National Human Exposure Assessment Survey (NHEXAS)

Arizona Study

Quality Systems and Implementation Plan for Human Exposure Assessment

The University of Arizona
Tucson, Arizona 85721

Cooperative Agreement CR 821560

Standard Operating Procedure **SOP-UA-D-26.0**

Title: Electronic Data QA Check (Hand Entry & Scanned)

Source: The University of Arizona

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.*

Electronic Data QA Check

1.0 Purpose and Applicability

As per SOP# UA-D-16.0, a data accuracy check is to be performed on a randomly selected 10% sample of all electronic data. This SOP gives details and a procedure for such checking. It applies to the cleaned, working databases generated by NHEXAS Arizona.

2.0 Definitions

DATA, ELECTRONIC: Data stored on some type of magnetic or optical medium (for example: floppy disk, hard disk, Bernoulli, tape).

DATA, PHYSICAL: A datum or data written on a physical data form.

DATA CLEANING: The process of locating and correcting data processing errors (see DATA PROCESSING ERROR below). They can be individual level errors in the electronic and physical data, or they can be system level errors in the data collection, packaging, coding, entry, and cleaning procedures themselves. This process is also referred to as "data validation."

DATA PROCESSING BATCH (DP BATCH): A collection of household packets or physical data forms reviewed for quality assurance and ready for data entry. Each DP batch receives a unique numeric or alphanumeric code that is written on all forms in the DP batch and is entered into the database corresponding to that form.

DATA PROCESSING ERROR: An error occurring at any level of data processing. It is a procedural mistake, such as a duplicate data record, a typographical error, a logical error, or missing information.

DATA RECORD: In the context of this SOP, this is a row of electronic data in a database.

DATABASE, MASTER: Accumulative database generated from the appendage of newly cleaned data processing batches. *Copies* of master databases are used in analyses and any data corrections made to copies are also made to the master databases. Thus, a master database is the most complete and accurate database of its kind.

DATABASE, WORKING: A database earmarked for or in the process of cleaning that contains one or more data processing batches. When cleaned, it will be appended to the master database.

FIELD: An area on a data entry form (i.e., screen) where a datum from a physical form is entered (see FORM, DATA ENTRY below).

FORM, PHYSICAL = The paper or "hard copy" version of a data form. This is also referred to as a "physical data form."

HRP SITE: The Health Related Professions building, located at 1435

North Fremont Avenue; Tucson, AZ 85719. This is an annex of the Respiratory Sciences Center and the primary site of NHEXAS Arizona.

NHEXAS Arizona: Acronym for National Human EXposure Assessment Survey, a research project conducted in Arizona by the University of Arizona/Battelle/Illinois Institute of Technology consortium.

3.0 References

SOP# UA-D-16.0

4.0 Discussion

4.1 General Introduction

A data accuracy check is to be performed on a randomly selected 10% sample of all electronic data. This check is to confirm the accuracy on two levels:

4.1.1 Correspondence between the electronic data and the physical forms, and
4.1.2 The cleanness of the electronic data (per the cleaning of H&E Data Assistant Students as described in SOP UA-D-16.0).

Each of these checks will involve examining each field in each data record and confirming its correctness (relative to the current level of inquiry). Thus, each field will be checked for correspondence to its physical form as well as its correctness within its context.

In an attempt to expediate the process, the two levels of checking will be performed simultaneously (in one phase). This will avert the need to re-examine each field for each level.

4.2 The Two Levels

4.2.1 Typing Accuracy

This level of checking examines the correspondence between the electronic data and the corresponding physical forms. Note that this check should be done with respect to the coding of the physical form and not the actual responces given by the subject. Also, if there are multiple codes given for a particular field, then the latest code should be used (as it is the most correct with respect to the current coding protocols).

Thus a data field is considered to be invalid if it disagrees with the most correct coding on the physical form. If an invalid field is found, it should be recorded on a LEVEL ONE ERROR REPORT form (Figure 2), then fixed in the database (this may require filing a Working Database Change Log Form or Master Database Change Log Form).

4.2.2 Data Cleaning Accuracy

This level of checking examines the cleaness of the data in its context (i.e., does each field in a given record fall within the range of acceptable values for that particular instance?). A data field is considered to be invalid if one of the following holds:

(a) The field is null
(b) The field's code is invalid (i.e., the code does not exist)
(c) The field was miscoded (i.e., the wrong code used)
(d) The field's code is logically incorrect (i.e., the field is dependent on another field, and its value is contradicted by that field)

If an invalid field is found, it should be recorded on a LEVEL TWO ERROR REPORT form (Figure 3), fixed in the database, and the coding changed on the physical form (this may require filing a Working Database Change Log Form or Master Database Change Log Form).

