Appendix G

STATISTICAL METHODS
INFECTIOUS METHODS
STATISTICAL ROADMAP

Prepared in Support of:

CDC/NCEH Cross Sectional Assessment Study

Prepared by:

Centers for Disease Control and Prevention
National Center for Environmental Health

Statistical Methods for Analyzing Data Collected During the Churchill County Study

1: Background and Introduction

An Expert Panel convened by the Nevada State Health Officer, Dr. Mary Guinan, requested that National Center for Environmental Health (NCEH), Centers for Disease Control and Prevention (CDC) assess exposure to a variety of chemicals, active elements, and infectious agents among children in Churchill County, Nevada diagnosed with acute lymphocytic leukemia [ALL] and acute myelocytic leukemia [AML]. The twelve initial cases of ALL reported in Churchill County between 1997 and 2000 suggested that there might be a localized increase in the incidence of ALL for this area.

The primary goal of this study is to assess whether known or probable carcinogens are present at previously unrecognized high levels in Churchill County, Nevada as evidenced by levels measured in the environment and in biological samples from residents and former residents. In total, 15 cases of childhood leukemia were detected in children living or formerly living in Churchill County between 1997 and 2001.

It was hypothesized that there might be known or suspected carcinogens present in the county that could be related to this disease. The secondary goals of this study are to assess these hypotheses by analyzing the cross-sectional data collected. To assist with the complex statistical analysis, investigators from the CDC contracted with Battelle for their assistance. Battelle developed a compact disc of statistical analysis results (Churchill County Analysis CD) that:

- Provides descriptive assessments and statistical analyses of environmental exposures in Churchill County, and, to the extent possible, compares them to national levels published in the Second National Exposure Report [1].
- Provides descriptive assessments and statistical analyses of information collected from questionnaires and interviews that include questions on current occupation and recreational activities of parents/caregivers.

To accomplish these goals, data were investigated from a variety of sources, including (1) the results of tests of environmental samples collected from current and previous Churchill County residences of the children including indoor air, play yard soil, drinking water, and household dust; (2) the results of tests of blood and urine samples collected from the case children and their families and from community-based comparison children and their parents; and (3) information collected from questionnaires and interviews.

Fifteen children with leukemia were identified in Churchill County between 1997 and 2001. Of these, one case family declined to participate in the CDC investigation. Nine case children meet the *restricted case definition*, which reflects the most homogeneous case population with respect to age, diagnosis, and duration of residence in Churchill

County. Community-based comparison children were *matched* to 15 case children (13 living children; 1 deceased child; and 1 child whose family refused to participate) on the basis of year of birth and sex. The comparison children were recruited through random digit dialing of households in Churchill County. They were required to be residents of the county at the time of enrollment. Children were excluded from the community based comparison group if they were siblings of a case child or if they had ever been diagnosed with any form of cancer. There were 55 comparison children enrolled from 54 households.

The following sections outline the information contained in the tables and figures presented on the Churchill County Analysis CD. The document is divided into several sections corresponding to the different types of data analyses that were performed. Section 2 gives a brief introduction to the relevant statistical methods and to conditional logistic regression, the primary analysis tool used to contrast the case and comparison populations. Section 3 describes the summaries available and analyses completed for the questionnaire and interview responses. Section 4 describes the summaries and analyses completed for the biological and environmental samples, and Section 5 briefly describes the summary tables available on the CD. Finally, Section 6 describes some of the limitations associated with these analyses.

2: Statistical Methods

For each of the variables analyzed, univariate descriptive statistics provide an overall picture of the data. For continuous variables, sample sizes, geometric means, and selected percentiles are presented to summarize the range and distribution of the data for the various sub populations of interest. Geometric means (with 95% confidence intervals) are constructed under the assumption that the data approximates a log-normal distribution. The estimate of the mean and confidence interval is based on a statistical model that controls for the possible correlation of observations within a family (i.e. a variance components model), when appropriate. In some cases, comparison values (based on Second National Exposure Report [1] and/or other published levels) are presented in order to give an indication of what constitutes a “large” or “small” value for a selected summary statistic. For categorical variables, frequency counts and percentages are presented as summary statistics for the subpopulations of interest. The denominator for calculating percentages is related to all relevant study subjects with valid responses, and omits those with missing, invalid or incomplete information.

