10)	-0.21	5.07	99.39
Age at Sampling		-0.05	0.33	99.73
123678-HxCDF	(7)	0.04	0.16	99.89
123789-HxCDD	(11)	-0.03	0.11	100.00
123478-HxCDF	(6)	0.00	0.00	100.00

Final Communality Estimates

		Final Communality Estimates			
Variable (Analyte No.)*1		PC1	PC1+PC2	PC1+PC2+PC3	
2378-TCDD	(2)	0.60	0.60	0.74	*
23478-PeCDF	(4)	0.69	0.71	0.76	
12378-PeCDD	(5)	0.73	0.74	0.88	*
123478-HxCDF	(6)	0.88	0.88	0.88	
123678-HxCDF	(7)	0.84	0.85	0.85	
234678-HxCDF	(8)	0.60	0.60	0.79	*
123478/123678-HxCDD	(10)	0.81	0.83	0.87	
123789-HxCDD	(11)	0.84	0.84	0.84	
1234678-HpCDF	(12)	0.75	0.78	0.82	
1234678-HpCDD	(14)	0.74	0.74	0.88	*
OCDD	(16)	0.77	0.77	0.82	
Age at Sampling		0.0004	0.86	0.86	
Collection Year		0.11	0.81	0.86	
Total Communality		8.37	10.02	10.85	
Percent Variance Explained		64.4	77.1	83.5	

* next to last column marks those variables showing largest improvement in final communality estimates from 2 to 3 principal components

Note 1*: All concentrations were log-transformed

Note 2: Shaded area within each set highlights variables with a variance contribution above average (1/13 or 7.7%)

12.7%, and 6.39%, respectively. Generally, principal components with eigenvalues less than 1.0 are not considered in the interpretation of PCA results. Past experience indicates that principal components with eigenvalues below 1.0 explain little correlated behavior among the measured variables and that they often reflect random variance contributions due to measurement and sampling error.

This reasoning should prompt us to discard the third principal component with a low eigenvalue of 0.83. However, other information in the PCA results was considered in the selection of the final number of principal components.

Each selection of a number of principal components has associated with it a set of final communalities for the 13 variables considered. They are calculated as the sum of the squares of the loadings (correlations) of a given variable on the principal components. Final communality estimates are the multiple correlations for predicting the variables from the estimated principal components and will therefore be limited by 1. They show how well the variables considered in the PCA are accounted for by the number of components selected. Thus, if one were to consider all 13 principal components generated by the PCA on the 13 variables, then the final communality estimates would each be 1 and therefore add up to 13, explaining 100% of the total variance in the data.

The final communality estimates of each variable were carefully examined in the successive consideration of one, two, and three principal components. These results are shown in the last part of Table 8. As shown in the column entitled "PC1+PC2", it is clear that 2,3,7,8-TCDD and 2,3,4,6,7,8-HxCDF with final communalities of 0.60 each, are "poorly" accounted for by a two-principal component model. Adding the third principal component to the model considerably increases the final communalities (compare column 3 with column 4 entitled "PC1+PC2+PC3") for these two analytes, as well as those of 1,2,3,7,8-PeCDD and 1,2,3,4,6,7,8-HpCDD. Therefore, it was found that the third principal component is important to the interpretation of the data in spite of its low eigenvalue (0.83) and that a three-principal component model should be considered.

3. Interpretation

Overall, three principal components, explaining approximately 84% of the total variance, were judged to be important in the interpretation of the data. The selection of three rather than two principal components was made in spite of the low eigenvalue of the third component which contributes little (6.4%) to the variance in the data. The reason for its inclusion was discussed above.

Since the PCA was performed with 13 variables, variables with contributions above 7.7% (=1/13), the average relative contribution should all variables be equally important within a given principal component, were considered significant. This cutoff is indicated by the shaded area in each set in Table 8.

Principal component 1, which explains 64.4% of the total variance in the data set, consists mainly of the higher chlorinated furans and dioxins. This principal component is purely a chemical principal component--the first 9 of the 13 variables considered account for about 84% of the variance explained by this principal component. These nine analytes each contribute an above-average (1/13 or 7.7%) share of the variation that can be explained by principal component 1; they have about equal loadings ranging from 0.94 to 0.83. Principal component 1 was thus labeled the "Higher Chlorinated PCDDs and PCDFs" component. This labeling is used on all PCA plots. Due to the loading pattern, this principal component is approximately the sum of the concentrations of these nine analytes.

An interesting fact should be noted in relationship to the outcome of the cluster analysis (see Figures 5 and 6). In these figures, the list of analytes in the large cluster of nine analytes (similarity of 0.83) is identical to the list of the nine analytes listed in order in the first principal component in Table 8. This fact is not surprising. Considering that principal components are orthogonal to each other--that is, variables contributing significantly to one principal component do not to another and vice versa--one would expect that highly correlated analytes, such as those seen in the large cluster from Figure 7, would all contribute to the same principal component. Were this not the case, then some of the analytes would significantly contribute to the next principal component, meaning that some of the remaining variance could be explained by these analytes. This would be in contradiction with the fact that these analytes are highly correlated.

The second principal component, which explains considerably less (12.7%) of the total variance than does the first component, is interesting in the fact that it is a purely "non-chemical" component. The two variables, age of subject at time of death or surgery and specimen collection year, account for 94% of that principal component. This principal component was naturally labeled the "Age and Collection Year" component. It was shown earlier that these variables, taken one at a time, contributed very little to the variability in 2,3,7,8-TCDD levels. It is shown here, however, that considering all chemical and demographic variables simultaneously, can improve the predictive power of a combination of variables with respect to the whole data set. It should be noted, though, that an overall contribution of 12.7% leaves considerable room for unexplained variance, were one to look at principal component 2 only.

The third principal component, which explains only a small proportion (6.4%) of the overall variance in the data, consist mainly of four analytes. Two of these, 2,3,4,6,7,8-HxCDF (positive loading) and 2,3,7,8-TCDD (negative loading), were those not contributing above average (1/13 or 7.7% of variance) to the first principal component. However, the other two, 1,2,3,4,6,7,8-HpCDD (positive loading) and 1,2,3,7,8-PeCDD (negative loading), were contributing to the first component as well. It is for these 4 analytes that the addition of the third principal component resulted in the largest increase in final communalities (see last part of Table 8.) This principal component was labeled the "Lower Chlorinated PCDDs and PCDFs" component.

It should be noted that the third principal component is the only component to which 2,3,7,8-TCDD contributes a considerable amount of variation (16.62%). However, the third principal component contributes only a low 6.39% to the total variance in the data. Overall, when considering the three principal components described above, 2,3,7,8-TCDD contributes approximately 5.7% to the variation in the whole data set determined by the 195 specimen results. The figure 5.7% is obtained as the sum of the weighted percent variances explained by the three principal components. The weights are the percent variance attributed to 2,3,7,8-TCDD within each principal component. Thus 5.7% is obtained as 7.20x64.4% + 0.06x12.7% + 16.62x6.39%. Of these 5.7%, the largest proportion of variance (4.6% = 7.20x64.4%) attributed to 2,3,7,8-TCDD is contributed by the first principal component. It should not be expected that any pattern in the specimens will be visible along the axis determined by this principal component.

4. Plotting of Principal Component Scores

The three selected principal components were used to compute factor scores for each of the 195 specimens. These standardized scores were calculated by using the respective loadings as the coefficients in the linear combinations of the variables in each principal component. Plotting these scores in the space spanned by the three principal components will provide a means of detecting visible patterns or grouping in the specimens, should they exist. Due to the disproportion of variance explained by the three principal components, the specimen scores were plotted on two principal components at a time, rather than in a three-dimensional figure. The latter approach, in which the third dimension would add only about 6% to the 77% contributed by the first two dimensions, would tend to obscure rather than enhance possible patterns.

Figures 7 through 15 show nine selected principal component plots. These are two-dimensional plots of principal component 2 vs. 1, 3 vs. 1, and 3 vs. 2, each using a different label for encoding the data points. The principal component explaining the highest proportion of variance in a given pair is always plotted on the X-axis. Plots of principal component 2 vs. 1 will display 77.1% of the variation in the data; plots of principal component 3 vs. 1 will display 70.8% of the variation with most of the variation along the first component; and plots of principal component 3 vs. 2 a total of 19.1%.

The axes representing principal components 1, 2, and 3 have been labeled "Higher Chlorinated PCDDs and PCDFs", "Age + Collection Year", and "Lower Chlorinated PCDDs and PCDFs", respectively, on all 24 plots. The 24 principal component plots are shown in four set of six as follows:

- Figures 7 through 9: **military/civilian status** was used as the labeling code.
- Figures 10 through 12: **geographic region** was used as the labeling code.
- Figures 13 through 15: **race** (Caucasian vs. non-Caucasian) was used as the labeling code.

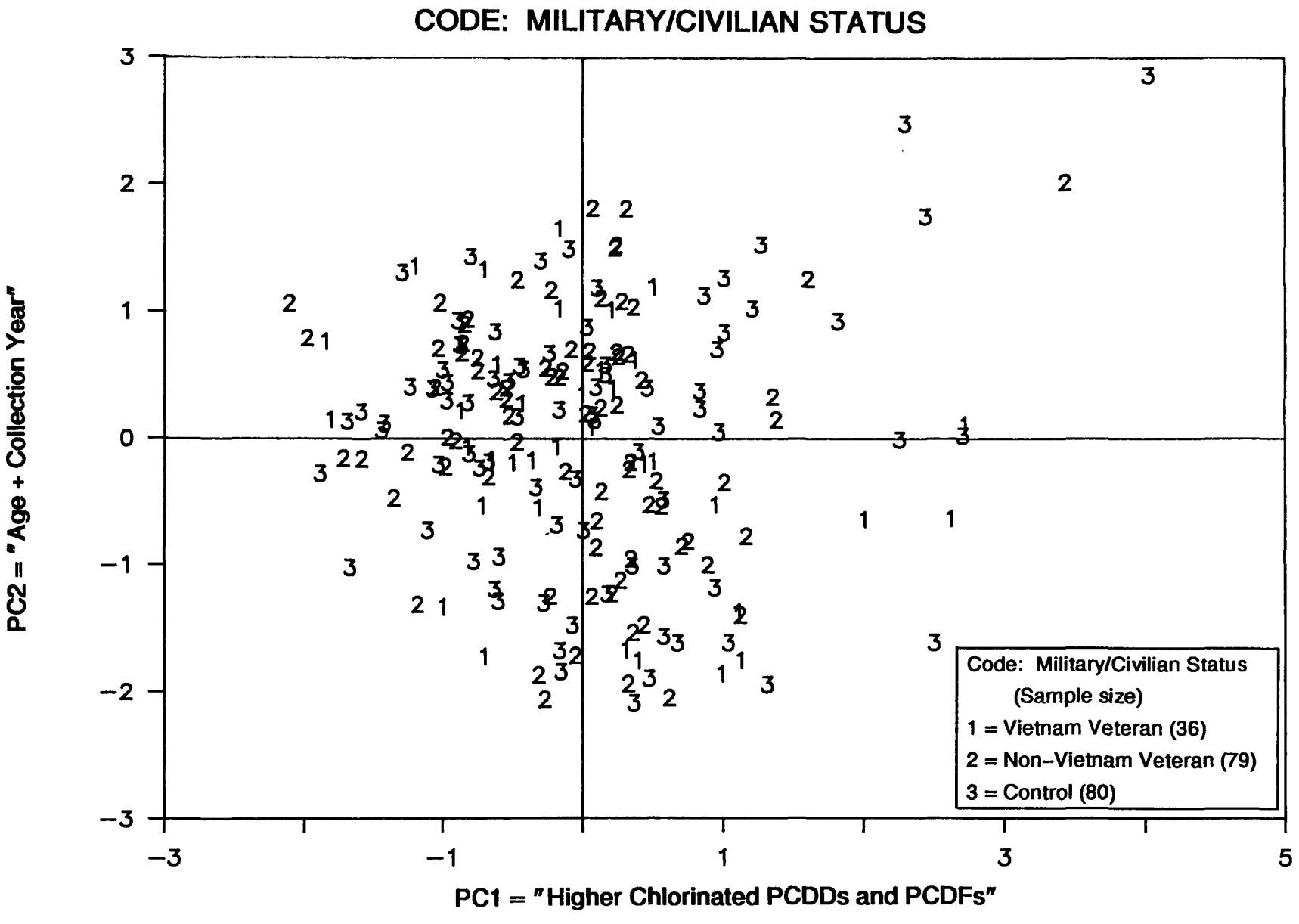


Figure 7. Plot of principal component 2 vs. 1--all 195 specimens coded by military/civilian status.

