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

Linda Andrews

Health Policy and Data Analysis, 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 (and their parents) to examine the environmental influences on child health and development. After the initial household enumeration recruitment strategy yielded projections for cost and duration greater than expected, NCS Vanguard Study developed and evaluated the feasibility, acceptability, and cost of four alternate strategies for recruiting a large prospective national probability sample of pregnant women and their newborn children (1). The Initial Vanguard Study (IVS) and the Alternate Recruitment Substudy (ARS) and the post-recruitment follow-up (PRF) were conducted between 2009 and 2014 and ultimately enrolled and followed over 5400 birth families in 43 counties across 31 states throughout the United States.

This presentation focuses on the many challenges with data management and editing for the ARS Phase and the post recruitment follow-up data collection activities. The ARS Phase simultaneously evaluated a decentralized model for data collection which utilized several different information management systems (IMSs) to test for potential alternatives to the IVS's centralized electronic case management system (1). This decentralized approach introduced many inconsistencies and data quality issues across the 40 Study Centers. Additionally, the study fielded multiple instrument versions for the study visits. These versions needed consolidation and harmonization to produce instrument files useful for analysis. Forty study centers submitted up to 700 datasets bi-weekly, totaling 3.5GB by the study end.

This paper covers the tools and methods used to validate study data submitted and the post-submission data management and editing to prepare datasets for future use.

METHODS

MASTER DATA ELEMENT SPECIFICATIONS (MDES)

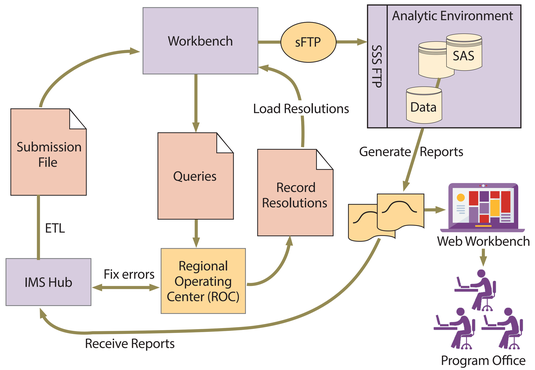
During the ARS Phase, Study Centers (SCs) were tasked to build their own infrastructure to collect data utilizing either their existing study management and data collection systems, adapting new technology, or working with other SCs. The goal was to deliver a set of pre-defined data to the Vanguard Data Repository (VDR) for analysis via Extensible Markup Language (XML). The MDES, an excel-based metadata tool, stored information for all 15,000 data elements, code lists, release notes, version information, and more to inform Study Centers on the central study protocol (2).

MDES Information

- 712 datasets
- 86 Operational tables storing linkages, contact information, event dispositions, consent data, demographics, sampling frame information, study center information, including staff roster and specimen collection information.
- 624 Instrument tables for the 77 instruments fielded for 39 event types covering recruitment, pregnancy, birth, post-natal through 42 months.
- 15,622 data elements
- 12 MDES versions over 4 years of data collection

VANGUARD WORKBENCH TOOL

To evaluate the data quality of data submissions by users of the different IMSs, data quality summary reports were generated on key variables. The data quality checks from these reports were incorporated into the Vanguard Workbench Tool, a local tool that supported NCS Study Center activities by providing validation methods and all related metadata used to manage the development of VDR XML submission files. This allowed for the debugging and diagnosing of problems with the XML submission files prior to submitting a file for inclusion to the database. With each new version of the Vanguard Workbench, the data quality checks were expanded to include newly identified gaps within the data or inconsistencies. The coordination of the Vanguard Workbench development team and the data management/reporting team resulted in ongoing improvements in data quality.



STACKING

The first challenge in processing the data once received by the VDR and converted to SAS was to consolidate the instrument versions. Using information from the MDES, the data analysis team constructed and maintained a master table to store the metadata necessary to append instrument tables with multiple versions into a single table (e.g, the four versions of the Birth Visit Instrument are appended to a single dataset).

1. Create new Primary Key Variable by concatenating the Source Table character string
   a. This ensures uniqueness of the primary key
   b. Easy method to identify source table
   c. Maintains existing linkages among the associated tables
2. Recode data elements with different code lists across the instruments. Use the information captured in the metadata master table to retain reference to original instrument table:

   New Value = Value + [Table Order/1000 (if value > 0) or

   Value = value – order/1000 (if value < 0)]
3. Data elements with different attributes are renamed with the version number concatenated to the variable name

   (e.g., *varname -> varname_V1* from Version 1).
4. Update all formats so that the new values links to their original formats and code lists.
5. Append the datasets per the metadata master table.

Data Challenges and Solutions

Issue	Solution
How to address data quality across multiple IMS users	Provide local tool to evaluate data prior to submission and include robust data quality checks (see Vanguard Workbench).
How to consolidate multiple instrument versions	Stacking Algorithm to append multiple versions into one dataset while maintaining information on the source file and formats.
Establishing universal identifiers across multiple Study Centers	Establish Universal Participant ID and distribute pools of potential IDs to SCs and store crosswalks between identifiers.
Determining Due Date	Build algorithm to cluster due dates within a short range and delete others as potentially related to different pregnancies or entry errors. Select last reported value from this cluster of valid dates.
Linking children to pregnancy	In future studies, we recommend establishing pregnancy identifiers to enable easily linking children to the corresponding pregnancy.
Consistency for Static Demographic Characteristics	Maintain participant demographic file with last reported values for demographic characteristics.
How to handle removing date fields for confidentiality of PII	Replace with the weeks pregnant, gestational age or child's age on the date of interest.