In the secondary analysis, we compared exposure among the case and comparison populations using conditional logistic regression. In general, logistic regression models are used to model the association between disease and exposure. For studies that recruit study subjects independently from their disease status, logistic regression models are used to model the probability of disease as a function of measured exposure. In this study, because the study subjects were recruited based on their disease status, the logistic regression models are used to model the probability of increased exposure as a function of disease status. For example, does exposure to a particular chemical tend to be higher among the case population than the comparison population? For this case/control study in

which each of the case subjects are *matched* to several comparison subjects, conditional logistic regression models are used to model the association between exposure and case-control status while conditioning on the *matching* that is a feature design of this study.

Odds ratios, based on logistic regression models (or conditional logistic regression models), provide a means of exploring the relationship between disease and exposure by estimating the ratio of the odds of experiencing the event for one group versus the odds of experiencing the event for another group. In this study, the event is increased exposure and the groups are defined as being a member of the designated case or comparison study populations. An odds ratio of 1.0 indicates no differences between the two groups with respect to the exposure variable, in which both groups are equally likely to have been exposed. Odds ratios that are statistically significantly different from 1.0 (as assessed by a p-value) indicate that one group has higher (or lower) odds of having been exposed. For this study, small p-values (e.g. p-values less than 0.01) are indicative of statistically significant differences in the exposure variable between the case and comparison groups. Likewise, odds ratio confidence intervals that do not contain the number one indicate statistically significant differences between the case and comparison populations for the exposure variable. For categorical exposure variables, odds ratios are used to assess the association between disease and exposure between two specific levels of the categorical variable. For these continuous exposure variables, the odds ratios are based on data that were standardized prior to analysis. In essence, the exposure measures are standardized by dividing each individual response by the standard deviation observed among the entire study population. The LogXact software from Cytel Corporation was used to fit the conditional logistic regression models.

For this study, many of the logistic regression models compare current levels of exposure among the case and comparison populations. Because both populations were sampled well after diagnosis and onset of disease among the case population; treatment, past diagnostic events, changes in behavior and changes in chemical exposures over time may all have significant though immeasurable impact on the relationships that are being explored in this secondary analysis. Thus, the resulting odds ratios cannot necessarily be interpreted as rate ratio, risk ratio, or other causal measures.

3: Questionnaire and Interview Responses

Similar statistical analysis methods were applied to variables appearing on the questionnaire and interview tabs of the CD. Thus, they are jointly described in the following sections, with Section 3.1 describing the analysis performed for categorical questionnaire and interview variables, and Section 3.2 describing the analysis performed for continuous variables from the questionnaire and interview.

3.1 Categorical Variables

For each of the analyzed categorical variables from the questionnaire and interviews, summary statistics, figures, and conditional logistic regression models are computed and presented. At the top of each univariate analysis page, a short title describing the variable

of interest is presented. Just below the title and above the first table the full description of the variable of interest can be found. The following list describes the information found on the remainder of the page:

- A table displaying cell frequencies and percentages of total number of valid respondents (non-missing) for each of the possible question responses and for each subgroup of interest. Some questions being analyzed pertain to a particular family member (the reference child, biological father, etc.) and other questions relate to current or previous residences of the family. The subgroups of the study population are characterized as case or control based on the status of the enrolled child and include members of the case or control family to whom the question applies.
    - All Subjects – data reported here describe all children, regardless of case or control status.
    - Case Subjects – data reported here describe the case children.
    - Comparison Subjects – consists of all children recruited into the study as comparisons.
    - Restricted Definition Case Subjects – consists of all children classified as a restricted definition case child, based on the age, diagnosis and duration of residence in Churchill County.
    - Comparison Subjects Matched to Restricted Definition Case Subjects – consists of all children classified as comparison matched to a case respondent classified as a restricted definition case.
- The table is followed by a figure displaying a side-by-side bar chart of the percentages of valid responses (non-missing) for the case and comparison subjects. The chart enables a graphical contrast of the case and comparison groups.
- The second table displays the results of a conditional logistic regression model with case/comparison status as the outcome, and the response to the question of interest as the explanatory variable. Odds ratios are constructed using exact methods as implemented by the LogXact software from Cytel Corporation. The p-value column corresponds to an exact test of the overall significance of the selected variable in the model. The conditional logistic regression model is fit to two subgroups of the study population. They are:
    - Case Subjects vs. Comparison Subjects – compares all case and comparison children
    - Restricted Definition Case Subjects vs. Comparison Subjects Matched to Restricted Definition Case Subjects – compares only those subjects classified as cases or matched to a restricted definition case subject.