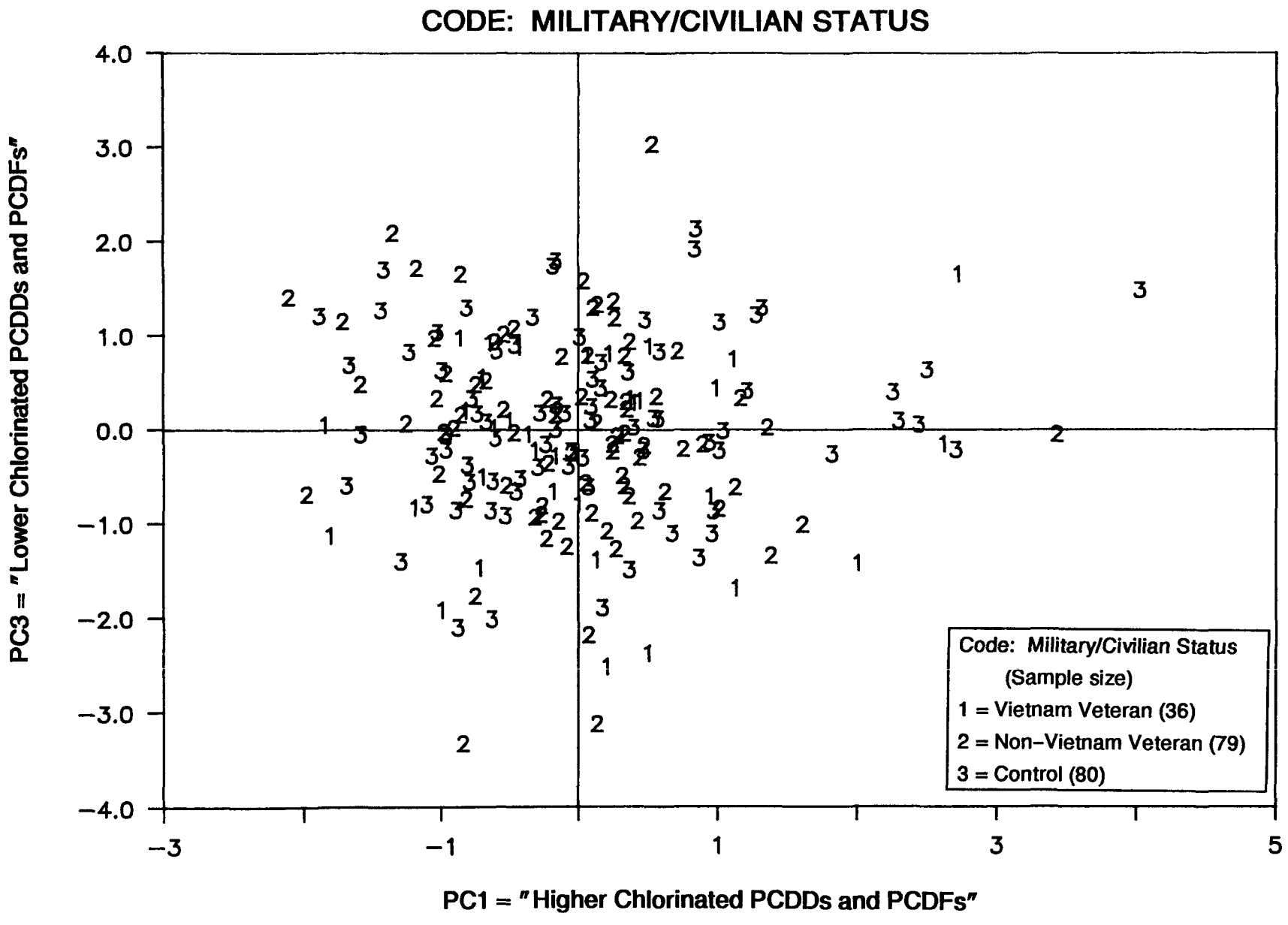


Figure 8. Plot of principal component 3 vs. 1--all 195 specimens coded by military/civilian status.

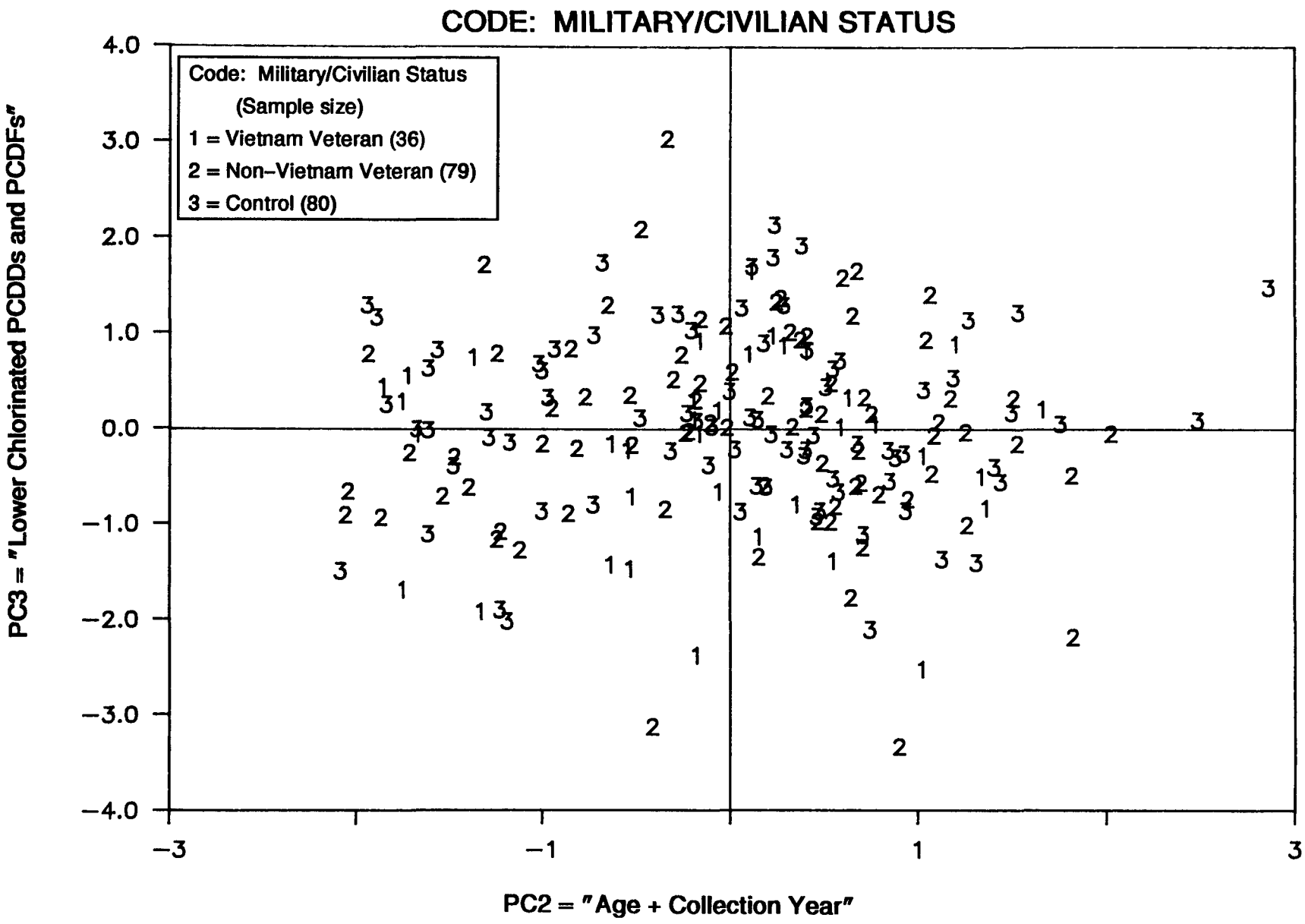


Figure 9. Plot of principal component 3 vs. 2—all 195 specimens coded by military/civilian status.

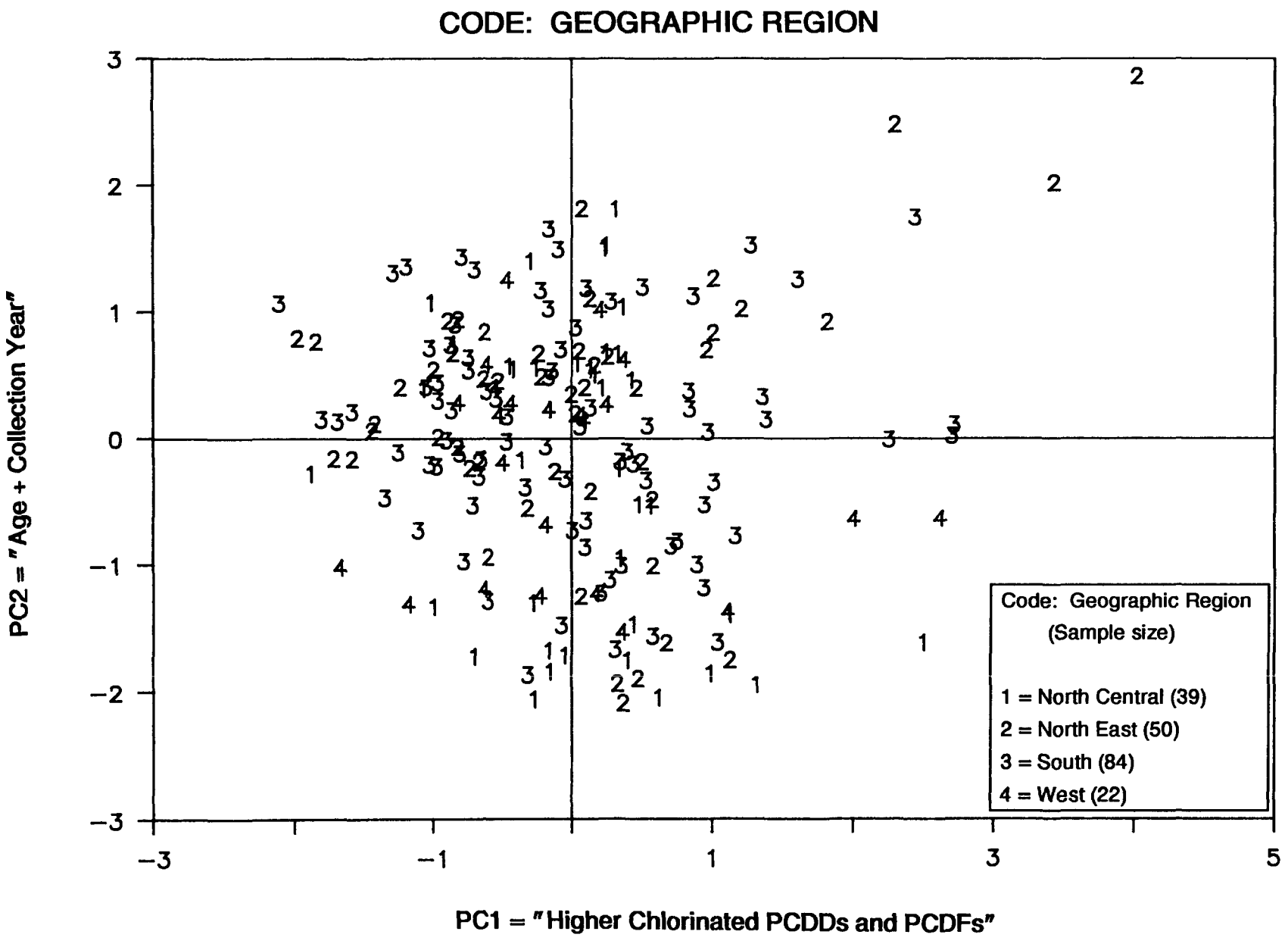


Figure 10. Plot of principal component 2 vs. 1--all 195 specimens coded by geographic region.

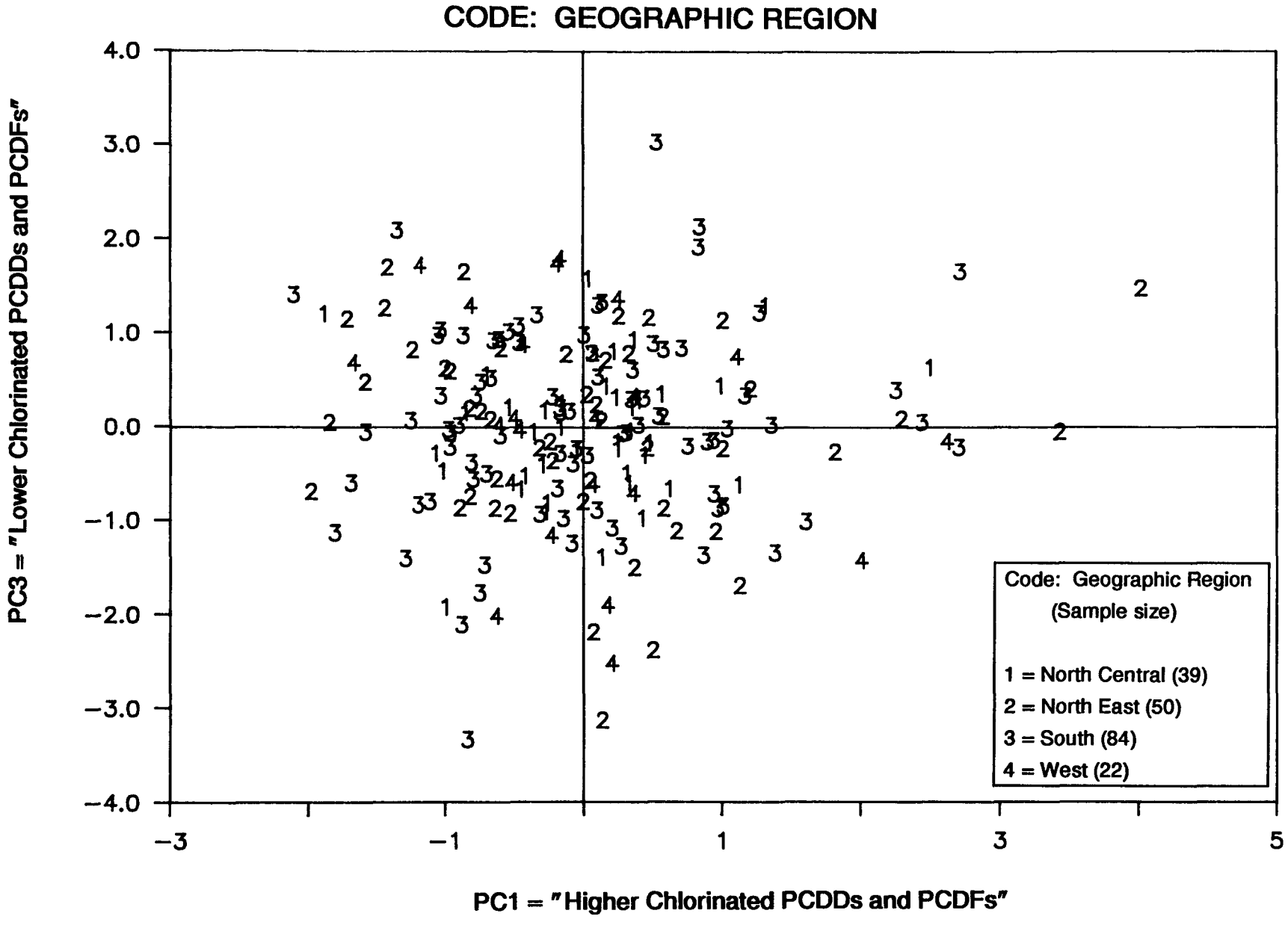


Figure 11. Plot of principal component 3 vs. 1--all 195 specimens coded by geographic region.