5.0 Responsibilities

5.1 The Project Data Manager is responsible for:

5.1.1 Approving suggested corrections to the master database, and

5.1.2 Making any approved corrections to the master database.

5.2 Student Data Assistants are responsible for performing the procedures detailed in this SOP.

6.0 Materials and Reagents

6.1 Materials

6.1.1 Cleaned Electronic Data
6.1.2 Physical Forms
6.1.3 Two-Level Checking Summary forms
6.1.4 Level One Error Report forms
6.1.5 Level Two Error Report forms
6.1.6 Quality Assurance Check for Electronic Data forms

6.2 Equipment

6.2.1 Hardware:

It is necessary to work on the LAN. This requires use of PC that is a workstation in the Respiratory Science Center LAN network where the Sun SparcStations "Ipomea" and "Lonicera" are network nodes, as well as access to one of these SparcStations (not necessarily from the console).

6.2.2 Software:

(a) R:Base for DOS, Version 2.11
(b) count (UNIX binary)
(c) gencheck (UNIX binary)
(d) percent (UNIX binary)
(e) percent (UNIX binary)
(f) rbapp (UNIX binary)

6.3 Reagents

None

7.0 Procedure

7.1 Fill out Header Information on a Quality Assurance Check for Electronic Data form.

7.2 Randomly select a 10% sample of the current data.

7.2.1 From R:Base, unload all data to be considered into a file using the "unload data" command ("as ascii").

7.2.2 If the data records vary in size (i.e., in the number of elements), break up the file into components based on the length of each record.

7.2.2 Convert the DOS files into UNIX files.

7.2.3 If there are multiple data files (obtained by step 7.2.2), determine the length of the longest record and run "rbapp" on each of the other files to make all records the same length. Concatenate the files and obtain one large file.

7.2.4 Count the number of records to be considered.

7.2.5 Run "10_percent" on the data file to obtain a 10% sample and pipe the results into a target file.

7.3 Fill out a Two-Level Checking Summary form, gathering the information from the file obtained in step 7.2.5.

7.4 Run "gencheck" on the file obtained from step 7.2.5 and convert it into DOS format.

7.5 Enter R:Base (from DOS), and "run" the DOS file obtained from step 7.4. For each data field in each data record displayed:

7.5.1 Assure that the electronic code matches the most recent code given on the physical form (if there is a single code given, the electronic data should match it; if there are multiple codes given, they should each be dated, and the electonic data should match the most recent coding). If an invalid variable is found, a Level One Error Report form must be filled as well as either a Working Database Change Log Form or a Master Database Change Log

Form.

7.5.2 Assure that the field does not fall under one of the following catagories:

(a) The field is NULL (-0-).
(b) The field has been assigned an invalid code (i.e., a code that does not exist).
(c) The field has been coded inappropriately (i.e., the code does not match the form response).
(d) The field has a logically inconsistent code (i.e., the field depends on some other variable, and the code given cannot occur under the circumstances dictated by that variable).

If such a variable is found, a Level Two Error Report form must be filled as well as either a Working Database Change Log Form or a Master Database Change Log Form.

7.6 Fill out Part I of the Quality Assurance Check for Electronic Data form. If the error rate is less than 5%, then the process is complete. Otherwise, proceed to step 7.7.

7.7 Run "50_percent" on the file obtained in step 7.2.2 (or 7.2.4, if applicable) to obtain a 50% random sample and pipe the output to a file.

7.8 Run "gencheck" on the file obtained from step 7.7 and convert it into DOS format.

7.9 Fill out a Two-Level Checking Summary form, gathering the information from the file obtained in step 7.8.

7.10 Enter R:Base (from DOS), and "run" the DOS file obtained from step 7.8. For each data field in each data record displayed, follow steps 1 and 2 as described in 7.5. Then, record the results on a Two-Level Checking Summary form.

7.11 Fill out Part II of the Quality Assurance Check for Electronic Data form. If the error rate is less that 5%, then the process is complete. Otherwise, proceed to step 7.12.

7.12 Re-enter and re-clean the data. Re-apply this procedure to the new data and record the results in Part III of the Quality Assurance Check for Electronic Data form.