3.2 Continuous Variables

For each of the analyzed continuous variables from the questionnaire and interviews, summary statistics, figures, and conditional logistic regression models are computed and presented. Again, at the top of each analysis page, a short title describes the variable of interest. Just below the title and above the first table the full description of the variable of

interest can be found. The following list describes the information found on the remainder of the analysis page:

- A table displaying summary statistics (arithmetic means, standard deviations and percentiles) for responses to the question of interest. The statistics are calculated for the same subgroups of the study population as described for the categorical variables in the questionnaire and interview datasets.
- A figure displaying side-by-side box plots for the subgroups of the population described above. The upper and lower edges of the box represent the 75^{th} and 25^{th} percentiles, respectively. The bar through the middle of the box identifies the median or 50^{th} percentile, and the lines extending from the box are referred to as whiskers. The upper and lower ends of the whiskers are drawn from the box to the most extreme point within 1.5 interquartile ranges (the distance between the 25^{th} and the 75^{th} sample percentiles). Any value more extreme than this is marked with a star.
- A second table displaying the results of a conditional logistic regression analysis with case versus comparison status as the outcome variable and the response to the question of interest as the explanatory variable. The model is fit to the same two subgroups as described for the corresponding categorical variables table. The displayed odds ratio estimates provided in the table for continuous responses correspond to a one-unit change in the continuous response and are constructed using asymptotic methods as implemented by the LogXact software from Cytel Corporation.

4: Biologic and Environmental Samples

Statistical analyses of the biologic and environmental samples are broken into two analysis sections, one for the comparison subjects group, and one for the case subjects group. The comparison group summaries provide a snapshot of the potential exposure in the entire Churchill County population, while the case population summary helps summarizing the case group and in contrasting the case and comparison group (via conditional logistic regression models and summary statistics). An additional complexity in the analysis of biological and environmental samples is the inclusion of the limit of detection (LOD). The LOD is generally considered the lowest concentration of an analyte that the analytical process can reliably detect. Thus, concentration values that fall below the LOD may be considered below the range that a measurement is considered reliable. For purposes of analysis, if a large percentage of the measurements falls below the LOD, some of the reported summary statistics may be uninformative. For this reason, the summary tables for the biological and environmental samples include the percent of observations that were below the LOD, and the median and maximum LOD value reported. Geometric means and conditional logistic regression models for target biologic and environmental samples were computed when less than 40 percent of the data were observed below the detection limit. When more than 40 percent of the observed data were below the detection limit, a conditional logistic regression model was fit to a binary variable that indicated whether that sample was above or below the LOD. LODs for lipid-

adjusted and creatinine adjusted sample results were computed based on methods used and documented in the Second National Exposure Report [1].

4.1 Biologic Samples

The biologic samples are obtained from blood and urine samples collected from the case children and their families and from community-based comparison children and their parents. For the comparison population, the summary page consists of the following:

- Title describing the analyte of interest and the media in which it was measured
- A summary table presents descriptive statistics for the analyte of interest among each group or subgroup of interest, and comparisons to summary statistics obtained from the Second National Exposure Report [1]. Descriptive statistics from the Second National Exposure Report are provided in the first row of this table, and are based on a national population-based survey of exposure. The groups or subgroups of the study population are defined as:
    - All Study Subjects – consists of **all** study subjects regardless of case/comparison status and family status
    - Comparison Subjects and their Family Members – consists of all comparison study children, their parents or guardians and other related adults living with the study child
    - Comparison Subjects – consists of all comparison study children
    - Comparison Family Members Only – consists of the parents or guardians and other related adults living with the study child
    - Restricted *Case* Definition Comparison Subjects – consists of comparison subjects matched to a case subject who meets the restricted case definition (with respect to age, diagnosis and duration of residence in Churchill County)
    - Restricted Definition Comparison Family Members Only – consists of family members of restricted definition comparison study children.
- A histogram of the natural log transformed analyte concentration values for comparison subjects and their family members. The histogram gives an overall picture of the distribution of concentration of the analyte of interest.
- Box and whisker plot provided (see section 3.2 for description).

