Paradata Production Post-Data Collection: Challenges and Best Practices

Authors: Tonja Kyle, MS, Nicole Frascino, BS, Maria Prosviryakova, MS, Baibai Chen, MS, Robert Tortora, PhD, and Ronaldo Iachan, PhD

Overview

- Brief history of the NCS study
- Paradata and its uses
- NCS paradata types and variables
- Steps in preparing NCS paradata datasets
- Challenges
- Lessons Learned and Conclusions

NCS Objectives and History

- The National Children's Study (NCS) was initially designed as a 21-year longitudinal study of U.S. children, and their primary caregivers
- The goal of the study was to examine environmental influences on child health and development
- Enroll and follow 100,000 children from birth to age 21
- Vanguard Study : A pilot study was undertaken to determine the feasibility, appropriateness, and cost of different recruitment strategies, study protocols, and procedures
- This Vanguard Study collected information about participants through self-administered or caregiver-reported questionnaires, biological specimens, environmental samples, and physical assessments

Vanguard Study Overview

- Data were collected for the Vanguard Study from 2009 to 2014
- Approximately 5,000 children were enrolled
- Forty study locations across the United States
- Each study location was assigned to one of four regions: East, Central, South and West
- Each region contained 10 study locations and was managed by one of four Regional Operating Centers (ROCs)
- Similar data collection protocols were implemented
- Different operational modes across ROCs: mobile vans, at-home collections and store fronts
- Different information management system across ROCs

Data Collection Components: Instruments and units

- **The NCS Study ended in December 2014**
- **NIH Director recommended that Vanguard Study data be archived and made available for secondary analyses**
- **There are two categories of data to archive:**
  - Survey data collected about children and their parents
  - Operational data collected about the processes used to collect survey data

Data Collection Components: Instruments and units

- **Questionnaire/Instrument data including in-person (CAPI, computer-assisted personal interviewing), phone (CATI, computer-assisted telephone interviewing) and self-administered questionnaires (SAQ)**
  - Mothers, fathers and children
  - 60 interview instruments and 80 additional data collection forms
- **Neuro-psychological and cognitive assessments**
- **Direct assessments (e.g., physical measures such as height and weight, blood pressure, circumferences and skinfold thickness)**
- **Environmental samples, e.g., air, water, and dust from participants homes**
- **Biospecimens, e.g., blood, urine, saliva**
- **Operational data e.g., recruitment strategy, geographic data**

What is paradata?

- **Data collected about the actual data collection method**
- **Provides understanding of the quality of survey data collected**
- **Can assist in gauging measurement error and survey non-response error**
- **Generally defined prior to data collection**
- **Varies based on the mode of data collection: web, mail, phone, in-person**
- **Web paradata**
  - Number of visits to the survey, time spent in each visit, IP address
- **Phone paradata**
  - number of callbacks to complete survey, time of the call, corrections in the data entry
- **In-person paradata**
  - Field interview experience, barriers to access, number of attempts, time of interview

NCS Paradata Production Overview

- **NCS paradata was defined post data collection**
- **For NCS, paradata is focused on operational data**
  - Geographic location, recruitment strategy, number of contacts made, and time to complete
- **Several steps were required to identify and produce operational data for NCS:**
  - Literature review to identify common paradata variables
  - Review of available data from NCS Vanguard Study
  - Definition of variable concepts that can be defined using existing data
  - Identification of data gaps
  - Computation and organization of operational data variables
  - Creating link between operational data and study data

NCS Paradata Production Process

- **Conduct literature review**
  - Identify common operational data
  - Guidance from other agencies that produced similar datasets from complex studies, e.g., National Health Interview Survey
- **Identification of ideal paradata concepts**
  - Study Location
  - Participant ID (Child, Mother, Father)
  - Final Disposition
  - Completion status for each event
  - Completion status for each component of each event
  - Time to complete each component
  - Data collection mode (CAPI, CATI, PAPI)
  - Data collection location (home, van, clinic)
  - Data collector information