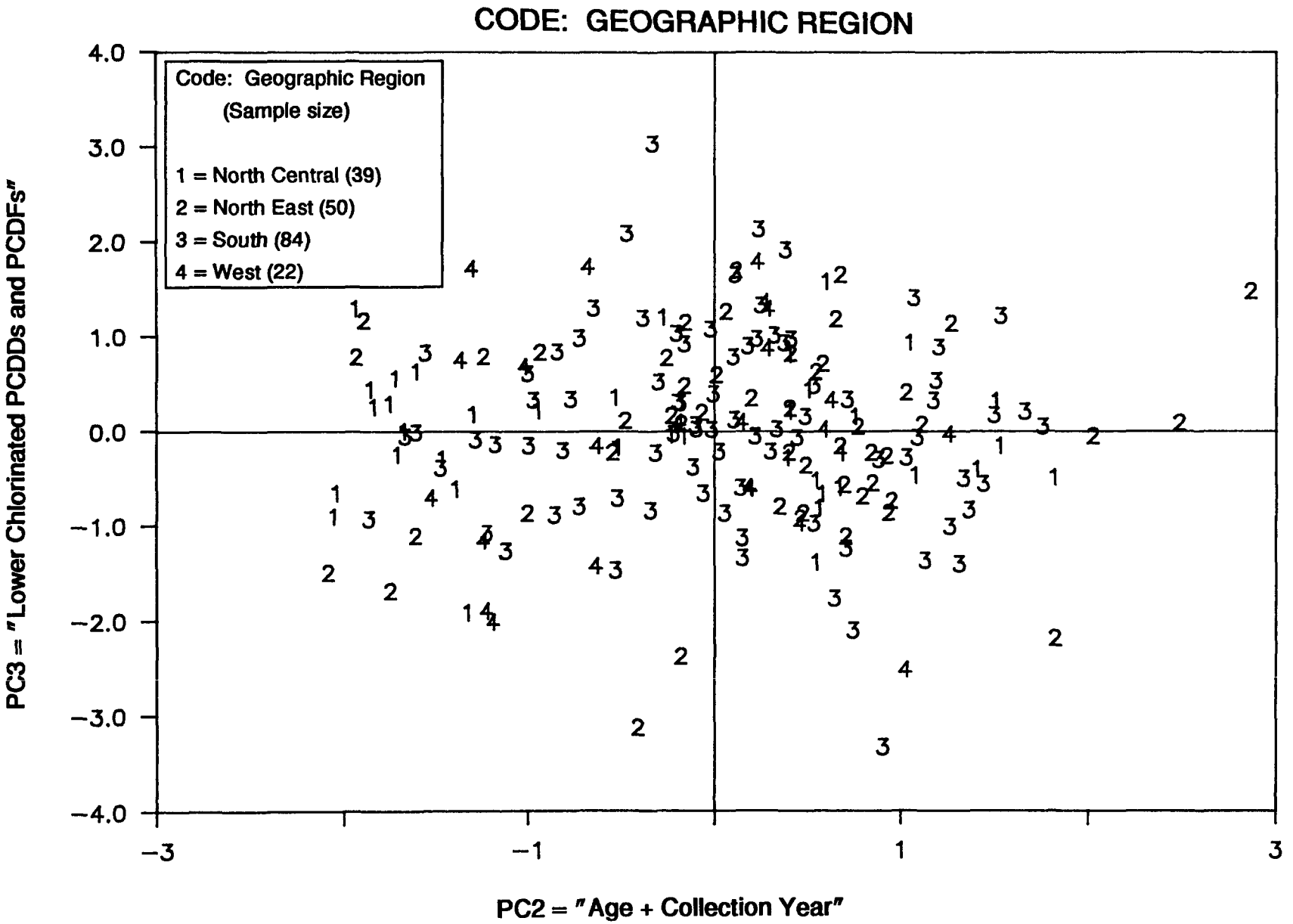


Figure 12. Plot of principal component 3 vs. 2--all 195 specimens coded by geographic region.

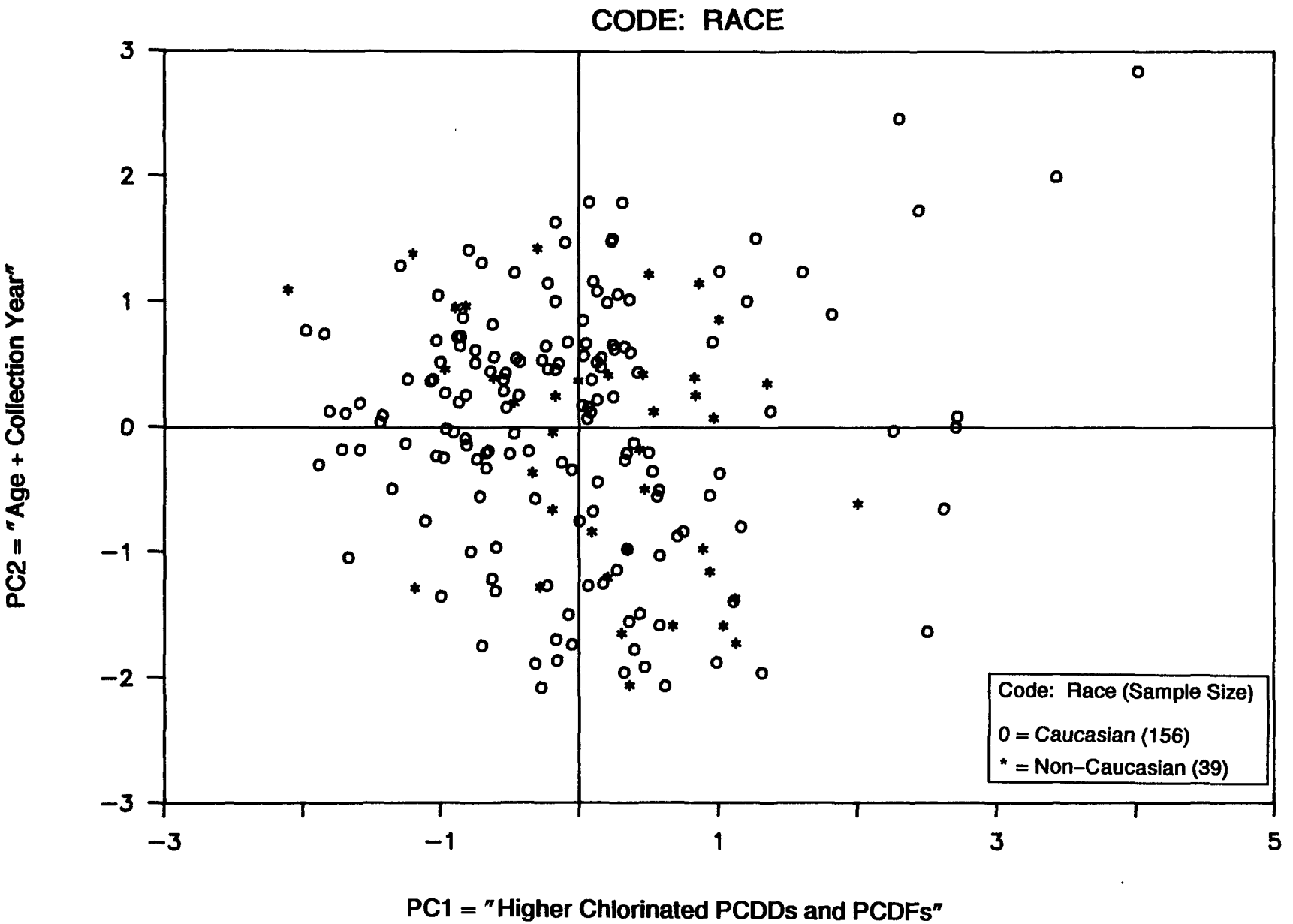


Figure 13. Plot of principal component 2 vs. 1—all 195 specimens coded by race.

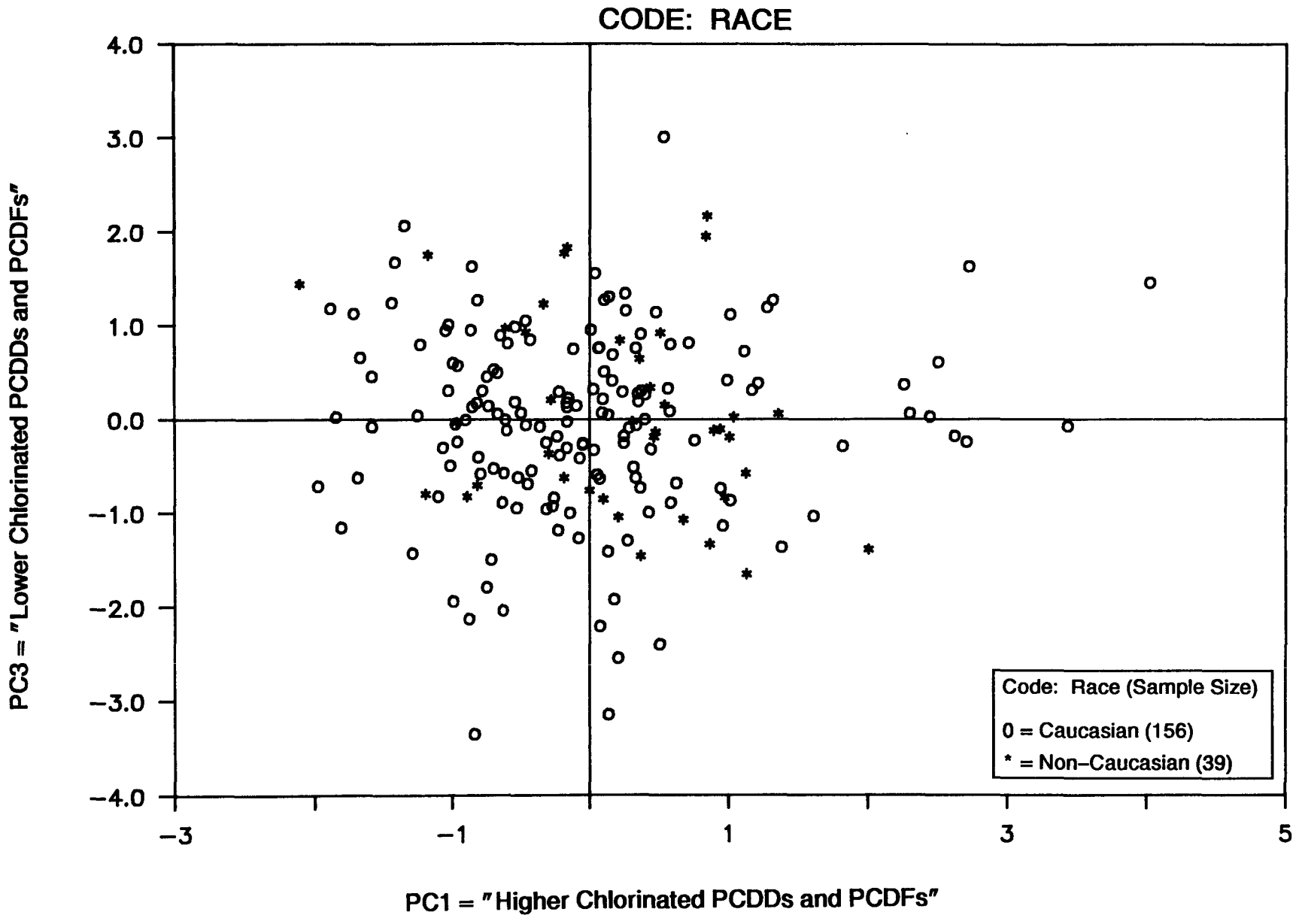


Figure 14. Plot of principal component 3 vs. 1--all 195 specimens coded by race.

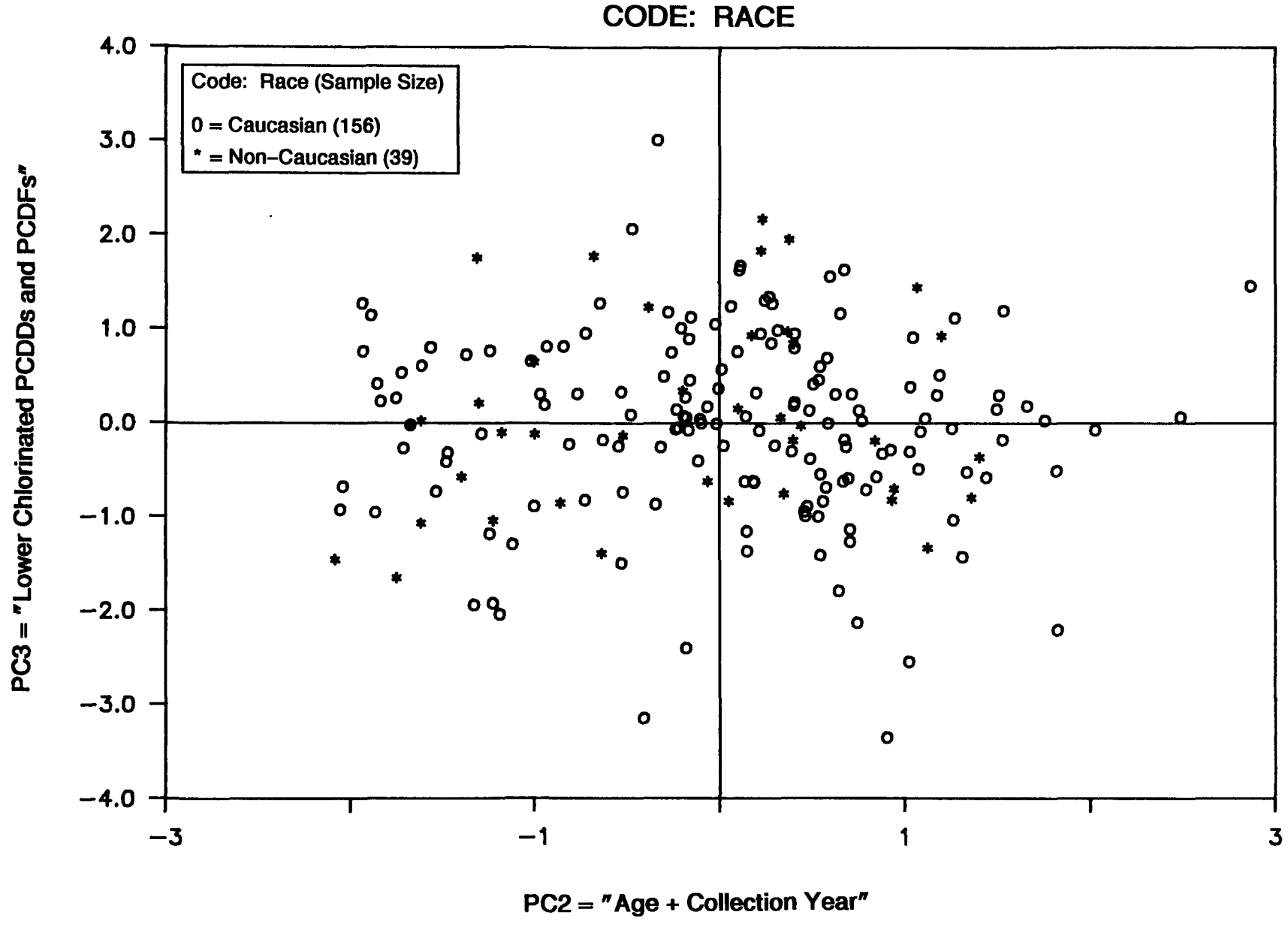


Figure 15. Plot of principal component 3 vs. 2--all 195 specimens coded by race.