For the case population biologic samples, the summary information is analogous to that displayed by the comparison population page with the exception of a table displaying the results of a conditional logistic regression analysis to contrast the case and comparison populations. The case subjects summary page consists of the following:

- Title describing the analyte of interest and the media in which it was measured
- A summary table presents descriptive statistics for the analyte of interest among each group or subgroup of interest, and comparisons to summary statistics obtained from the Second National Exposure Report [1]. Descriptive statistics from the Second National Exposure Report are provided in the first row of this table, and are based on a national population based survey of exposure. The groups or subgroups of the study population are defined as:

    - All Study Subjects – consists of **all** study subjects regardless of case/comparison status and family status
    - Case Subjects and their Family Members – consists of all case study children, their parents or guardians, siblings, and other related adults living with the study child
    - Case Subjects and their Family Members Excluding Siblings – consists of parents or guardians of case study children and other related adults living with the study child
    - Case Subjects – consists of all case study children
    - Case Family Members Only, Excluding Siblings – consists of all parents/guardians of case study children
    - Case Siblings – consists of siblings of all case participating children
    - Restricted Definition Case Subjects – consists of case subjects classified as a restricted definition case (with respect to age, diagnosis and duration of residence in Churchill County)
    - Restricted Definition Case Family Members Only, Excluding Siblings – consists of all parents/guardians of restricted definition case study children
    - Restricted Definition Case Siblings – consists of siblings of restricted definition case study children.
- A histogram of the natural log transformed analyte concentration values for case subjects and their family members (see description above).
- A figure displaying side-by-side box plots for the subgroups of the population described above (see box plot description above).
- A second table displaying the results of a conditional logistic regression analysis with case versus comparison status as the outcome variable and the value of the exposure assessment as the explanatory variable. The odds ratio estimates in the table correspond to an increase in dose corresponding to one standard deviation in the study population and are constructed using asymptotic methods as implemented by the LogXact software from Cytel Corporation. The subgroups are defined as (refer to specific group definitions above):
    - Case vs. Comparison Subjects and their Family Members Excluding Siblings
    - Case vs. Comparison Subjects
    - Case vs. Comparison Family Members Only Excluding Siblings
    - Case vs. Comparison Restricted Definition Subjects
    - Case vs. Comparison Matched to the Restricted Definition Family Members Only Excluding Siblings.

4.2 Environmental Samples

The environmental samples were collected from Churchill County residences of the study children (both comparison children and case children) including indoor air, play yard soil, drinking water, and household dust. For the case families, environmental sampling was attempted at all current and previous residences of the subject child within Churchill County. The current residence for case families is defined as the housing unit where the subject child lived at the time of diagnosis. For the comparison families, environmental

sampling occurred at the current residences for all recruited families. In addition, one of the four comparison families matched to each case family was randomly selected for attempted environmental sampling at previous residences within Churchill County. For the comparison population, the summary page consists of the following:

- Title describing the analyte of interest and the media in which it was sampled
- A summary table presents descriptive statistics for environmental samples obtained from among each group or subgroup of interest. The group or subgroups of the study population are defined as:
    - All Subjects-All Residences – consists of all collected environmental samples from all residences
    - Current Residence of Comparison Subjects – consists of all samples collected from a current residence of a comparison subject
    - Historical Residence(s) of Comparison Subjects – for each case family, a single matched comparison family was randomly selected from the comparison families matched to each case for environmental sampling among all previous residences within Churchill County.
    - Current Residence of Restricted Definition Comparison Subjects – consists of all samples collected from a current residence from a comparison subject that is matched to a restricted definition case subject
    - Historical Residence(s) of Restricted Definition Comparison Subjects – consists of those comparison families selected for environmental sampling in historical residences that were matched to Restricted Definition Case Families.
- A histogram of the natural log transformed analyte concentration values for all samples collected from a current or historical residence of a comparison subject. The histogram gives an overall picture of the distribution of the exposure to the analyte of interest in the community.
- A figure displaying side-by-side box plots for the subgroups of the population described above (see box plot description above).

