Chapter 7
CTEPP Database

7.1 Overview

The CTEPP database was configured similarly to the database developed in the National Human Exposure Assessment Survey (NHEXAS)-Arizona study (14). The database followed the general format that was used in EPA's exposure database that was current at the start of the CTEPP study. The database, which comprises the two databases for the North Carolina (NC) and the Ohio (OH) field studies, contains the questionnaire data, the analytical data, and metadata. Sufficient detail was provided so that the data can be understood by a diverse set of users.

The CTEPP database is one of the largest current databases containing information on the environmental exposures of preschool children. The study's documentation, which includes the study design, Standard Operating Procedures, and Quality System Implementation Plan, will be placed in EPA's Environmental Information Management System (EIMS; http://oaspub.epa.gov/edr/eims$.startup). In addition, the metadata, which include abstracts, acronyms, keywords, and related entries will be placed in EIMS. The CTEPP data will be stored in the Human Exposure Database System (HEDS; http://www.epa.gov/heds/) . The CTEPP database will be made available to interested federal agencies, state and local agencies, non-governmental organizations, academia, and the general public.

7.2 Quality Assurance Procedures for the Database

Quality assurance and quality control(QA/QC) procedures were implemented within both the NC and OH databases. The QA/QC summaries are given in Appendix D. The following subsections provide information on the types of QA/QC procedures associated with the questionnaire data, analytical data, and metadata collected in this study.

7.2.1 Questionnaire Data

A comprehensive QA/QC plan was implemented to ensure data quality in all phases of questionnaire data collection. During the pre-data collection phase, each hard copy data form was tested by trained project staff for consistency and accuracy. Mock interviews and field data collection simulations were conducted to evaluate the effectiveness of the data forms. Once revisions were made to the data forms based on the outcome of these activities, final drafts were sent to EPA for review and approval. The data forms were further updated after receiving EPA's comments. The updated forms were reviewed and approved by the Battelle Institutional Review Board, the U.S. EPA Human Subjects Research Protection Official, and the U.S. Office of Management and Budget.

After final approval of the data forms, software components were programmed for use in the recruitment telephone survey and to allow double entry of the data. Standardized programming methods were used which inserted QC checks in all of these programs, including range checks, consistency checks, and skip pattern rules. These programs enforced the rules upon data entry. Before the programs were approved for actual data entry, they went through strict QA/QC checks for programming errors.

Before data collection began, telephone interviewers and field staff were trained in the study procedures, according to the SOPs, to ensure high data quality. Telephone interviewers were required to be certified for the study by passing a series of tests before they could initiate any contact with the study subjects. Training for the field data collection team members included a 40-h training session which incorporated at least one day of actual supervised field sample collection experience. Field staff were allotted additional time to practice their field data collection techniques.

After data collection began, data collection activities were monitored routinely. These efforts included the use of computer software and phone monitoring systems to monitor telephone recruitment data, and periodic internal field audits to ensure high quality of data collected in the field. In addition, external field audits were conducted by an EPA auditor and the EPA Task Order Project Officer (TOPO).

During field data collection, the field staff also reviewed the collected information while at the sampling site, to identify missing data items or questionable information. Any identified issues or problems were resolved at the sampling site before the field data collection team returned to Battelle. A Daily Activity Check List was also used to assist the field staff in conducting data collection activities and field edits.

After a data collection event at a sampling site was completed and the data forms were returned to Battelle, the receiving team conducted QC checks on each participant's data forms and study materials. The team used a Participant Data QC Check List to verify a standard list of important items. All data forms were then entered twice and verified, using the CTEPP Double Data Entry Program. Two data entry teams performed the data entry work and entered the data into two separate databases. These separate databases were compared for consistency, corrected if necessary, and combined into one database. As mentioned above, these data entry programs included range checks, consistency checks, and skip pattern rules.

Finally, after completing all the data collection tasks, the project staff conducted final QA/QC checks by reviewing data frequency reports and verifying randomly selected participant files. Data items in the database were checked against the data documentation manual and the actual participant data in the original data form. Personal identifiers were removed from the database to ensure participant confidentiality.