Legends and sample sizes within each group as defined by the labeling variables are given in a box on each plot.

The standardized scores of all 195 specimens vary between -2.11 and 4.02 along the first principal component, between -2.09 and 2.86 along the second principal component, and between -3.33 and +3.04 along the third principal component. Very few data points fall outside the -3 to +3 boundaries, in fact, none along principal component 2, an indication that the plots are very "tight." Of the 195 specimens, only 5 have scores falling outside the ±3 range. These five subjects consist of four non-Vietnam veterans and one civilian.

These plots do not provides any clear stratification of the specimens. Particularly Figures 7 through 9, which use military/civilian status as the coding label, demonstrate that the group of Vietnam veterans in this study exhibits no difference from the two other groups of non-Vietnam veterans and civilians.

In summary, the plots of the two most important principal components, the "Higher Chlorinated PCDDs and PCDFs" and the "Age Plus Collection Year" components, explain 77.4% of the total variation in the data set. However, these plots do not reveal any stratification of the specimens into groups, whether defined by military/civilian status of the individual, geographic region, or race.

The result of most interest, based on this set of matched NHATS specimens, is that the group of Vietnam veterans does not exhibit higher levels of PCDDs and PCDFs, including 2,3,7,8-TCDD, than either the group of non-Vietnam veterans or the control group of civilians. However, to the extent that the NHATS population is not representative of Vietnam veterans, the results are not representative of 2,3,7,8-TCDD levels in the general population of Vietnam veterans. The relatively small sample size (36) of Vietnam veterans would limit the significance of results relating to 2,3,7,8-TCDD levels in the population of Vietnam veterans in any case. However, within the restricted NHATS population, the comparative results are valid.

VI. REFERENCES

Dunn JW, De Vault D, Rappe C, Wiberg K, Bergqvist, PA. 1988. The application of SIMCA pattern recognition to dioxin and dibenzofuran residue data in Great Lakes fish. The 8th International Symposium on Chlorinated Dioxins and Related Compounds, Umeå, Sweden, August 21-26, 1988.

Lindström G, Rappe C. 1988. Multivariate data analysis applied in studying the distribution of PCDDs and PCDFs in human milk. The 8th International Symposium on Chlorinated Dioxins and Related Compounds, Umeå, Sweden, August 21-26, 1988.

Pirkle JL, Wolfe WH, Patterson DG, Needham LL, Michalek JE, Miner JC, Peterson MR, Phillips DL. 1989. Estimates of the half-life of 2,3,7,8-tetrachlorodibenzo-*p*-dioxin in Vietnam veterans of Operation Ranch Hand. Journal of Toxicology and Environmental Health 27:165-171.

Pitea D, Bonati L, Lasagni M, Moro G, Todeschini R, Chiesa G. 1988. The combustion of municipal solids wastes: PCDD and PCDF in MSW and in emissions. A chemometric approach. The 8th International Symposium on Chlorinated Dioxins and Related Compounds, Umeå, Sweden, August 21-26, 1988.

Sielken, Jr. RL. 1987. Statistical evaluations reflecting the skewness in the distribution of TCDD levels in human adipose tissue. *Chemosphere* 16:2135-2140, No. 8/9.

Sjström M, Nygren M, Hansson M, Rappe C. 1988. A multivariate comparison of PCDDs and PCDFs levels in Vietnam veterans and controls. The 8th International Symposium on Chlorinated Dioxins and Related Compounds, Umeå, Sweden, August 21-26, 1988.

Tysklind M, Rappe C. 1988. Multivariate data analysis of operation and emission variables from scrap metal melting at a steel mill. The 8th International Symposium on Chlorinated Dioxins and Related Compounds, Umeå, Sweden, August 21-26, 1988.

USEPA. 1986. U.S. Environmental Protection Agency, Office of Toxic Substances. Analysis for polychlorinated dibenzo-*p*-dioxins and dibenzofurans in human adipose tissue: method evaluation study. Washington, DC. USEPA. EPA 560/5-85-022.

VA and USEPA. 1989. The Department of Veterans Affairs, Office of Environmental Epidemiology, and the U.S. Environmental Protection Agency, Office of Toxic Substances. Dioxins and dibenzofurans in adipose tissue of U.S. Vietnam Veterans and controls. Washington, DC. USEPA. EPA 560/5-89-002.

Wold S. 1988. An overview of multivariate data analysis. The 8th International Symposium on Chlorinated Dioxins and Related Compounds, Umeå, Sweden, August 21-26, 1988.

BMDP Statistical Software. 1983. 1985 Printing. University of California Press, Berkeley.

SAS: Statistical Analysis System, SAS Institute, Inc.
SAS® User's Guide: Basics, Version 5 Edition, 1985.
SAS® User's Guide: Statistics, Version 5 Edition, 1985.

Lotus 1-2-3 Release 2.01. 1986. Lotus Development Corporation, 55 Cambridge Parkway, Cambridge, Massachusetts 02142.

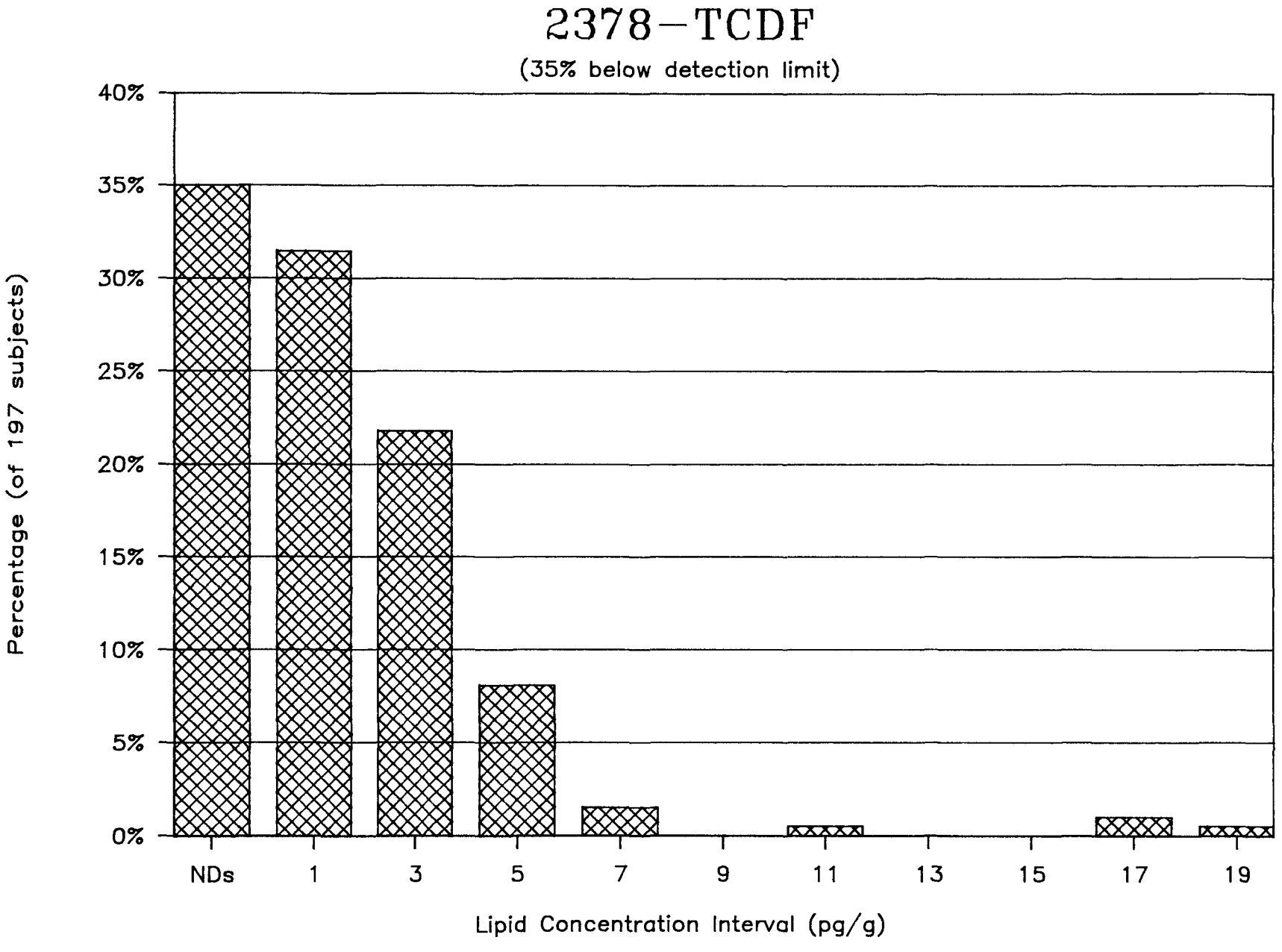
APPENDIX

HISTOGRAMS OF CONCENTRATION LEVELS OF ALL 16 ANALYTES

HISTOGRAMS OF CONCENTRATION LEVELS OF ALL 16 ANALYTES

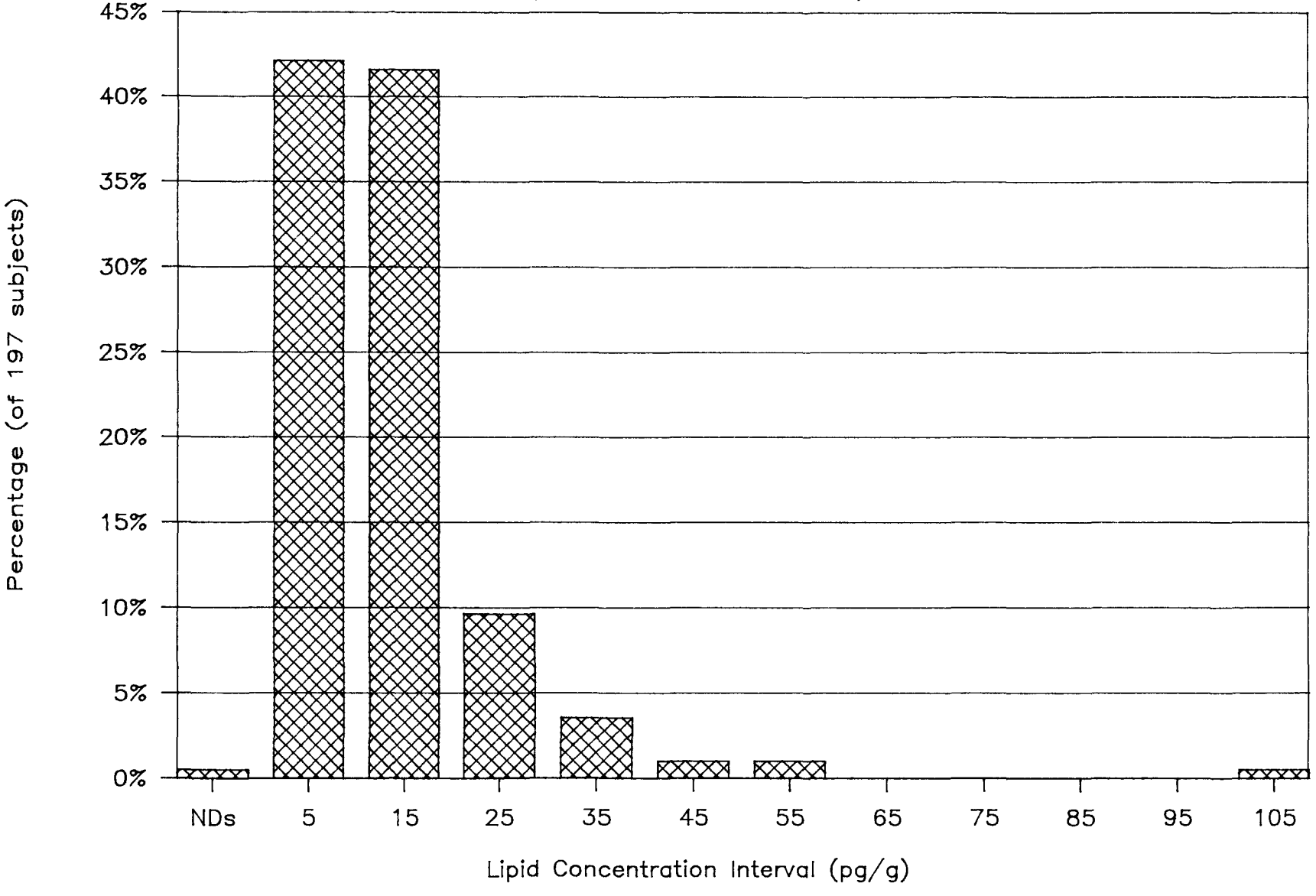
Each vertical bar on a histogram represents the percentage of the total number of specimens whose concentrations fall within a given concentration range. The first bar in each histogram indicates the percentage of specimens with concentrations below detection limit, if applicable. The ranges are represented by their midpoints on the graphs. The width of each range is consistent within a given histogram. However, they change from histogram to histogram, depending on the concentration range of any given analyte.