For the case population environmental samples, the summary information is analogous to the comparison population samples with the exception of a table displaying the results of a conditional logistic regression analysis to contrast the case and comparison populations. The case summary page consists of the following:

- Title describing the analyte of interest and the media in which it was sampled
- A summary table presenting a variety of descriptive statistics for each subgroup of interest. The subgroups of the study population are defined as (for descriptions see above):
    - All Subjects-All Residences
    - Current Residence of Case Subjects
    - Historical Residence(s) of Case Subjects
    - Current Residence of Restricted Definition Case Subjects
    - Historical Residence(s) of Restricted Definition Case Subjects.
- A histogram of the natural log transformed analyte concentration values for all samples collected from a current or historical residence of a case subject. The

histogram gives an overall picture of the distribution of the exposure to the analyte of interest

- A figure displaying side-by-side box plots for the subgroups of the population described above (see box plot description above).
- A final table displaying the results of a conditional logistic regression analysis with case versus comparison status as the outcome variable and the value of the exposure assessment as the explanatory variable. The odds ratio estimates provided in the table correspond to an increase in dose corresponding to one standard deviation in the study population and are constructed using asymptotic methods as implemented by the LogXact software from Cytel Corporation. The subgroups are defined as (descriptions can be inferred based on the above subgroup descriptions):
    - Current Residence of Case Subjects vs. Current Residence of Comparison Subjects
    - Current Residence of Restricted Definition Case Subjects vs. Current Residence of Comparison Families Matched to Restricted Definition Case Subjects

  Note that conditional logistic regression models were not pursued for the historical residences, due to the fact that a number of case and selected comparison families did not have any data from historical residences.

5: Summary Tables

The summary tables give a general overview of all the analytes (broken down by category) on a single table. They are a subset of the more detailed univariate analyses (described above) that are produced for the biologic and environmental samples only.

6: Limitations and Caveats

Several limitations are present in the collected data and in the analyses presented. They are as follows:

- Although the goal of the study was to match four community-based comparison children to each case child, this was not always achieved. Further, comparison children were enrolled for the deceased case child from whom no biologic samples were collected and for the case child whose family refused to participate. These comparison children/families provide useful information for the primary goal of characterizing exposures among the Churchill County community; however, their samples do not contribute to the secondary analyses that were pursued with the conditional logistic regression models.
- The environmental measures collected represent a single point in time and were obtained after diagnosis. They may not correlate with the analyte levels present before a child was diagnosed with leukemia (e.g., pre-diagnosis levels), and they may not correlate with the amount of environmental exposure before the development of leukemia.
- According to the literature, there can be a low correlation between personal and ambient monitoring for environmental exposures.

- Some of the chemicals measured have short half-lives.
- The laboratory findings in case children may have been affected by the treatment of their leukemia. It is not possible to determine how treatment will have altered the presence of analytes or infectious agents. Therefore, current measurements may not accurately reflect pre-treatment exposures.
- We will have very low power to detect any differences between the cases and community-based comparisons due to the small sample size.
- The number of measurements (including survey, biological and environmental samples) is far larger than the number of study subjects.
- Several case children spent a majority of their lives outside Churchill County.

Approach to Analyzing Potential Relationships between Viral Testing Results and Leukemia Outcomes

