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
Define paradata concepts as NCS-specific paradata

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

<table>
<thead>
<tr>
<th>Concept</th>
<th>Variable Number</th>
<th>Variable Name</th>
<th>Variable Label</th>
<th>Format</th>
<th>Variable levels</th>
<th>Derived</th>
<th>Justification</th>
</tr>
</thead>
<tbody>
<tr>
<td>Study-Data Management Info</td>
<td>1</td>
<td>_roc</td>
<td>Regional Operational Centers</td>
<td>rocs.</td>
<td>CROC 19710 EROC 7956 SROC 6731 WROC 16507</td>
<td>Yes, using CURRENT_PSU_ID per guidance in Data User Manual (pages 17-19).</td>
<td>This variable can be used to compare data collection strategies / acquisition and retention rates across four regional centers managed by three contractors.</td>
</tr>
</tbody>
</table>
## NCS Paradata Example – Case Level and Contact Strategies

<table>
<thead>
<tr>
<th>Concept</th>
<th>Variable Number</th>
<th>Variable Name</th>
<th>Variable Label</th>
<th>Format</th>
<th>Variable levels</th>
<th>Derived</th>
<th>Justification</th>
</tr>
</thead>
<tbody>
<tr>
<td>Case Level</td>
<td>2</td>
<td>P_ID</td>
<td>Participant ID</td>
<td>$idfmt.</td>
<td>Valid Non-Missing ID 50904</td>
<td>No</td>
<td>For linkage with analytic and demographic caregiver datasets.</td>
</tr>
</tbody>
</table>

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