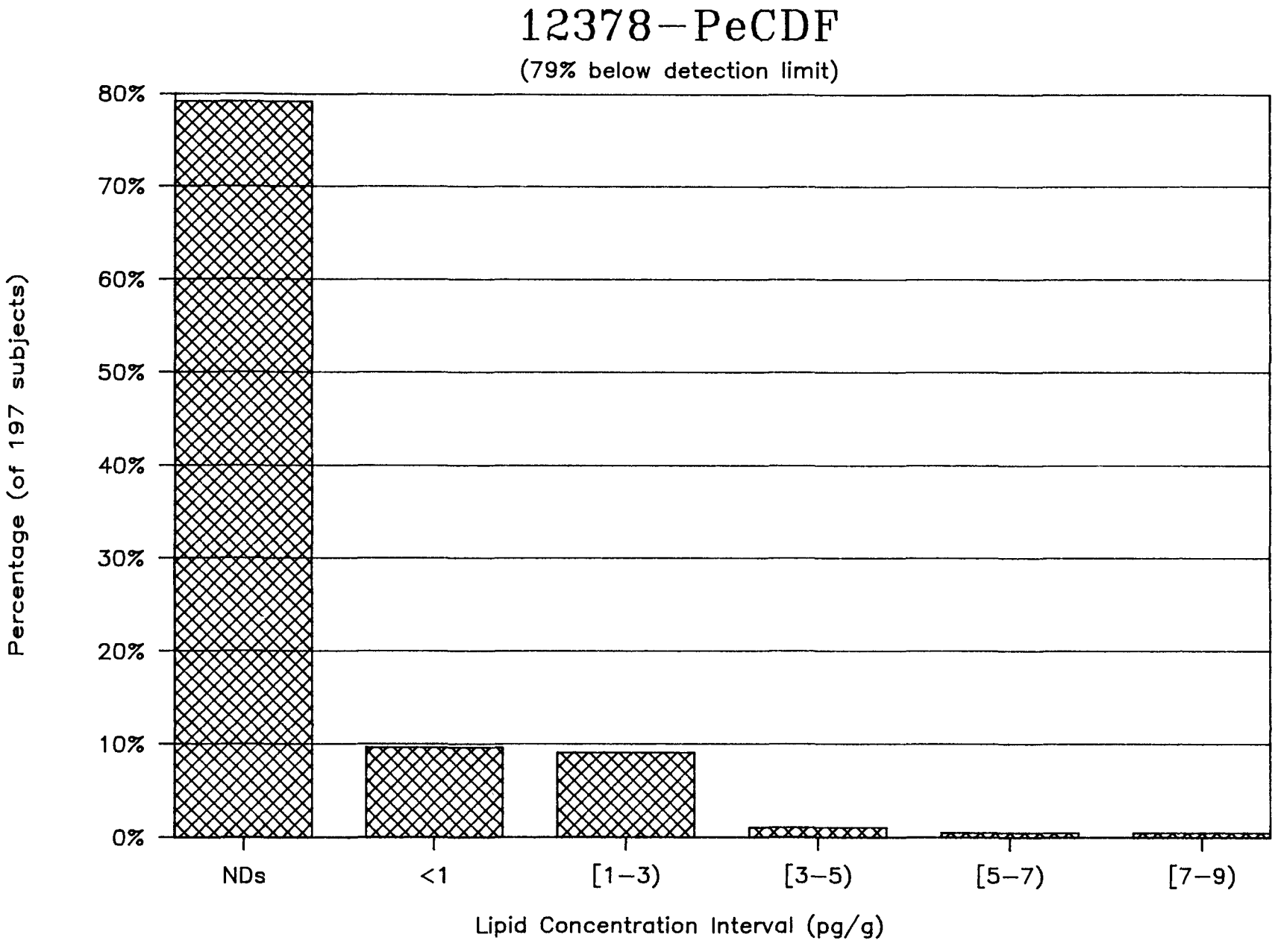
To determine the width of a range, simply subtract two consecutive midpoints from each other. The ranges are defined in such a way that the lower end concentration is included while the upper end is excluded. Using 2,3,7,8-TCDD as an example, there are approximately 43% of the subjects with concentrations above detection limit but below 10 pg/g; about 42% of the subjects have concentrations equal to or greater than 10 pg/g but less than 20 pg/g, etc.

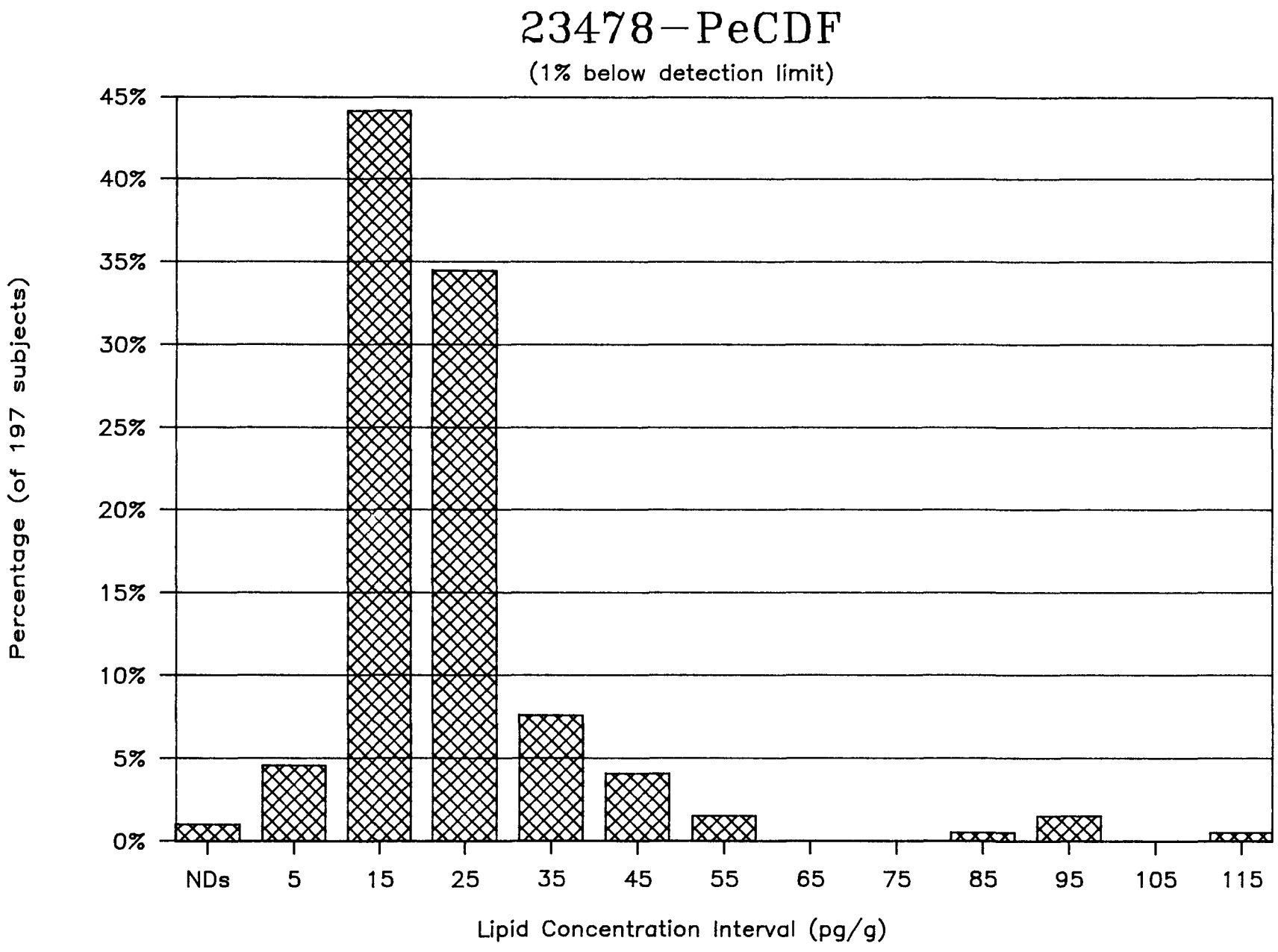

2378-TCDF
(35% below detection limit)
40%
35%
30%
25%
20%
15%
10%
5%
0%
Percentage (of 197 subjects)
NDs
1
3
5
7
9
11
13
15
17
19
Lipid Concentration Interval (pg/g)

2378-TCDD

(0.5% below detection limit)

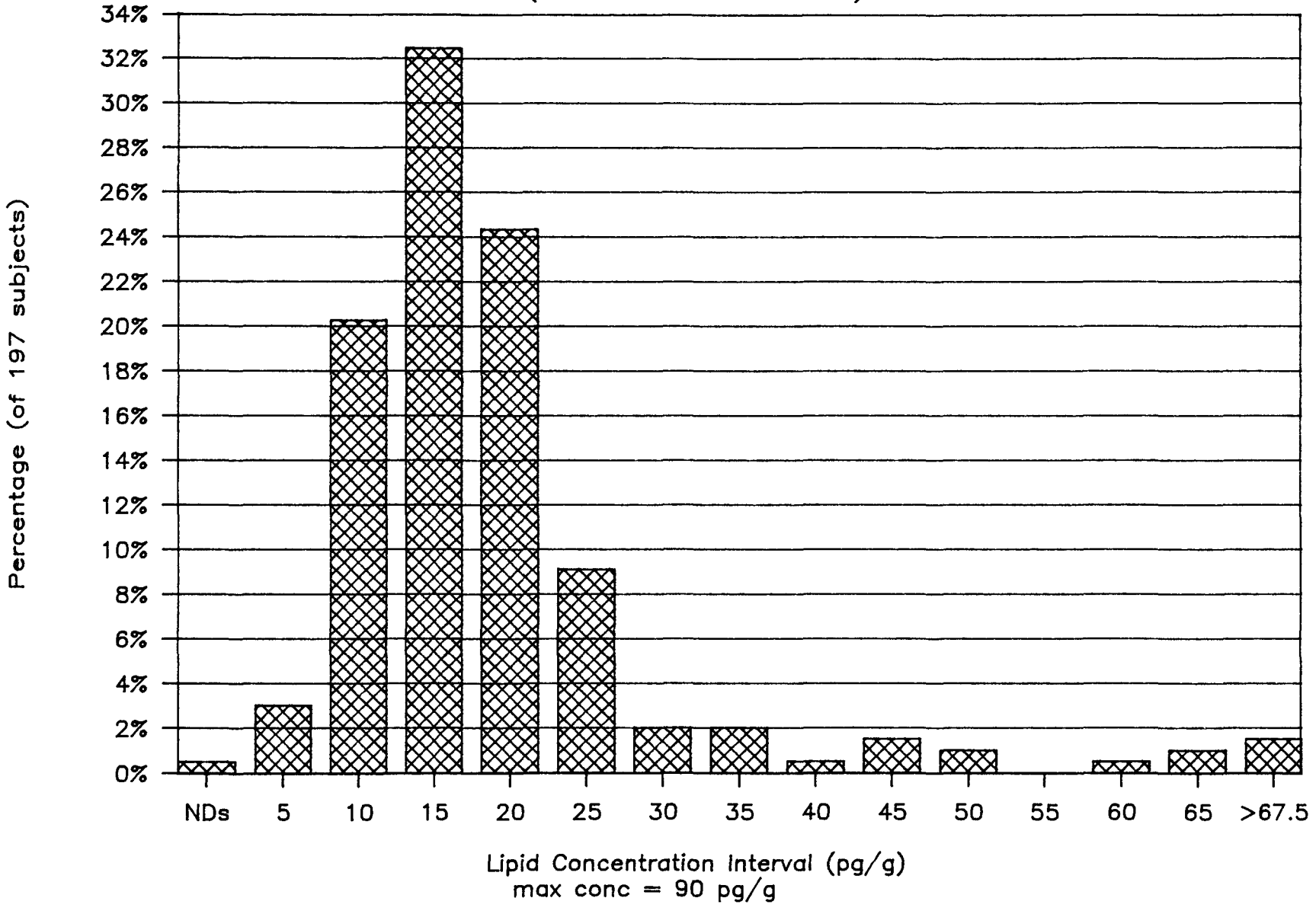



12378-PeCDF
(79% below detection limit)
80%
70%
60%
50%
40%
30%
20%
10%
0%
Percentage (of 197 subjects)
NDs
<1
[1-3)
[3-5)
[5-7)
[7-9)
Lipid Concentration Interval (pg/g)


23478-PeCDF
(1% below detection limit)
45%
40%
35%
30%
25%
20%
15%
10%
5%
0%
Percentage (of 197 subjects)
NDs
5
15
25
35
45
55
65
75
85
95
105
115
Lipid Concentration Interval (pg/g)

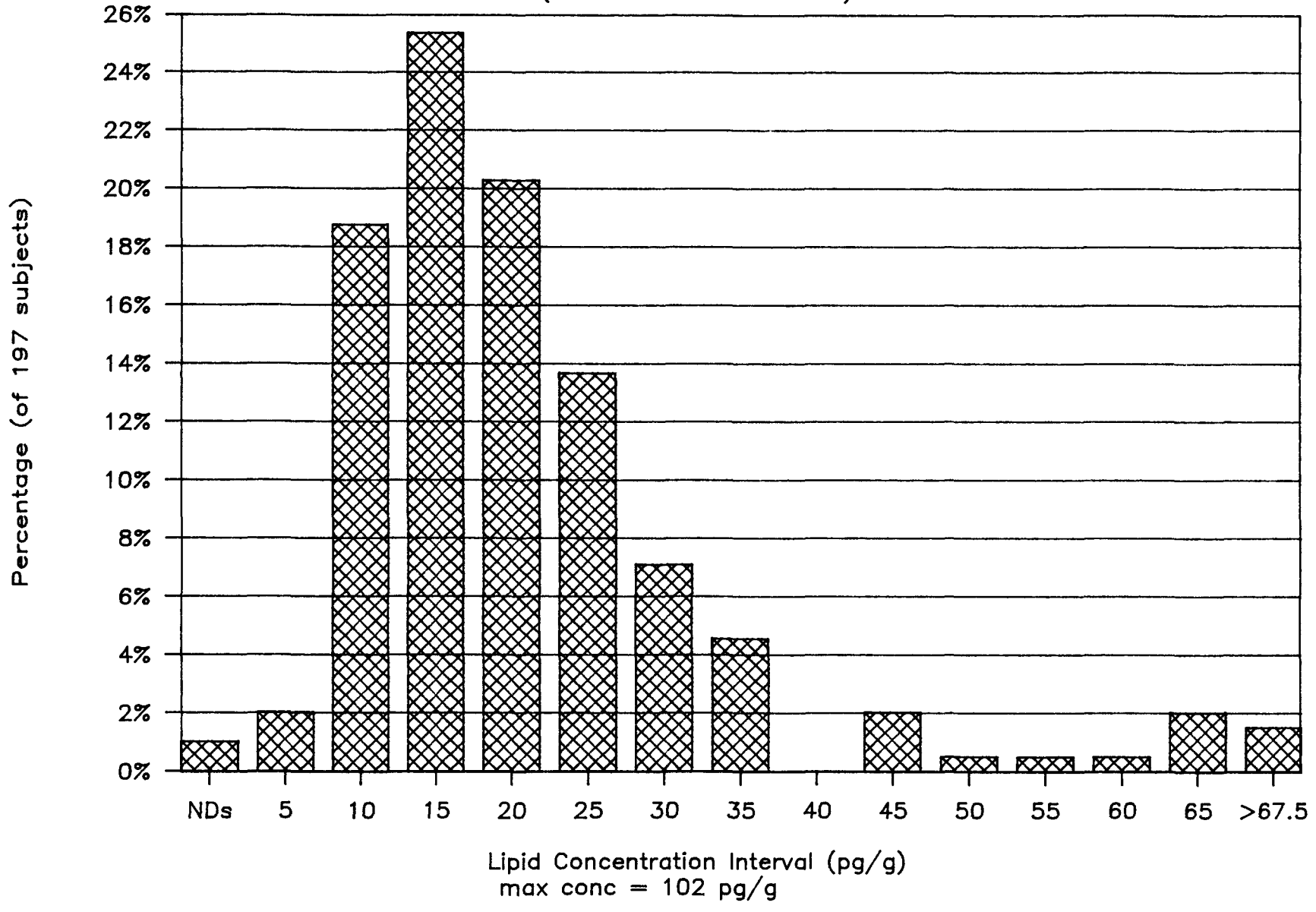
12378-PeCDD

(0.5% below detection limit)