The overall seroprevalence of Epstein-Barr virus (EBV) among the study participants was compared with previously published seroprevalence rates in U.S. and other populations. Individual results for participants with a history of childhood leukemia (case children) were then compared with the individual results of their matched comparison participants (comparison children). The most meaningful evaluation for a potential causal link between EBV and the development of childhood leukemia requires that infection occur before the leukemia. However, serology results from this community assessment only indicate the EBV infection status of the individuals at the time each blood sample was drawn. They do not definitively predict infection status at the time of the reference case child's leukemia diagnosis unless diagnosis was made within 3 months of the blood draw date. Therefore, for analysis purposes only, we estimated the likely EBV infection status at the reference case child's date of diagnosis. When the test results of a case child suggested recent EBV infection (<3 months before the blood draw date) and the leukemia diagnosis date preceded his or her blood draw date by 3 months or more, that recent EBV infection was classified as negative at the time of diagnosis. Similarly, the EBV infection status of each comparison child was classified according to the leukemia diagnosis date of the case child to whom he or she was matched (reference case child). Recent EBV infection in a comparison child was classified as negative at the time of the reference case child's diagnosis if that diagnosis date preceded the comparison child's blood draw by 3 months or more. "Probable negative" results were treated as negative. "Probable past infection" was treated as past infection. Past and negative serology results were not reclassified. All subsequent analyses used this dichotomous scheme of infection status classification – all leukemia case children and comparison children were classified as positive or negative for EBV infection at the leukemia diagnosis date of the appropriate reference case child. The positive or negative infection status of all leukemia case children was compared to their matched comparison children. To maximize the chance of detecting any potential relationships between EBV infection and childhood leukemia, the same analyses were also performed on these subgroups of case children and their matched comparison children: the subgroup referred to as restricted case definition; B-cell leukemias; T-cell leukemias; AML; B-cell leukemias diagnosed at <18 years old; B-cell leukemias living in Churchill County at least 6 months before diagnosis.

The retrovirus testing results were not further analyzed because all results were negative.

Roadmap to Information Contained in the Analysis of Interview Variables

The first three pages of this document provide an overview of the information contained in the tables corresponding to the interview variables.

1: Categorical Variables

Figure 1 displays an example template for the categorical variables that appear in the interview datasets.

Study Child Adoption Status

Table of Question: Is the study child adopted

Subgroup of Study Population	N	Number Missing or Don't Know Responses	Possible Responses	
			Yes (%)	No (%)
All Subjects	68	0	3 (4.41%)	65 (95.59%)
Case Subjects	14	0	0 0.00%	14 (100.00%)
Comparison Subjects	54	0	3 (5.56%)	51 (94.44%)
Restricted Definition Case Subjects	9	0	0 0.00%	9 (100.00%)
Comparison Subjects Matched to Restricted Definition Case Subjects	34	0	3 (8.82%)	31 (91.18%)

Bar chart for both comparison and case populations

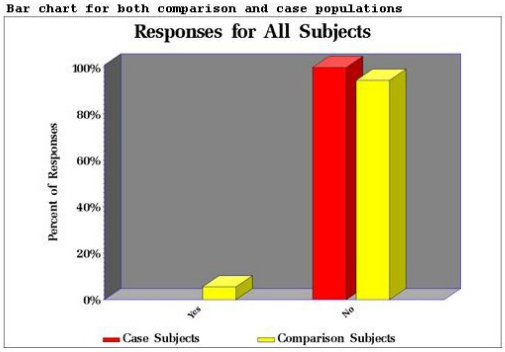


Table of Odds Ratios Based on Conditional Logistic Regression

Subgroup of Study Population	Yes vs. No Odds Ratio (95% CI)	P-value[a]
Case Subjects vs. Comparison Subjects	0.938 [b] (0.000, 8.771) [b]	0.600
Restricted Definition Case Subjects vs. Comparison Subjects Matched to Restricted Definition Case Subjects	0.938 [b] (0.000, 8.771) [b]	0.600

[a] P-value for overall significance of variable in conditional logistic regression model
[b] Unstable estimates

Figure 1: Example of categorical variable template

- The first table on the page displays a frequency table of the responses for each subgroup of interest. The column labeled "N" denotes the total number of respondents with valid responses to the question of interest, and the column labeled "Number Missing or Don't Know Responses" denotes the total number of respondents for which a valid response could not be determined. The remainder of

the table consists of counts and percentages for each of the possible responses. The subgroups of the study population are defined as:
    - All Subjects – consists of all respondents regardless of case/comparison status.
    - Case Subjects – consists of all respondents classified as cases
    - Comparison Subjects – consists of all respondents classified as comparisons
    - Restricted Definition Case Subjects – consists of all respondents classified as a restricted definition case (with respect to age, diagnosis and duration of residence in Churchill County)
    - Comparison Subjects Matched to Restricted Definition Case Subjects – consists of all respondents classified as comparisons with the matching case respondent having been classified as a restricted definition case
- The displayed figure presents a bar chart of the percentage of each response for all case (red bar) and comparison (yellow bar) subjects.
- The second table displays results based on a conditional logistic regression model with case/comparison status as the outcome, and the response to the question of interest as the explanatory variable. Odds ratios are constructed using exact methods as implemented by the LogXact software from Cytel Corporation. The p-value column corresponds to an exact test of the overall significance of the selected variable in the model. A variety of notes corresponding to the odds ratio estimates are possible (see Table 1 for description of notes). The conditional logistic regression model is fit to two subgroups. They are:
    - Case Subjects vs. Comparison Subjects – compares all case and comparison subjects
    - Restricted Definition Case Subjects vs. Comparison Subjects Matched to Restricted Definition Case Subjects – compares only those subjects classified as or matched to a restricted definition case subject.

