EPA United States Environmental Protection Agency

Office of Research and Development

National Human Exposure Assessment Survey (NHEXAS)

Maryland Study

Quality Systems and Implementation Plan for Human Exposure Assessment

Emory University
Atlanta, GA 30322

Cooperative Agreement CR 822038

Standard Operating Procedure **NHX/SOP-D03**

Title: Lab Results Data Entry and Preparation

Source: Harvard University/Johns Hopkins University

U.S. Environmental Protection Agency
Office of Research and Development
Human Exposure & Atmospheric Sciences Division
Human Exposure Research Branch

Notice: *The U.S. Environmental Protection Agency (EPA), through its Office of Research and Development (ORD), partially funded and collaborated in the research described here. This protocol is part of the Quality Systems Implementation Plan (QSIP) that was reviewed by the EPA and approved for use in this demonstration/scoping study. Mention of trade names or commercial products does not constitute endorsement or recommendation by EPA for use.*

1. Title of Standard Operating Procedure
Harvard University/Johns Hopkins University Standard Operating Procedure:
D03 Lab Results Data Entry and Preparation, Rev 1.1

2. Overview and Purpose
The purpose of this SOP is to describe how lab results are organized and processed into the official database known as the Complete Dataset (CDS); to describe the structure and creation of the Analysis-ready Dataset (ADS); and to describe the structure and process of creating the worksheets which will be used in the analysis of duplicate samples and laboratory comparison.

3. Discussion
Two types of data collected that pertain to physical samples and the corresponding analysis of the samples: the lab results and logsheets. The handling of logsheets is described in SOP D04. Lab results, or parts thereof, will make up part of the CDS, the ADS, and the Quality Assurance Database (QADS). All information returned in the lab results will be stored in the CDS. The ADS will contain a subset of this data transformed so as to be useful for analysis. The data in the ADS will be the information necessary to carry out hypothesis testing of lab results. Both the CDS and ADS will be stored in *Paradox* database. The QADS will contain only data for quality assurance purposes, e.g co-located samples, field blanks, to estimate variability in field equipment and lab analysis. The QADS will be stored in Excel workbooks. A more detailed description of these databases are given below.

Structure of the CDS.
The CDS is the official database. The CDS will be used to create any database used for analysis, *e.g.* used to construct the ADS. Unlike the questionnaires, when data is requested from other investigators typically the ADS files will be sent because the measurements used in analysis will only be calculated at the ADS stage.

The results returned from each lab will be stored in a separate table named after the lab from which they came. These files will reflect exactly what is returned from the labs. The results from each lab will be stored in a separate data file. Since there will be little control in the amount of data returned in each data file, the data file itself will be considered the unit of data for lab results.

Structure of the ADS
The data frame of interest is the combination participant ID-cycle Number. The ADS for lab results will naturally be in that data frame. There will be two categories of ADS data files: media-based and compound-based. Media based files contain all the compounds analyzed for a particular sampling method, *e.g.* indoor air monitoring or water samples. Compound-based data files will contain the results for a specific compound regardless of the media sampled, *e.g.* lead in water, duplicate diet, dust, *etc.* will all be in the same ADS file.

The CDS lab data files will contain many variables which are needed only for lab purposes. The essential variables which are needed for the purposes of hypothesis testing are: Sample ID, exposure media, exposure compound, quality of result comment, numerical result, units of measure, and type of sample. These are the only variables carried through to the ADS.

Duplicate samples and results from labs other than the designated lab, which were generated only as a check on the designated lab, will not be included in the ADS. The calculations necessary to transform the lab results into the standard units of measurement used in discussing results will be done during the creation of the ADS.

The transformations of the CDS into the ADS are stored as Paradox queries. This makes creation and updating of the ADS a simple procedure which can be repeated quickly and easily as new data arrives and is entered into the CDS. Typically, links between the tables in the ADS will be on the participant-cycle level. The ADS is derived from the official database CDS, and unlike questionnaires the ADS for lab results will be considered an official, although not necessarly complete, database.

Duplicate Analysis and Lab Comparisons
Duplicate analysis studies the variability of measurements taken in the field. Only participant-cycle samples which have duplicate equipment will be included in this database. Each record will contain the variables: Participant ID, cycle, compound, media, original result and duplicate result. The data for each analyte from every media will be stored on a separate worksheet in the QADS.

