Data management and editing approaches for solving data challenges in the National Children’s Study Vanguard Study

Linda Andrews
Health Policy and Data Analytic, Social & Scientific Systems, Silver Spring, Maryland, USA

BACKGROUND

The National Children’s Study (NCS) Vanguard Study was a pilot for a planned large-scale epidemiological birth cohort study of children including parents to investigate the influences on child health and development. After the initial household enumeration, recruitment strategy was modified to reduce follow-up burdens on child and families. The original data collection period was from mid-2006 to mid-2011. Primary data collection methods included in-person visits and data reporting using a Web-based venue and over-the-phone interviews. The data quality assurance (QA) process was designed to ensure feasibility, acceptability, and cost of four alternate strategies for recruiting a large number of participants in the Vanguard Study. The methods included used to manage the development of the NCS Vanguard submission files. This allowed for the data collection and reporting of the NCS submission files by submitting a file for inclusion to the database. With each new version of the Vanguard Workbench, the data quality challenges were expanded to include newly identified gaps within the data management team and reporting team resulted in ongoing improvements in data quality.

DATA ELEMENT SPECIFICATIONS (MDES)

• 15,622 data elements
• 624 Instrument tables for the 77 Vanguard Study Centers
• 86 Operational tables storing 712 datasets
• 540 tables for data linking and specimen collection
• 12 tabular files
• 1228 Study Center data files
• 1386 SAS log files
• 184 SAS macro files
• 1550 SAS data set files

The MDES was a comprehensive catalog of the data elements, including metadata and code lists, that were defined for data collection and submission for the study. The MDES was a hierarchical document that provided a complete definition of the data elements, including their data type, data format, system of units, description, mandatory status, and other information.

METHODS

To evaluate the data quality of data submissions by users of the different IMSs, data quality summaries were generated for each user. The data quality checks for each user’s data were incorporated into the Vanguard Workbench, a tool that supported not only the MDES Study Center activities and data management/editing methods and related data elements used to manage the development of the NCS Vanguard submission files. This allowed for the debugging and diagnosis of problems with the NCS submission files prior to submitting a file to the submission server database. With each new version of the Vanguard Workbench, the data quality challenges were expanded to include newly identified gaps within the data management team and reporting team resulted in ongoing improvements in data quality.

DEMOGRAPHICS AND DUE DATE

• 2.8 MB in the National Children’s Study Vanguard Study

igor.layer("Layer-0").visible = false

Figure 1: Younger siblings (Y.S.) born between 1999 and 2002, older siblings (O.S.) born before 1999 and postnatal data (P.D.) born after 2002. The classification of these three groups by their birth years is shown in Table 1.

SUBMISSION FILE SPECIFICATIONS

• 46 Operational tables listing unique column definitions for each instrument or event type, including case level identifiers, data imputation rules, and other information.
• 375 tables for the 77 Vanguard Study Centers
• 188 event types
• 38 instrument versions
• 948 datasets
• 4277 data elements
• 1550 SAS data set files

• 1550 SAS data set files

The MDES Information Table is a comprehensive catalog of the data elements, including metadata and code lists, that were defined for data collection and submission for the study. The MDES Information Table provided a comprehensive definition of the data elements, including their data type, data format, system of units, description, mandatory status, and other information.

WORKBENCH TOOL

• WebWorkbench
• File Processing
• WebPages
• Data Management
• Data Editing

The Vanguard Workbench Tool was a Web-based application that provided access to the data management and editing tools for the NCS. The WebWorkbench Tool was designed to be a centralized, dynamic, and user-friendly tool for managing data. It included features such as data browsing, data validation, and data editing, as well as tools for data management and reporting.

STAGING

• 1. Create new Primary Key Variable by concatenating the
• 2. If any of the absolute differences were more than 28 days, the
• 3. The absolute difference between the

The staging process was the first step in processing the data. The primary key variable was created by concatenating the values of the primary keys from the different IMSs. This was necessary to ensure the uniqueness of the primary key. It was also necessary to maintain the integrity of the data.

CONSISTENCY FOR STATIC DEMOGRAPHIC CHARACTERISTICS

• 1. Create new Primary Key Variable by concatenating the
• 2. If any of the absolute differences were more than 28 days, the
• 3. The absolute difference between the

The consistency check for static demographic characteristics was used to ensure that the demographic data were consistent across all sources. It was necessary to verify that the data were consistent across all sources and that they were not contradictory or inconsistent.

How to handle removing data for confidentiality of PII?

• Replace with the weeks pregnant, gestational age or child’s age on the date of interest.

This provided complete and consistent data as possible per participant rather than relying on an individual data element. The post-recruitment follow-up data were stored on analytic summary files such as the study visit files or the participant demographic file. This allowed easy retrieval of the information for all reports and yielded consistent demographic counts by all team members and analysts.

PREGNANCY TO CHILD CROSSWALK

• Establishing accurate linkages between a child and theassociated pregnancy is crucial for the accurate capture of data.
• The absolute difference between the

It was necessary to verify that the linkages were correct. This provided complete and consistent data as possible per participant rather than relying on an individual data element. The post-recruitment follow-up data were stored on analytic summary files such as the study visit files or the participant demographic file. This allowed easy retrieval of the information for all reports and yielded consistent demographic counts by all team members and analysts.

MAINTAINING CONFIDENTIALITY FOR DATE FIELDS

• The 5-year course of operations, data collection activities and data submissions were migrated from the SCs to four Regional Operations Centers (ROC).
• During the 5-year course of operations, data collection activities and data submissions were migrated from the SCs to four Regional Operations Centers (ROC).
• During the 5-year course of operations, data collection activities and data submissions were migrated from the SCs to four Regional Operations Centers (ROC).
• During the 5-year course of operations, data collection activities and data submissions were migrated from the SCs to four Regional Operations Centers (ROC).