PREGNANCY TO CHILD CROSSWALK

Establishing accurate linkages between a child and the associated pregnancy is crucial for analysis, especially since not all pregnancies resulted in live births.

Constant validation of dates were necessary to verify that the linkages were correct.

Establishing a pregnancy identifier and maintaining a crosswalk to the child's ID, or even using that as the child's identifier, could solve this problem.

UNIVERSAL PIDS AND CHECK DIGITS

During the 5-year course of operations, data collection activities and data submissions were migrated from the SCs to four Regional Operating Centers (ROCs), each with a different IMS, with data still submitted to the VDR. The identifiers varied in format across these systems and if similar formats, they could potentially be duplicates. To ensure that the new P_IDs were unique and uniformly formatted, the data management team constructed a pool of universal P_IDs and coordinated with the Study Centers to develop a crosswalk between these and the original P_IDs. Once implemented all subsequent data submissions utilized the universal P_IDs or UPIDs.

IDs were constructed as follows:

Participant Type Values 1-4 || random 6-digit identifier || 2-digit check value

The 2-digit authentication known as check digits (3) allows for easy verification that IDs are entered correctly. In the case of the Vanguard Study UPIDs, the check digit function was as follows:

Check Digits = MOD [98 - MOD (6-digits*100, 97), 97]

ASSIGNING DEMOGRAPHICS AND DUE DATE

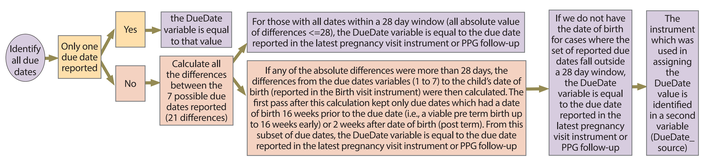
Information used across reports, such as recruitment and retention data, and demographic characteristics of the participants were stored on analytic summary files such as the study visit files or the participant demographic file. This allowed easy retrieval of the information for all reports and yielded consistent demographic counts by all team members and analysts.

Demographics

Some demographics are static, such as race, and others are changing, such as education or marital status. Additionally information is reported multiple times throughout the study, and respondents may not necessarily be consistent on how they respond to even the static demographic questions. For analysis repeatability, the data management team produced a demographic file as a source for all analyses. Algorithms were developed culling information across multiple instruments to identify reported values to assign participants demographic characteristics, using the last reported value as the one to keep. This provided complete and consistent data as possible per participant rather than relying on an individual data element within a given dataset or ad hoc algorithms.

Due Date

Determination of the best due date presented a challenge in that a woman may report multiple due dates for a given pregnancy and then may also have multiple pregnancies, some resulting in live births and others not (e.g., miscarriages). Determining the most accurate due date is crucial for subsequently calculating the gestational age of the child at birth, and the weeks pregnant for bio-specimens and data collection from the mother. Up to seven data elements contained due dates.



Results from the Due Date Algorithm:

- 62% of the cases had one due date value reported, either reported once or multiple times.
- 32% of the cases had multiple due dates reported but all were within 28 days of each other.
- Just over 1% of cases had multiple due dates more than 28 days apart and no Child's DOB was available to verify the dates.
- The remaining 5% had due dates more than 28 days apart and Child's DOB was used to verify the set of feasible dates before assign the best estimated due date.

MAINTAINING CONFIDENTIALITY FOR DATE FIELDS

Full date of birth is crucial in determining a child's age during particular events or postnatal interviews, especially for the very young. Likewise, the full due date is crucial in determining the gestational age for sample collection or at the time of the mother's pregnancy interviews. Since these dates represent personally identifiable information (PII), they cannot be released. Instead, relative ages were provided (i.e., child's age in weeks or months or the gestational age in weeks).

.

ACKNOWLEDGEMENTS

I would like to thank Kathleen McCann, Peter Gilbertson, Susan Forrester, Lydia Rogers, and Lauren Prelewicz at Social & Scientific Systems, Inc., Dr. John Moye, Jr. at the *Eunice Kennedy Shriver* National Institute of Child Health & Human Development, and Dr. Christina Park at the Office of the Director at NIH for their assistance with the review and poster preparation, and all the families who participated in The National Children's Study. I would like to thank system developers Lew Gaul and Michael Laux for their work on developing the Vanguard Work bench.

BIBLIOGRAPHY

1. Park CH, Winglee M, Kwan J, Andrews L, Hudak ML Comparison of Recruitment Strategy Outcomes in the National Children's Study. Pediatrics 2017; 140; DOI: 10.1542/peds.2016-2822 originally published online July 19, 2017.
2. Master Data Element Specifications (MDES) for the Vanguard Data Repository (VDR) V4.0. National Children's Study Alternate Recruitment Study.
3. Foley MJ Check Digits: Avoiding Errors in Clinical Trials, University of North Carolina at Chapel Hill, NC *https://www.lexjansen.com/pharmasug/2000/DMandVis/dm08.pdf*.

For more information, please contact:

NICHD Website
https://www.nichd.nih.gov/research/NCS/

NCS Archive Website
https://ncsarchive.s-3.net/

Archive Email
NCSArchive@s-3.net

THE NATIONAL CHILDREN'S STUDY

HEALTH GROWTH ENVIRONMENT

NIH *Eunice Kennedy Shriver* National Institute of Child Health and Human Development