7.2.2 Analytical Data

Analytical data were electronically imported into the database according to CTEPP SOP 4.12. The analytical raw data (QUAN report) were generated from each instrument by a qualified analyst. The QUAN report was then reviewed by the Task Order Leader (TOL) for all the identified pollutants. The QUAN report was then electronically transferred into a custom report and saved as a "crd" file. The "crd" file was then electronically parsed into an Excel spreadsheet template. Data such as sample extraction weight and quality assurance codes were manually entered and saved as an Excel file with an extension of .xls by the first data reviewer. The TOL reviewed all the Excel files before they were imported into the analytical database. If any anomalous results were observed in the data, every effort was made to identify any problems in the sample collection, sample preparation, and/or analysis, which could have contributed to the anomaly. Data dictionaries and code sets for core analytical data, QA/QC data, and ancillary data were developed for the analytical database. The completed Excel spreadsheets were then electronically imported into the analytical database by the database staff.

Database queries were developed to perform QA/QC checks on the NC and OH analytical databases. These included (1) sample ID checks, (2) missing data checks, (3) duplicate data checks, (4) out-of-range checks, and (5) upper- and lower- concentrations checks.

The sample ID checks were performed to verify that all Sample IDs with reported data were valid Sample IDs, that is, that they were logged as being received from the field. If invalid sample IDs were detected, the database staff traced back to the original raw data, including laboratory record books and GC/MS logbooks, to identify the transcription error and to make the corrections accordingly. All corrections were documented in the database importing log book.

Missing data checks were performed to verify that all Sample IDs received from the field had a complete set of analytical data reported. Those samples that were received but did not have a complete set of analytical data and/or ancillary data for a stated reason in the electronic Chain of Custody (CoC) data were identified, and either the analytical data for these samples were found and imported into the database, or the samples were located, processed, analyzed, and reviewed, and the analytical data were imported into the database.

Duplicate data checks were performed to verify that the same analytical data were not imported into the database twice for a given sample. The database staff traced the sample results back to the laboratory record books, the GC/MS sequence logs, and/or the QUAN reports to confirm that duplicate data were the result of a double import, and not a QA/QC re-analysis (e.g. duplicate sample or duplicate injection). Once the duplicate data were identified as a double import, the set of results for the sample having the oldest sample import date were eliminated from the analytical database. If the duplicate data were identified as a QA/QC re-analysis, the proper QC code was added to the QC_Code data field, and the data for the first duplicate (only) remained in the Core_Analytical_Results table, and the data for the first and second duplicates were reported in the QA_QC_Results table.

Out-of-range checks were performed to verify that all data for data fields limited to a code set did not violate that code set. For data fields that were limited to a code set of values, queries were performed to identify data within those fields that did not belong to, or "violated", the code set. Once identified, the database staff traced the sample results back to the laboratory record books to identify the transcription error. The data in the database were corrected, and these corrections were documented.

Upper- and lower-level concentrations checks were performed on all results that were greater than plus or minus three standard deviations from the mean. Database queries were performed to identify those calculated results (Result1 and Result2) that were greater than or less than three standard deviations from the domain mean. Five percent of these data were reviewed again by the data reviewer. The data reviewer checked the QUAN report, all the parameters used for the results calculation, and the result calculation itself to make sure that identification and quantification were performed correctly. If the data reviewers detected any mis-identification and/or mis-quantification, corrections were made accordingly. The TOL approved the corrected data, and the database manager made the changes in the database. All activities were documented in the laboratory record books and database importing log.

After all checks were completed, the final calculation of results was performed within the database. A random subset (approximately 5%) of calculated results were recalculated using an independent calculation source (Excel) for validation. In addition, hand calculations were performed on one data set for each sample matrix using a calculator.

7.2.3 Metadata

Metadata were prepared in the format described in the "User Guide and Data Administration Guidelines for the USEPA's Environmental Information Management System (EIMS)," Version 1.3, Oct. 2001, including abstracts, related entries, key words, and acronyms, at the study, data table, and document level.