Approximately [RB: KH questioned these numbers] 5-10% of field samples will be split and sent to at least two other labs besides the main lab for analysis. The main lab samples and outside lab samples will be recombined into one worksheet with the observations from different labs as different records. Variables that will be included are: Sample ID, cycle, compound, media, lab ID, and result. The data for each analyte from every media will be stored on a separate worksheet in the QADS.

File naming conventions
The file naming conventions are the same as for the questionnaires. See SOP D02 for details. The names and contents of the templates are given in Appendix A of this SOP.

4. Personnel Responsibilities

4.1 Project Data Coordinator

The Project Data Coordinator is responsible for
assuring accuracy and consistency of coding
modifying coding protocol if necessary, in writing
notifying the Data Input Supervisor of any modifications that affect working databases

4.2 Data Input Supervisor

The Data Input Supervisor is responsible for
training and supervising Data Input Assistants
tracking questionnaires and other forms through coding and review
resolving problems and ambiguities that Assistants are unable to handle

4.3 Data Input Assistant

Data Input Assistants are responsible for
logging in arriving data, checking for ID, sorting forms
coding
reviewing coding
The same Assistant may do both coding and review, but no one may review a file s/he has coded.

5. Required Equipment

- computer with database software having appropriately labeled fields for entering:
  (a) the Data Input Assistants' initials and the dates when the file is coded and reviewed
  (b) codes next to the corresponding data (in the data field if the software will accept numbers and letters in the same field, otherwide in a code field next to the data field)
- data in magnetic format
- data in hardcopy format (e.g., photocopies of field logsheets or laboratory logbook pages) gathered into stacks (all within a stack are the same type)
- disks for data and backups. The system of filenames will tell the status of the file, e.g. coded but not reviewed. The first 8 digits of a filename will be its ID, and the last 3 will indicate its status, e.g., .COD, .REV, etc.
- stack card for each stack (large card that always stays with stack; identifies stack and has spaces for initials and dates when it is coded and reviewed) [similar to UA ownership card]
- coding and entry log (forms with initials and dates for coding and review of each stack of forms; kept in Supervisor's logbook)
  [copy of UA form attached to SOP "Coding ... Hand-Entered Data"]
- coding instruction sheet appropriate to the type of data being coded
- pen, purple (to distinguish coding from notes made by field staff, which are in black)
  [UA colors; don't know what color lab staff use]
- post-it notes

6. Procedure

6.1 When data arrive from the field and analytical laboratories in either magnetic or hardcopy format, a Data Input Assistant will log them in and check them to make sure that they have all the necessary identification. When samples arrive from the field, logsheet data are coded before the samples are analyzed, in order to catch any samples that are so seriously compromised that they need not be analyzed.

Each field logsheet will have a label identical to the label on the corresponding sample, showing the ID number in both bar-code and human-readable format.

Each computer file of analytical data will have a filename corresponding to the sample batch identification code.

Then paper forms are sorted by type and separated into stacks, each with a stack card.

A stack contains the forms corresponding to the data in one computer file.

6.2 The Data Input Supervisor assigns each stack and the corresponding file to a Data Input

Assistant for coding.

6.3 The Data Input Assistant who will code a file takes the disk, stack, stack card, pen, and instruction sheet; and marks the card and log with his/her initials and the date. S/he opens the file on a computer and types his/her initials and the date into the appropriate fields.

The Assistant codes the stack according to the instructions, finding each datum in the file and noting whether the paper form has any notes or flags. Wherever there are notes or flags, the Assistant types the appropriate code into the code field.

If there are any problems or ambiguities, the Assistant marks each one on the paper form with a post-it with an arrow and note pointing out the location of the problem; and types the ambiguity code "?" into the corresponding field.. When the stack is finished, the Assistant makes a backup copy of the file and submits the stack to the Supervisor.

6.4 The Data Input Supervisor assigns the stack to a different Assistant to review the coding.

6.5 The Data Input Assistant who will review the coded file takes the disk, stack, stack card, pen, and instruction sheet; and marks the card and log with his/her initials and the date. S/he opens the file on a computer and inserts his/her initials and the date in the appropriate fields.

The Assistant reviews the coding according to the instructions. Where a problem or ambiguity is marked, the Assistant resolves the problem or asks the Supervisor to resolve it, and replaces the ambiguity code with the appropriate code. When the file is finished, the Assistant backs it up and submits the stack to the Supervisor.