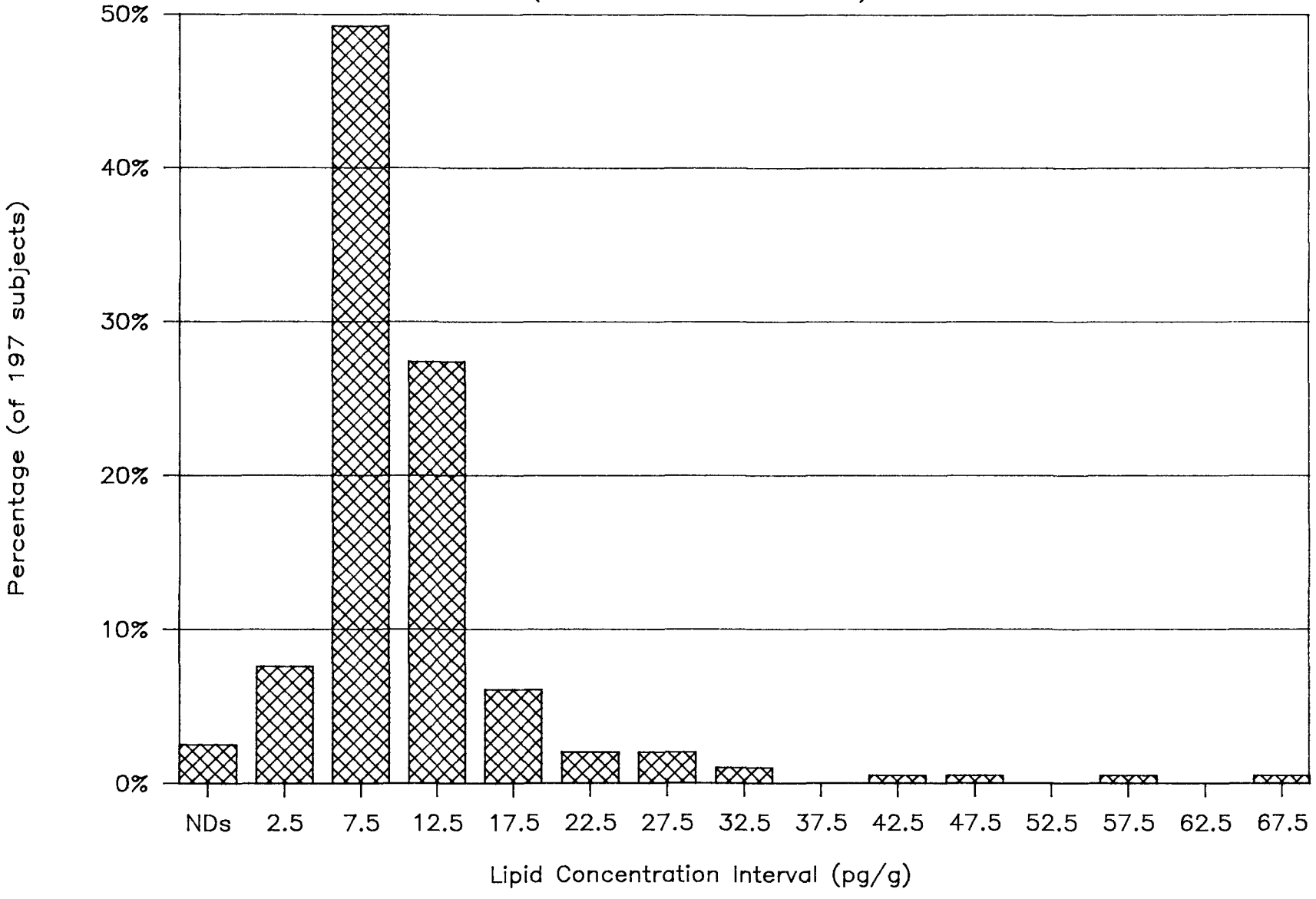
123478-HxCDF

(1% below detection limit)



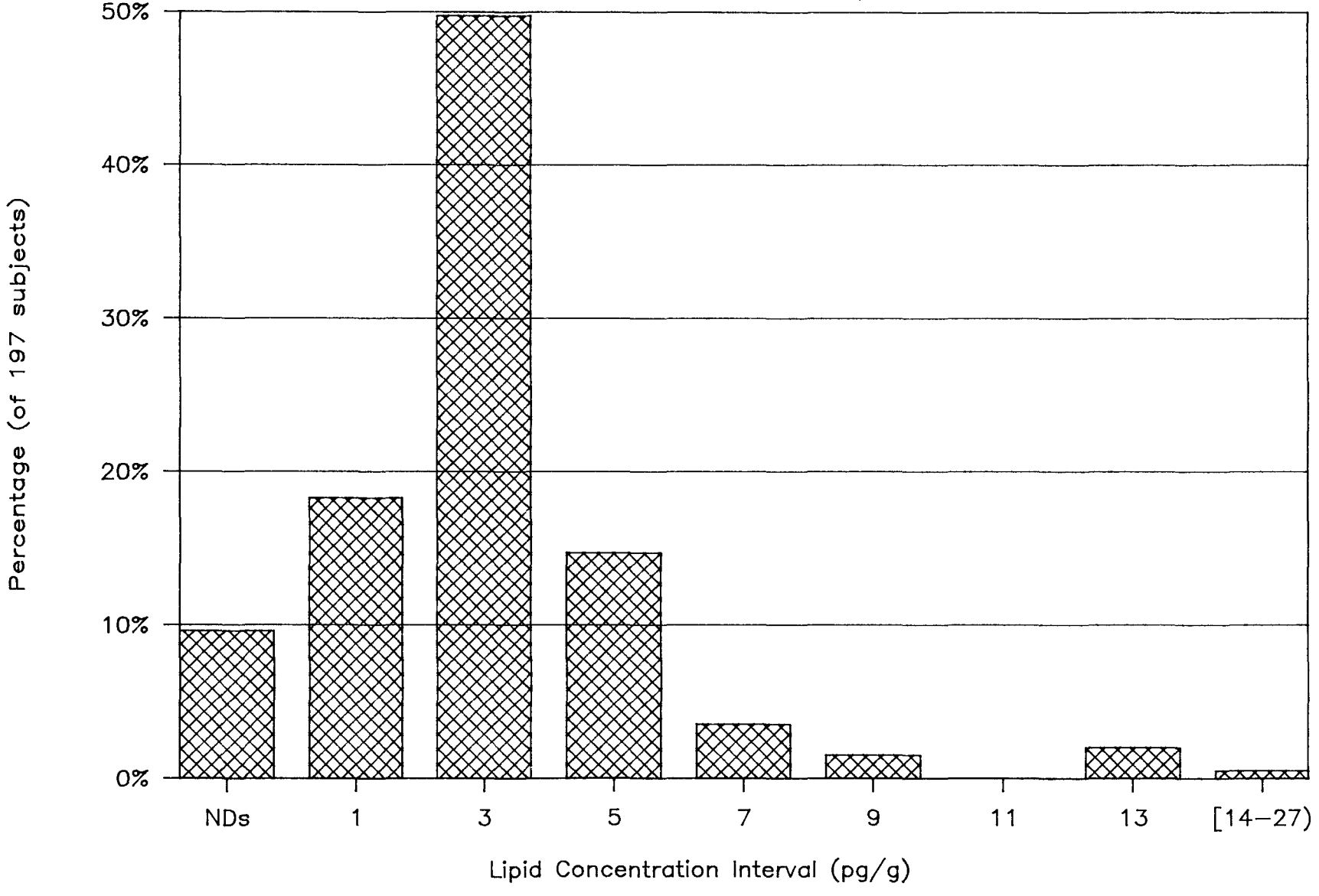
123678-HxCDF

(2.5% below detection limit)



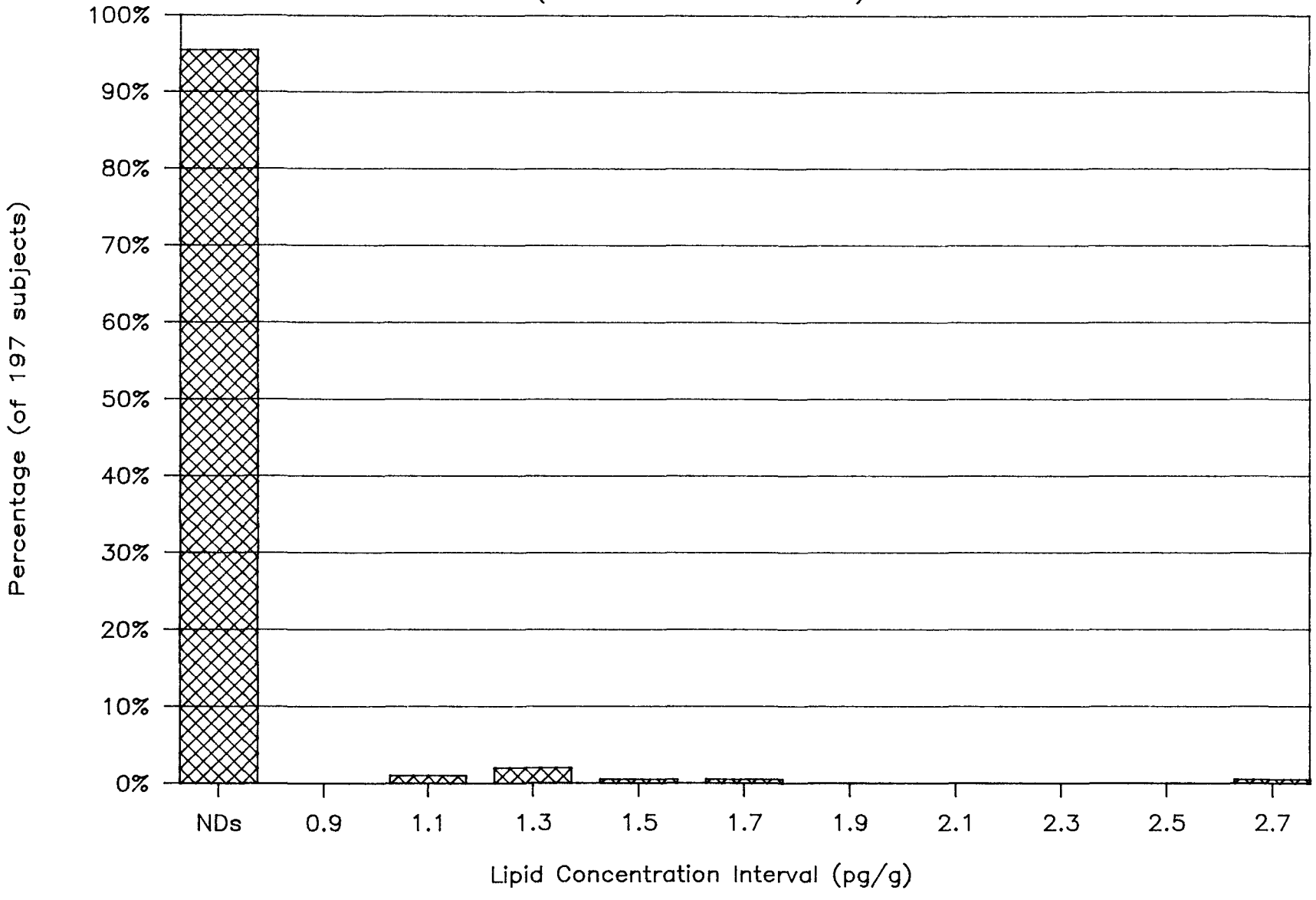
234678-HxCDF

(9.6% below detection limit)