7.13 Quality Control

7.13.1 Tolerance Limits

All working databases generated by NHEXAS Arizona will be QA checked according to the rules and procedures outlined in this SOP.

7.13.2 Detection Limits

The "Two-Level Checking Summary," "Level One Error Report,"

"Level Two Error Report," and "Quality Assurance Check for Electronic Data" forms will enable detection of any errors made during the electronic data QA check.

7.13.3 Corrective Actions

Any data not receiving a proper electronic data QA check will be brought to the attention of the Project Data Coordinator. She or he will then assign the task to the appropriate person.

8.0 Records

8.1 Data to Be Recorded from This Procedure

N/A

8.2 Record Forms (attached)

8.2.1 Figure 1: "Two-Level Checking Summary" form
8.2.2 Figure 2: "Level One Error Report" form
8.2.3 Figure 3: "Level Two Error Report" form
8.2.4 Figure 4: "Quality Assurance Check for Electronic Data" form

8.3 Location of Record Forms

The forms listed in section 8.2 above are filed in the "QA Checks on NHEXAS Arizona Databases" notebook in the lobby of room 128 at the HRP site.

Figure 1: Two-Level Checking Summary

Page:__ of __

Physical Form Name:_____

Database Name:_____

Data Entry Form Name:_____

Description of Data:

[] DP Batch #:__

[] Master Database

[] Other:_____

Key Variables	Valid	Failed Level 1	Failed Level 2

FORM D26-1

Figure 2: Level One Error Report

Page:__ of __

Physical Form Name:_____

Database Name:_____

Data Entry Form Name:_____

Description of Data:

[] DP Batch #:__

[] Master Database

[] Other:_____

Key Variables	Element	Problem/ Solution

Form D26-2

Figure 3: Level Two Error Report

Page:__ of __

Physical Form Name:_____

Database Name:_____

Data Entry Form Name:_____

Description of Data:

[] DP Batch #:__

[] Master Database

[] Other:_____

Key Variables	Element	Problem/ Solution

Form D26-3

Figure 4: Quality Assurance Check for Electronic Data

Page 1

Physical Form Name:_____
Database Name:_____
Data Entry Form Name:_____

Description of Data:
[] DP Batch #:_____
[] Master Database
[] Other:_____

I. 10% Random Selection

A. Total data records: _____
B. Total data records selected: _____
C. Total elements per record: _____
D. Total data records with error(s): _____
E. Total elements with error(s): _____
F. Record error rate: _____
G. Element error rate: _____

QA Check By: _____ Date:_____
Staff Initials:_____ Date:_____

Comments:

II. 50% Random Selection (If Applicable)

A QA check is completed for 50% of the data described on this form if the error rate in step I.G (above) is greater than five percent.

A. Total data records: _____
B. Total data records selected: _____
C. Total elements per record: _____
D. Total data records with error(s): _____
E. Total elements with error(s): _____
F. Record error rate: _____
G. Element error rate: _____

QA Check By: _____ Date:_____
Staff Initials:_____ Date:_____

Comments:

FORM D26-4

Quality Assurance Check for Electronic Data
Study of Health and the Environment

Physical Form Name:_____
Database Name:_____
Data Entry Form Name:_____

Description of Data:
[] DP Batch #:_____
[] Master Database
[] Other:_____

III. 100% Data Re-Entry (If Applicable)

If the error rate in step II.G (Page 1) exceeds five percent, then the data described on this form must be completely re-entered and re-cleaned. Once this has been done, record the results of the QA check below:

A. Total data records: _____
B. Total data records selected: _____
C. Total elements per record: _____
D. Total data records with error(s): _____
E. Total elements with error(s): _____
F. Record error rate: _____
G. Element error rate: _____

QA Check By: _____ Date:_____
Staff Initials:_____ Date:_____

Comments:

FORM D26-5

Addendum 1 - Using The UNIX Binaries

Program Usage:

```
  count <character> [< input_file] [> output_file]

  gencheck <formname> [< input_file] [> output_file]

  10_percent <# records> [< input_file] [> output_file]

  50_percent <# records> [< input_file] [> output_file]

  rbapp <cur # cols> <delimiter> <append string>
        [< rbase_unload_file] [> output_file]
```

Where

```
  <xxx>  denotes a mandatory string,
  [xxx]  denotes an optional string
  "<"    denotes pipe from input file
  ">"    denotes pipe to a file
```