4. Personnel Responsibilities

For responsibilities on getting Samples to the appropriate lab and results shipped to Emory see the relevant Field SOP.

4.1 Project Data Coordinator is responsible for
·assuring accuracy and consistency of database
·modifying data entry protocol if necessary, in writing
·tracking samples from collection to entry into database to storage.
·notifying the Data Entry Supervisor of any modifications that affect working databases.

4.2 Data Entry Supervisor is responsible for
·training and supervising Data Entry Assistants
·tracking disks and hard copy through data entry and review
·resolving problems and ambiguities that Assistants are unable to handle.

4.3 Data Entry Assistants are responsible for
·logging in arriving data, checking for ID, storing disks and hard copy of data
·data entry
·reviewing data entry.

5. Required Equipment and Reagents

- set of questionnaires and pen for marking the questionnaires.

- computer with Borland Paradox V5.0 for Windows and database directories
- data in magnetic format returned from the Lab
- data in hardcopy format
- CDS backup disks.
- data entry instruction sheet appropriate to the type of data being entered
- data entry log (forms with initials and dates for data entry; kept in Supervisor's logbook)

6. Procedures

Data Preparation at Emory

On receipt of each data file from a Lab the Data Entry Assistant will take the following steps to create the CDS and store new data:

1. The file will be copied to the archive directory. The file will be renamed if necessary to reflect the NHEXAS file naming standards.
2. If the file is renamed the file will be copied onto the same disk as the original file with the new name for cross-reference.
3. The disk with the original data will be stored in the designated location for archived disks. See SOP D01 for more details.
4. The data will be imported to the relevant table of the CDS and the resulting file will be renamed to identify it as the latest version of that table. Details are given below.
5. The new CDS will be copied to the CDS back-up disk. See SOP D01 for more details.
6. The old version of the CDS will be removed from the CDS directory so only the current version of the CDS is available.
7. A record of completion and cross-referencing of file names will be entered into the data entry logbook.

Step 4: Entering new data units to the CDS

New records are added to copies of the latest version of the tables by importing the new records into a template table and then using the Add/Append functions to put the results temporarily stored in the template table into the new version table. An introduction to these methods can be found on p. 180 and p. 187 of the *Paradox* User’s Guide.

The procedure will be the same for all tables. The steps in preparing the CDS tables for questionnaire are:

1. Gather all new units of initial data files to be entered. Then for each new unit of data:
2. Rename a copy of the last version of the CDS table to identify it as the latest version of the CDS table.
3. Import the new data to the relevant template table by using the Tools|Utilities|Import commands.
4. Take the newly created table from Step 3 and use the Tools|Utilities|Add command to add the new table to the latest version of the relevant CDS table created in Step 2.
5. Clear the template table by using the Table|Empty command.

Correcting the CDS

See Quality Assurance Procedures for methods used to detect errors in the CDS. If an error is observed by the Data Entry Assistant, they will notify the Data Entry Supervisor who will notify the Project Data Coordinator. The Project Data Coordinator will decide whether the correction will be made and will notify the Principle Investigator about the decision. If the CDS needs to be corrected a renamed copy of the CDS will be made as described in the File Naming subsection of the discussion above. An example: if the CDS name is techw4.db then the renamed copy will be techw4a.db. The Data Entry Assistant will make the correction directly to the renamed CDS table and the nature of the correction and the reason for the error if known will be noted in the file qxcom.db. The updated file and the questionnaire comment file will be copied and stored in the designated official database location. The old version of the CDS will then be archived and removed from the CDS directory. See SOP D01 for storage details.

Creating the ADS
For each ADS table and the duplicate analysis and lab comparison files (see description of procedures below) there is a short series of queries which need to be run. The number of queries is dependent on the ADS table being constructed. The queries are already constructed and have names *compoundx*.qbe, *mediax*.qbe, d*yzx*.dbe and l*yzx*.qbe, where *x* is the number in the series of queries for the respective table being created; *compound* is the type of compound ADS being created, *e.g* lead; *media* is the type of media ADS being created, *e.g.* water; d*yzx*.qbe leads to the duplicate analysis workbook for compound *y* in media *z*; and l*yzx*.qbe leads to the lab analysis workbook for compound of compound *y* in media *z*. The wild cards *y* and *z* represent three letter codes identifying compound and media respectively.