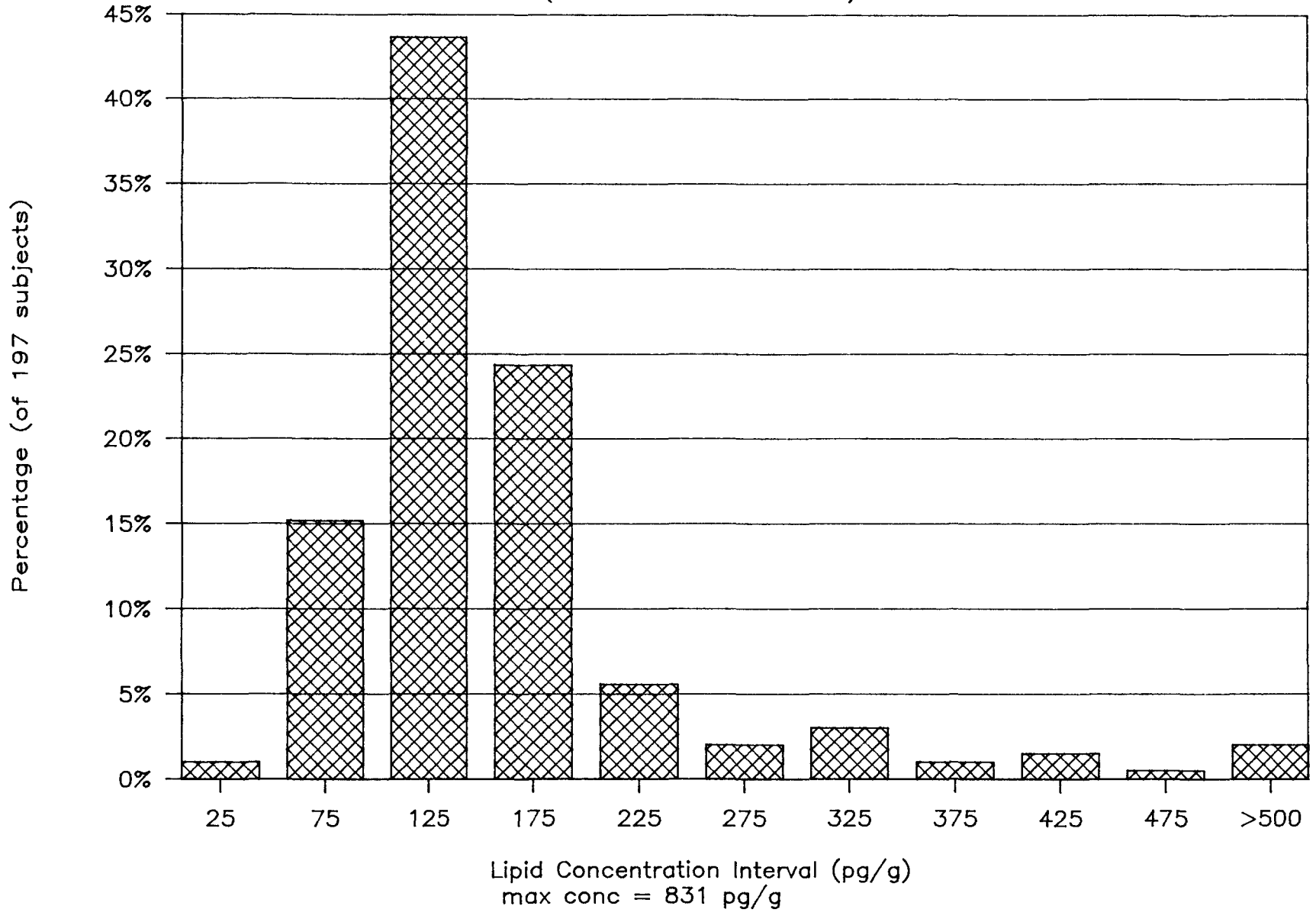
123789-HxCDF

(95% below detection limit)



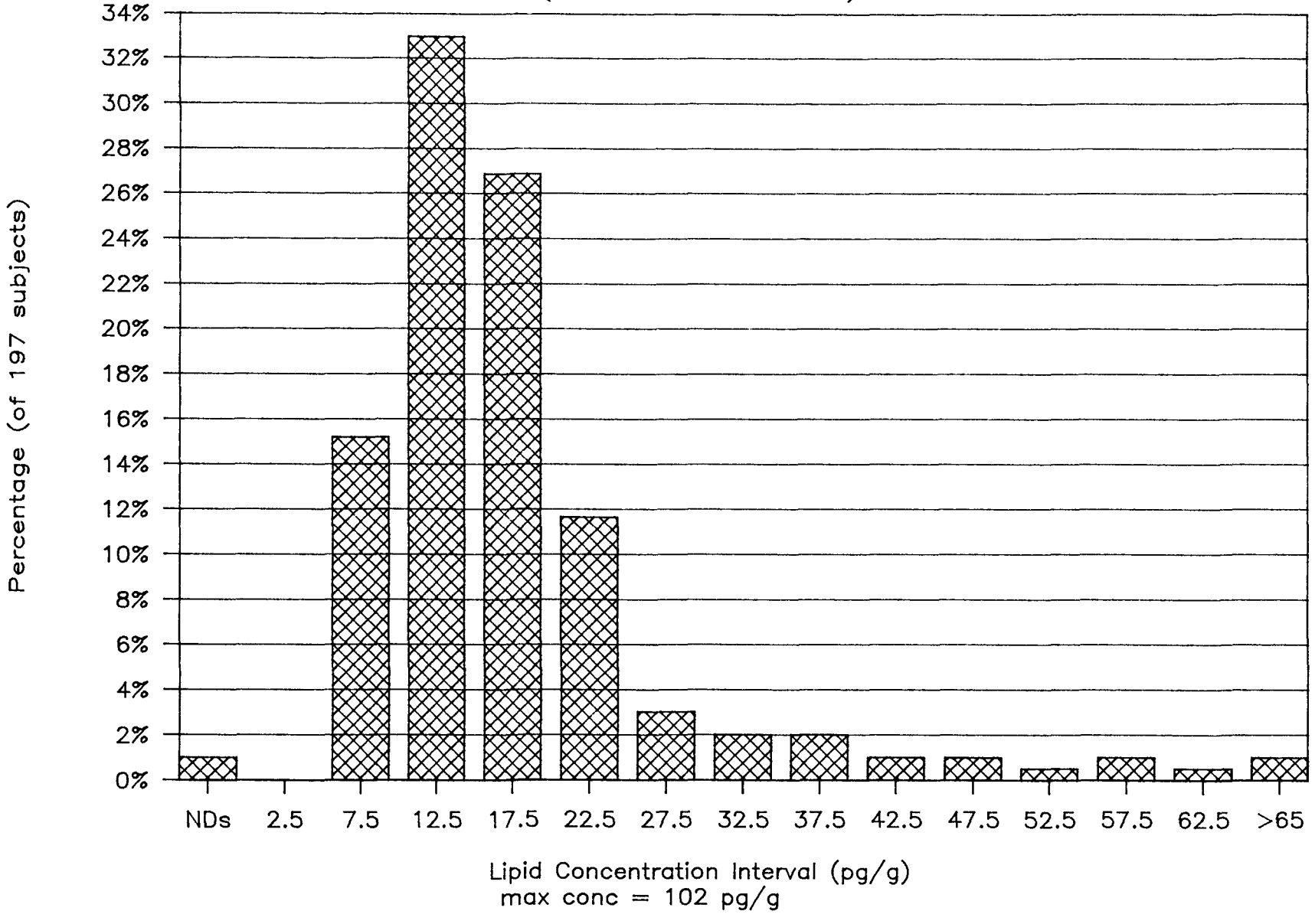
123478/123678-HxCDD

(0% below detection limit)



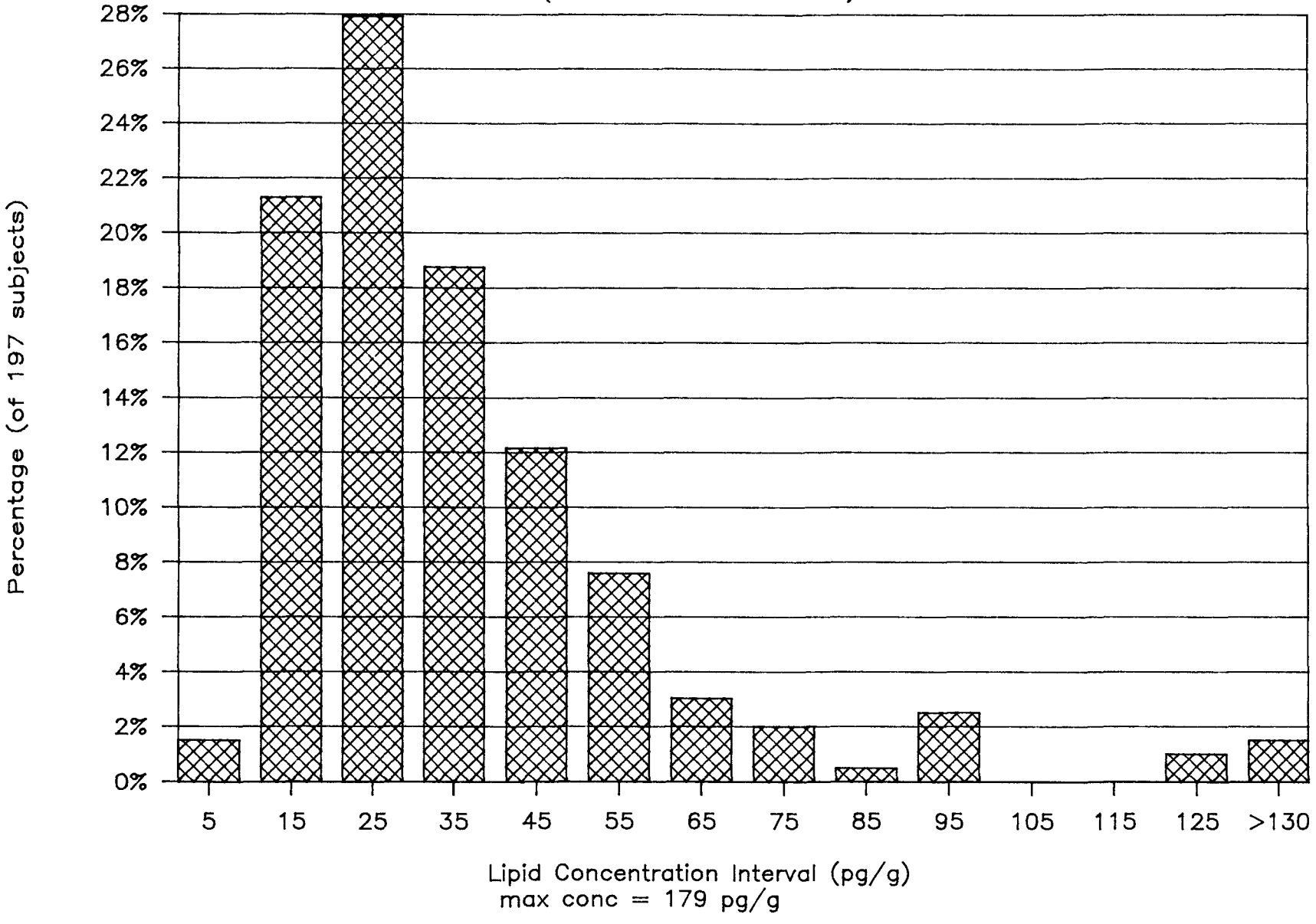
123789-HxCDD

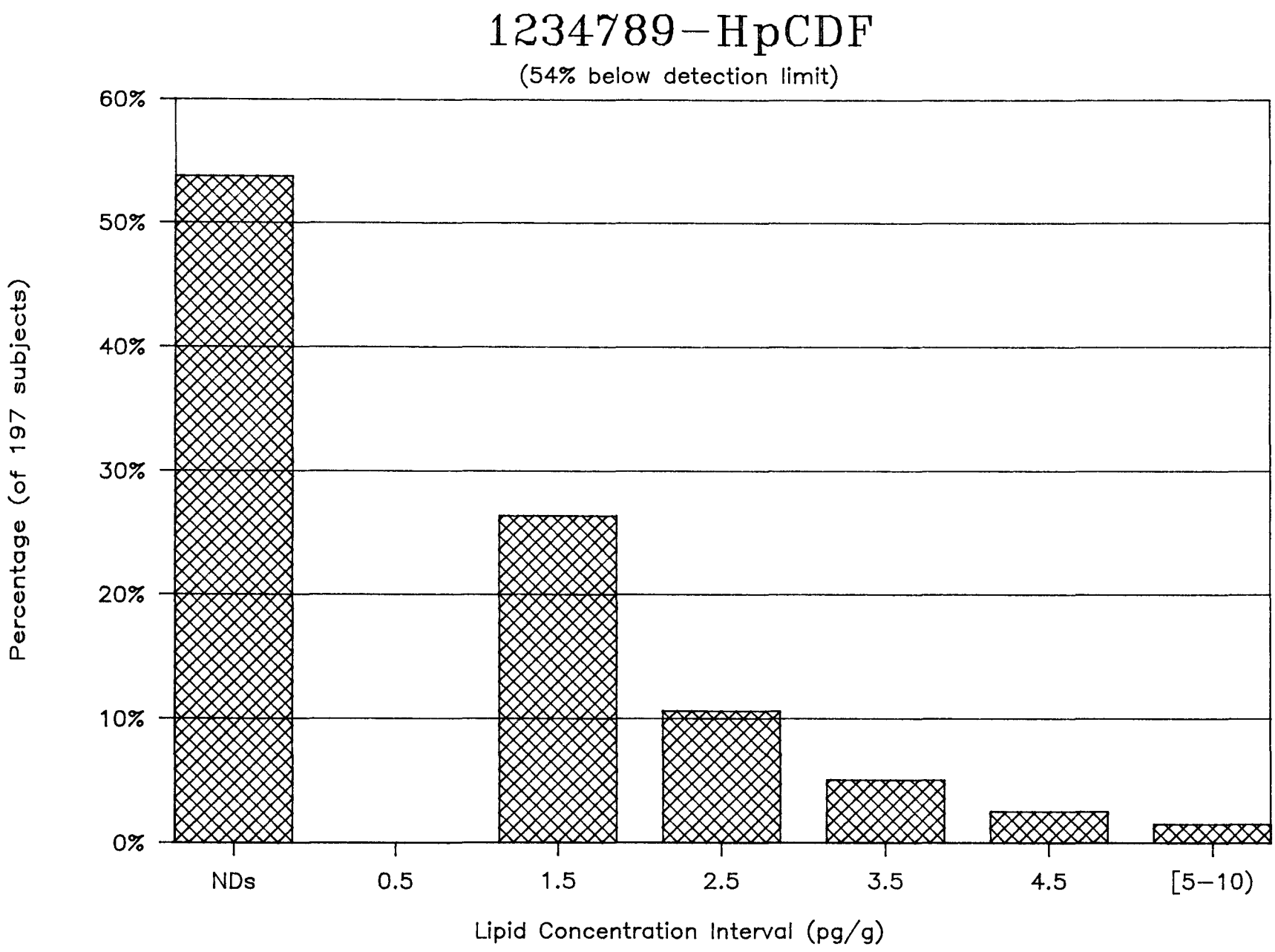
(1% below detection limit)