7.3 EPA Review

EPA conducted several independent QA/QC reviews of the early draft versions of the NC and OH databases. The EPA performed visual range checks. The data were normalized before identifying potential outliers. Outliers were identified based on whether they exceeded six standard deviations from the mean. For a randomly selected set of variables (about 5%), more extensive checks were performed including range checks, consistency, and skip pattern checks. When the EPA Database Manager or the EPA TOPO identified problems or errors (i.e., missing samples, duplicate samples, linkage problems) in the database, the EPA TOPO had Battelle verify that the data were correct and make any necessary changes to the data in the database.

After the draft final NC and OH databases were delivered to EPA, EPA conducted a more thorough QA/QC review of the databases. EPA repeated the extensive checks performed on the randomly selected variables. In addition, a new set of randomly selected variables

(additional 5%) were thoroughly checked. Furthermore, comparisons were made between earlier versions and the latest version of the database to assure that no unexplained changes had occurred. Any errors identified by EPA in the database were corrected by Battelle.

After EPA received the final versions of the NC and OH databases, EPA assigned data quality values to each sample in the Core_Analytical_Results table. The QA/QC protocol (SAS program) used to assess the quality of each sample is found in Appendix E. Each sample result was assigned one of the following data quality (QC_Flag code) values:

1 = good quality data
2 = questionable, but still acceptable data
3 = unusable data

Only sample results that had assigned QC_Flag code values of 1 or 2 were used in the statistical analyses discussed in Chapters 8 and 9. In addition, the data associated with one NC adult participant (PID972072) and accompanying child (PID972071) who withdrew from the study after Day 1 were excluded from the statistical analyses.

7.4 Evaluation

Within each record of the CTEPP database, the Participant Identification code (PID) was designed to be the key linking field that allows database users access to all of a given subject's study data, including questionnaire data and measured environmental and biological target compound levels. The PID was designed as a 6-digit composite data field. The first two digits contained the day care code (indicating whether or not the study participant attended day care, the specific day care, participant or non-day care participantand the state of origin). The next three digits were a unique participant identification number. The sixth digit contained the participant type code, which identified the sample as collected from a child or from an adult at home, or from a child at day care. Although the information contained within the PID was important, the first step database users had to do before querying the environmental, personal, or questionnaire data was to query the PID to separate it into its different pieces of information.

7.5 Recommendations

Based upon lessons learned from designing the CTEPP database, recommendations for designing large exposure databases on future studies are as follows:

(1) Avoid Composite Data Fields

Data fields should not be designed as a composite data field that contains several distinct pieces of data. If a piece of information is significant, it should be stored as its own separate data field. For example, the PID should actually have been separated into three separate data fields: 1) Day care ID, 2) Participant ID, and 3) Participant Type. Avoiding composite data fields would eliminate the need to write queries that separate out the bits of information contained within such fields, resulting in a more streamlined data extraction process.

(2) Design and Test Key Linking Fields

The key linking fields that allow a user access to all of the questionnaire, environmental, and personal data for a given subject should be planned and tested for a small pilot study prior to implementing them into a large study database.

(3) Add Link Tables for User Friendliness

Due to the complex nature of the CTEPP study design, it was not possible to have just one key field that linked all of the collected and calculated data for a given subject. As a result, several fields needed to be considered when bringing data together across all samples collected for a given subject. In the case of a study that has several "many-to-many" tables within its database (e.g., a single water sample is collected at a day care center, yet the analytical result for this sample is applicable to all study subjects attending the day care center, while conversely, a single study subject is associated with multiple samples such as urine, hand wipes, and food), an additional link table should be added to make the database more user-friendly. Designing a database containing many-to-many tables further complicates the relationships required to link the environmental and biological data with the questionnaire data. These relationships are not readily understood by those not intimately familiar with the study design. Link tables provide a user-friendly way of making use of the key linking fields without requiring the user to understand the relationships between those fields. An example of a useful link table is a table that lists all of the Sample IDs that are applicable for a given participant and makes the construction of the database queries that link environmental, biological, and questionnaire data much simpler.