Starting from the CDS the steps in preparing the media or compound tables for ADS are the same for each ADS table:

1. Using the latest version of all CDS files start with query 1 for the ADS file being constructed. Repeat for each query in the series running the queries in increasing numeric order.
2. After all queries have been run rename the final answer table to reflect the version of the ADS. See the file naming conventions in the discussion.
3. Remove the intermediary files which are created during the process.

Duplicate Analysis and Lab Comparisons
The methods are the same for every compound in every media. Assume that we are creating the duplicate analysis data set for compound *y* in media *z*. Starting from the CDS the steps in preparing the lab results for the duplicate analysis workbook are:

1. Run queries dup*yzx*.qbe.
2. Use the Tools|Utilities|Export commands to export data to Excel 3.0/4.0 worksheet.
3. Repeat for all compounds in all media.

7. Quality Assurance Procedures
All personnel have appropriate training, see SOP D04.

To guarantee no complete records have been missed, the number of records in the initial data files will be checked against the number of individual's responding to questionnaires for that unit of data. If the number of records do not match an investigation will be undertaken to determine the reason for the discrepency.

To guarantee no complete records have been missed, the Data Entry Assistant will check the number of records in the initial data file, which contains the new data, and the number of records in the old version of the official database to make sure the sum equals the number of records in the new version of the database. Discrepencies will be checked by looking directly at the initial data file sent by the Data Entry Company to see which records may have been lost or duplicated. If no error is observed the Data Entry Assistant will check to make sure that all old records were correctly incorporated into the new file.

At the conclusion of the study, a comparison of previous versions of the official database with the final version of the official database will be undertaken. Any discrepencies will be checked.

See SOP D05 for details on the use of exploratory data analysis in detecting unusual values. Duplicate analysis and lab comparison results can also be used as a spot check of results, since large discrepencies between co-samples and split samples will first be checked for incorrect data entry.

8. References

Borland Paradox Relational Database V5.0 User's Guide and Online Help.
Harvard University/Johns Hopkins University Standard Operating Procedure:
D01 Data Flow Procedures, Rev 1.1
Harvard University/Johns Hopkins University Standard Operating Procedure:
D02 Questionnaire Data Entry and Preparation, Rev 1.1
Harvard University/Johns Hopkins University Standard Operating Procedure:
D04 Logsheet and Confidential Questions Data Entry and Preparation, Rev 1.1
Harvard University/Johns Hopkins University Standard Operating Procedure:
D05 Exploratory Data Analysis and Summary Statistics, Rev 1.0

Appendix A - List of Lab Data files and description of contents

See Appendix A of the Data Management Qsip for a list of what results are coming from the different labs.

The CDS files

cdc.db -	Lab results from the CDC.
fda.db -	Lab results from the FDA.
herl.db -	Lab results from the EPA NHEERL.
hsph.db -	Lab results from Harvard School of Public Health.
swri.db -	Lab results from Southwest Research Institute.

The ADS files

airi.db -	Indoor Air - all compounds.
airo.db -	Outdoor Air - all compounds.
airp.db -	Personal Air - all compounds.
blood.db -	Blood Samples - all compounds.
derml.db -	Dermal Wipe Samples - all compounds.
dietb.db -	Duplicate diet: Beverages - all compounds.
diets.db -	Duplicate diet: Solids - all compounds.
dust.db -	Dust samples - all compounds.
markb.db -	Mini-market Basket Survey: Beverages - all compounds.
marks.db -	Mini-market Basket Survey: Solids - all compounds.
soil.db -	Soil samples - all compounds.
urine.db -	Urine Samples - all compounds.
water.db -	Water samples - all compounds.

Codes for duplicate analysis and lab comparison queries.

METALS	PESTICIDES	PAHs	MISC.
pb - Lead	cpy - chlopyrifos	ben - benzo(a)pyrene	lip - lipids
cr - Chromium	chl - chlordane	chr - chrysene	voc - VOCs
cd - Cadmium	die -dieldrin	ant - anthracene	cre - creatinine
as - Arsenic	mal - malathion	phe - phenanthrene	
	ddd - 4-4-DDD		
	dde - 4-4-DDE		
MEDIA	ddt -DDT		

die - Duplicate diet
wat - water
dus - dust
soi - soil
blo - blood
uri - urine
der - dermal wipe
ias - indoor air sample
oas - outdoor air sample
pas - personal air sample
mar - mini-market basket survey.