NCS Paradata Production Process, continued

- **Define paradata concepts as NCS-specific paradata**

Concept	Example Variables
1. Study/Data Management	Regional Operating Center, Information Management System
2. Case-Level	Participant IDs, Eligibility status, Enrollment status
3. Contact Strategies	Recruitment strategy, Prior knowledge of NCS
4. Measures of Cooperation	Consent status, Last interview conducted, Final status at study close
5. Measures of Contact	Number of contacts associated with each event, Total number of contacts
6. Dispositions	Disposition codes for each event, Presence of biologic specimens
7. Demographics	Age, Race/ethnicity, Primary sampling unit ID
8. Measures of Time	Time in study
9. Data Collection Mode	Mode (paper-and-pencil, computer, phone); Location of data collection (home, clinic, medical van)
10. Data Collector Information	Age, Race/ethnicity, Years of education

NCS Paradata Production Process, continued

- **Select subset of paradata**
  - Omit “Measures of Time,” “Data Collection Mode,” “Data Collector Information”
- **Conduct analyses to gauge variable utility and completeness**
- **Identify and produce variables requiring recoding or computation of new variable**
- **Create variables to link across datasets Mother ID and child IDs**
  - More than one child in a family could have enrolled in NCS
- **Prepare final dataset**
  - 63 paradata variables plus 12 additional “desirable” variables to support non-response and measurement error analyses
- **Prepare final data documentation**
  - Concept, variable number, variable name, variable label, format, variable levels, variable derivation, justification and comments

NCS Paradata Example – Study/Data Management Information

Concept	Variable Number	Variable Name	Variable Label	Format	Variable levels	Derived	Justification
Study-Data Management Info	1	_roc	Regional Operational Centers	rocs.	CROC 19710 EROC 7956 SROC 6731 WROC 16507	Yes, using CURRENT_PSU_ID per guidance in Data User Manual (pages 17-19).	This variable can be used to compare data collection strategies / acquisition and retention rates across four regional centers managed by three contractors.

NCS Paradata Example – Case Level and Contact Strategies

Concept	Variable Number	Variable Name	Variable Label	Format	Variable levels	Derived	Justification
Case Level	2	P_ID	Participant ID	$idfmt.	Valid Non-Missing ID 50904	No	For linkage with analytic and demographic caregiver datasets.

Concept	Variable Number	Variable Name	Variable Label	Format	Variable levels	Derived	Justification
Contact Strategies	13	RECRUITTYPE	Recruitment Strategy for PSU	$rec.	EH - Enhanced Household 27840 HL - Direct Outreach (High-Low Intensity) 19347 PB - Provider-Based 3717	No	To measure effectiveness of recruitment strategy for retention.

Challenges

- **Challenges occurred for several reasons:**
  - **Changes in Vanguard Study over time**
    - Varying IMS systems and changing recruitment strategies
    - Multiple contractors
    - Management reorganization and information transition
  - **Inconsistent data structures and data management**
    - Varying data collection modes (in-home, mobile vans, store front clinics)
    - Varying data structures over time
    - Lack of standardized capture of critical variables
  - **Variations of data quality, e.g., degree of missing data or paper vs. electronic capture**
  - **Availability of data**
    - Only data from the Alternate Recruitment Study (ARS) was used (2010-2014)
    - Initial Vanguard Study (IVS) data was not included (2009-2010)

Final Overview and Lessons Learned

- **NCS paradata production required:**
  - Examination and curation of hundreds of datasets
  - Review of hundreds of pages of documentation
  - Development and implementation of harmonization strategies
  - Development of analytical and computation procedures to create variables
  - Development of documentation for the final datasets
- **Created a paradata dataset containing 63 variables that can be used to inform future research and design questions related to longitudinal, large-scale studies**
- **Post data collection production of paradata may limit the utility of the dataset to support future research and examine bias and error issues**
- **Standardization of paradata components is key**

Questions and more information?

For more detailed information on the National Children's Study visit:

https://www.nichd.nih.gov/research/NCS/Documents/NCS_Archive_Study_Description.pdf

Questions?

Please contact:
Tonja Kyle
ICF International, Inc.
Tonja.Kyle@icfi.com
301-572-0820