1234678-HpCDF

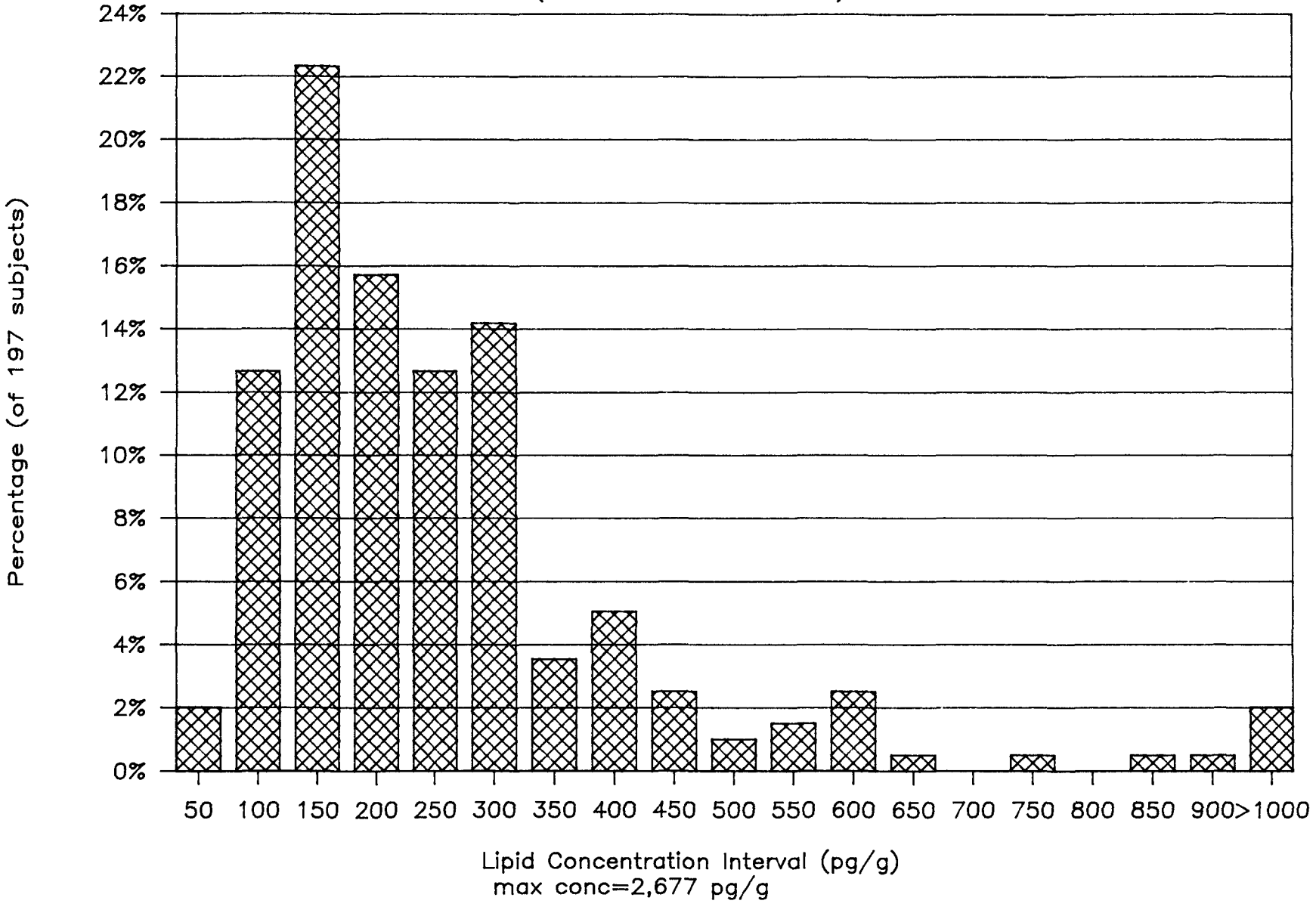
(0% below detection limit)

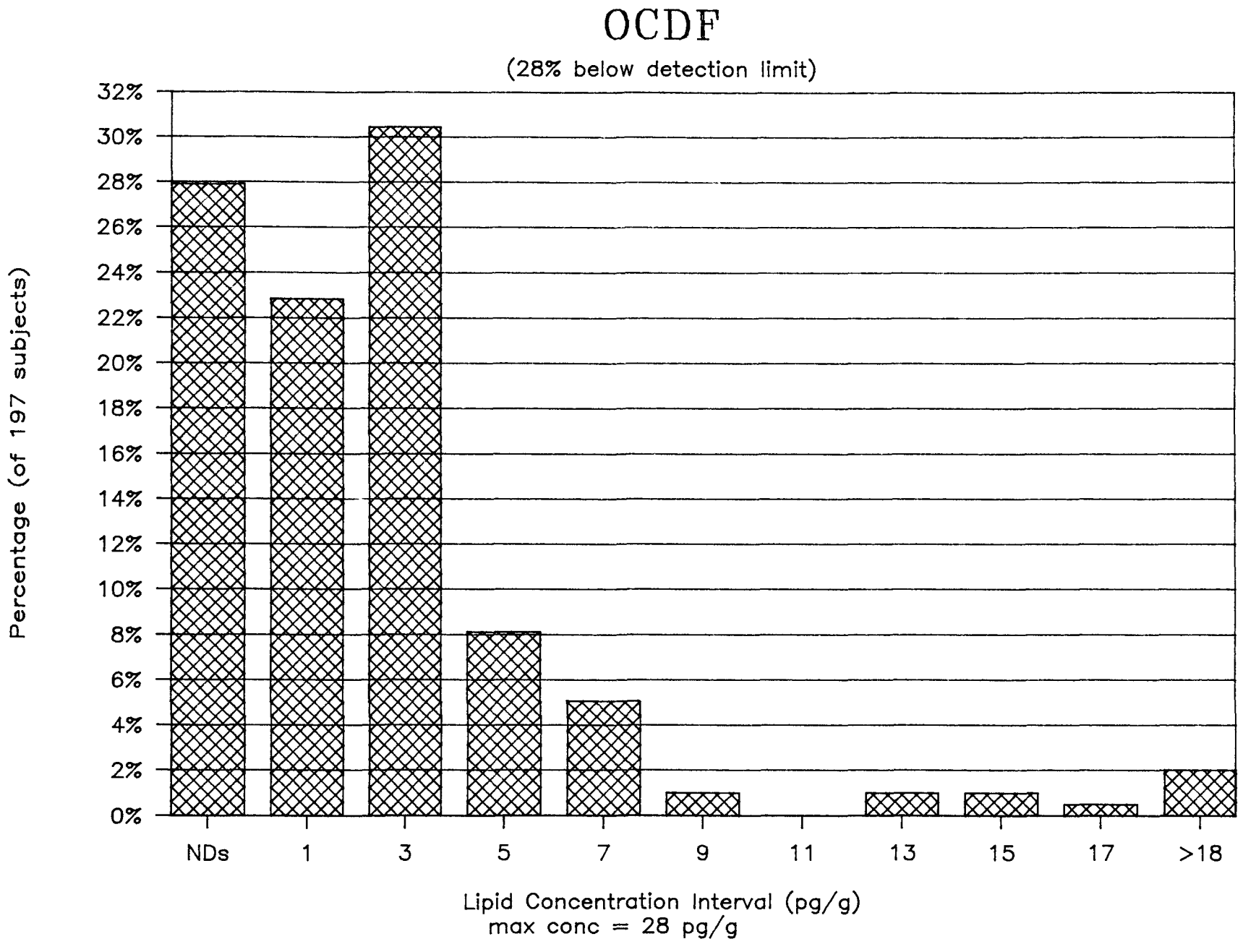


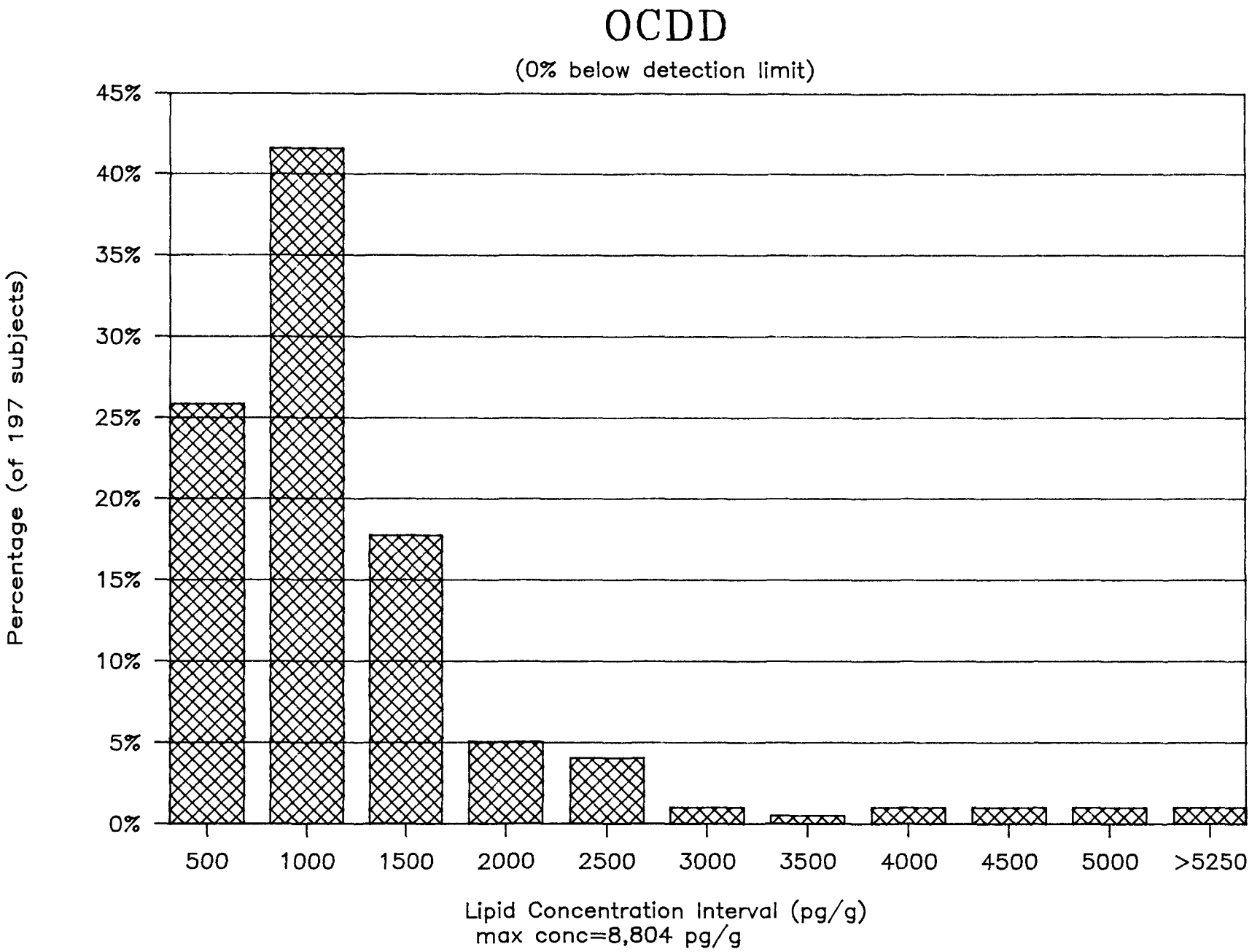

1234789-HpCDF
(54% below detection limit)
60%
50%
40%
30%
20%
10%
0%
Percentage (of 197 subjects)
NDs
0.5
1.5
2.5
3.5
4.5
[5-10)
Lipid Concentration Interval (pg/g)

1234678-HpCDD

(0% below detection limit)




OCDF
(28% below detection limit)
Percentage (of 197 subjects)
32%
30%
28%
26%
24%
22%
20%
18%
16%
14%
12%
10%
8%
6%
4%
2%
0%
NDs
1
3
5
7
9
11
13
15
17
>18
Lipid Concentration Interval (pg/g)
max conc = 28 pg/g


OCDD
(0% below detection limit)
45%
40%
35%
30%
25%
20%
15%
10%
5%
0%
Percentage (of 197 subjects)
500
1000
1500
2000
2500
3000
3500
4000
4500
5000
>5250
Lipid Concentration Interval (pg/g)
max conc=8,804 pg/g

50272 - 101

REPORT DOCUMENTATION PAGE	1. REPORT NO. EPA-560/5-90-006	2.	3. Recipient's Accession No.

4. Title and Subtitle
Pattern Recognition Analysis of VA/EPA PCDD and PCDF Data

5. Report Date
September 1990

6.

7. Author(s)
Karin M. Bauer and John S. Stanley

8. Performing Organization Rept. No.
8863-A

9. Performing Organization Name and Address
Midwest Research
Institute
425 Volker Boulevard
Kansas City, MO 64110

10. Project/Task/Work Unit No.
Work Assignment 25

11. Contract(C) or Grant(G) No.
(C) 68-02-4252
(G)

12. Sponsoring Organization Name and Address
Department of Veterans Affairs
Office of Environmental Epidemiology
1825 K Street, N.W.
Washington, DC 20006

U.S. Environmental Protection Agency
Office of Toxic Substances (TS-798)
Exposure Evaulation Division
401 M Street, S.W.
Washington, DC 20460

13. Type of Report & Period Covered
Final Report
March 1988 - Sept. 1990

14.

15. Supplementary Notes

16. Abstract (Limit: 200 words)

The objective of this study was to evaluate, by pattern recognition techniques, relationships of PCDDs and PCDFs in human adipose tissue among the analytes themselves and with respect to demographic variables (military/civilian status, age, collection year, race, geographic region.) The conclusions of this study provide a multivariate context for the data and strengthen the conclusions obtained in the univariate study (on 2378-TCDD) previously conducted by the VA and EPA (VA AND USEPA, 1989.)

From a principal component analysis, the result of most interest, based on a set of matched NHATS specimens (36 Vietnam veterans, 79 non-Vietnam veterans, and 80 civilians), is that Vietnam veterans in this study do not exhibit higher levels of 2378-TCDD, PCDDs, and PCDFs than the other two groups. Also, cluster analysis of the PCDDs and PCDFs demonstrates that the 2378-TCDD is not strongly correlated with any of the other compounds, indicating possible differences in exposure routes to these chemicals.

17. Document Analysis a. Descriptors

Pattern Recognition
Multivariate Statistical Analysis
Cluster Analysis
Principal Component Analysis
Adipose Tissue
Dioxins
Furans

b. Identifiers/Open-Ended Terms

c. COSATI Field/Group

18. Availability Statement	19. Security Class (This Report) Unclassified	21. No. of Pages 74
	20. Security Class (This Page) Unclassified	22. Price

(See ANSI-Z39.18)

OPTIONAL FORM 272 (4-77)
(Formerly NTIS-35)
Department of Commerce