Table 1: Description of notes corresponding to conditional logistic regression models

Note	Description
A	P-value for overall significance of variable in conditional logistic regression model
b	Parameter estimates are unstable
c	No estimate provided by logXact due to: all strata non-informative, all responses being equivalent, or lack of model convergence
*	Referent level differs for the two groups in the table. The referent level for the restricted definition group is indicated in the note.

2: Continuous Variables

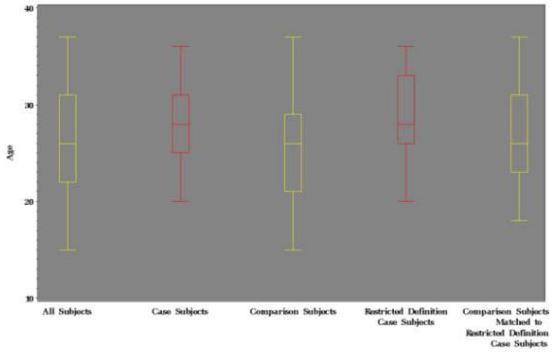
Figure 2 displays an example template for the continuous variables that appear in the interview dataset.

Age of Biological Mother

Table of descriptive statistics for: What was the age of the child's biological mother at the time the study child was born

Subgroup of Study Population	N	Number Missing or Don't Know Responses	Arithmetic Mean	SD	Selected Percentiles					P-Value [a]
					10th	25th	50th	75th	90th	
All Subjects	67	1	26.31	5.25	19	22	26	31	33	
Case Subjects	14	0	27.93	4.73	22	25	28	31	35	
Comparison Subjects	53	1	25.89	5.34	19	21	26	29	32	0.1702
Restricted Definition Case Subjects	9	0	28.78	5.54	20	26	28	33	36	
Comparison Subjects Matched to Restricted Definition Case Subjects	34	0	26.65	5.24	19	23	26	31	32	0.2735

[a] P-value for overall significance of variable in conditional logistic regression model



Plot Description: The upper and lower edges of the box represent the 75th and 25th percentiles, respectively. The bar through the middle of the box identifies the median or 50th percentile, and the lines extending from the box are referred to as whiskers. The upper and lower ends of the whiskers are drawn from the box to the most extreme point within 1.5 interquartile ranges (the distance between the 25th and the 75th sample percentiles). The stars beyond these ends represent outliers.

Table of Odds Ratios Based on Conditional Logistic Regression

Subgroup of Study Population	Odds ratio (95% CI)	P-Value [a]
Case Subjects vs. Comparison Subjects	1.094 (0.958, 1.249)	0.1702
Restricted Definition Case Subjects vs. Comparison Subjects Matched to Restricted Definition Case Subjects	1.084 (0.936, 1.256)	0.2735

[a] P-value for overall significance of variable in conditional logistic regression model

Figure 2: Example of continuous variable template

- The first table on the page displays several summary statistics (arithmetic means, standard deviations and percentiles) for responses to the question of interest. The statistics are calculated for the same subgroups of the study population as described for the categorical variable templates. The p-value column corresponds to the p-value to the test of the significance of the selected variable in a conditional logistic regression model (see second table on page).
- The box plot displays the distribution of values for the five subgroups of the population considered.
- The second table displays the results of a conditional logistic regression analysis with case vs. comparison status as the outcome. The model is fit to the same two subgroups as described for the corresponding table in the categorical variables template. The odds ratio estimates provided in the table for continuous responses correspond to a one-unit change in the continuous response (e.g. in the above example, the odds ratio corresponds to a one-year difference in the age of the biological mother).