National Center for Health Statistics Rapid Surveys System:

Navigating the data processing journey from online probability panel data collection to public-use file

Adam Lee[1] & Emily Terlizzi[2]

[1] RTI International

[2] National Center for Health Statistics - CDC

10/23/2024

Outline

- Background
- Challenges
- Planned Approach
  - Schedule and Team Structure
  - Technical Considerations
- Results
- Lessons Learned

Background: Rapid Surveys System (RSS)

- **Objectives**
  - **Time-sensitive data of known quality about emerging and priority health concerns.**
  - Evaluate the quality of public health estimates generated from commercial online panels.
  - Improve methods to appropriately communicate the fitness for use of public health estimates generated from commercial online panels.
- **Rounds**
  - 2023: RSS-1; RSS-2
  - 2024: RSS-3; RSS-4; RSS-5
  - In future years, the number of rounds will vary depending on public health needs.

Background: Rapid Surveys System (RSS)

- **Post Data Collection Tasks:**
  - Data Quality Review
  - Data Processing
    - Data cleaning
    - Harmonization
    - Creation of new variables
    - Restricted-Use File (RUF) development
  - Disclosure Review/Public-Use File Development
  - Reporting
    - Producing statistical tables and dashboards

Challenges in Producing Timely Data

- Two panel providers performing data collection:
  - Different data collection platforms (Voxco vs. SPSS Dimensions)
  - Panel providers program their surveys separately (from the same survey instrument)
  - Different modes (Web vs Web/CATI)
  - Different profile variables collected of respondents
  - Two different panel data processing teams

Challenges in Producing Timely Data (Cont.)

- Survey complexity:
  - Topics can change every round
  - Variables: ~600
    - 250 questionnaire variables
    - 150 paradata variables
    - 200 panel profile variables
  - Sample of 15,000 records (~8,000 respondents)
  - Possible complex skip logic and subgroup targeted questions

Challenges in Producing Timely Data (Cont.)

- Schedule drivers:
  - Contractual requirement for RTI, 45-day after panel providers deliver their data, RTI is required to produce a restricted-use files (RUF) and public-use files (PUF).
  - NCHS Disclosure Review Board (DRB) needs to review and approve of the PUF before public release.
    - DRB reviews need to be scheduled months in advance
    - Requires 21 days review period
    - Estimates would be released at the same day as the PUF dataset

How Do We Address These Challenges?

Our approach:

- Optimizing our schedule
- Building parallel teams to support development
- Addressing the technical complexity through planning

Planned Approach

Schedule and Team Structure

Example Schedule (if completed in series)

- Completing all the tasks linearly would take **8 months**!

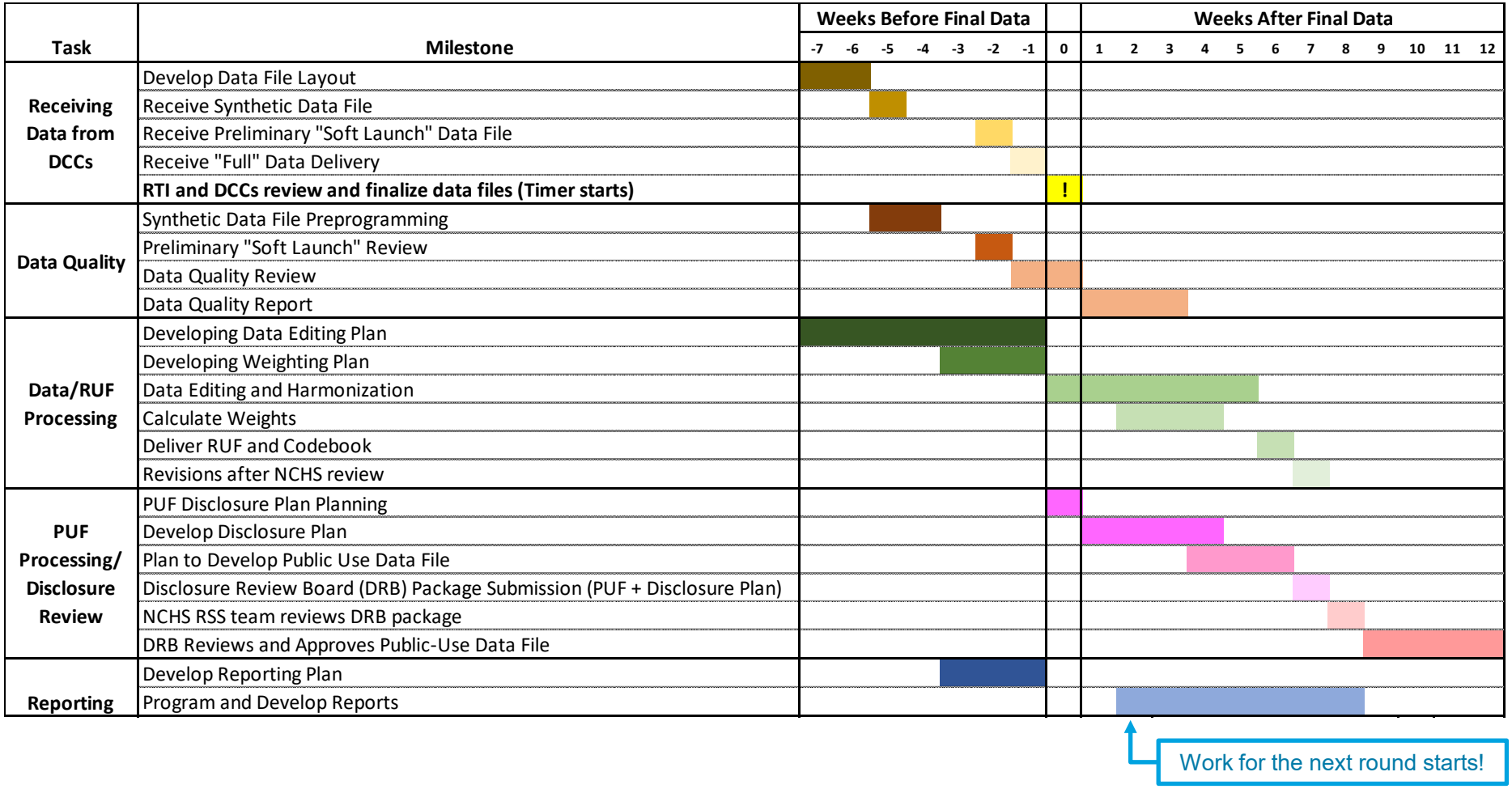
Task	Milestone	Weeks After Final Data																												
		0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29
	RTI and DCCs review and finalize data files (Timer starts)	!																												
Data Quality Review	Data Quality Review																													
	Data Quality Report																													
Data/RUF Processing	Developing Data Editing Plan																													
	Developing Weighting Plan																													
	Data Editing and Harmonization																													
	Calculate Weights																													
	Deliver RUF and Codebook																													
	Revisions after NCHS review																													
PUF Processing/ Disclosure Review	PUF Disclosure Plan Planning																													
	Develop Disclosure Plan																													
	Plan to Develop Public Use Data File																													
	Disclosure Review Board (DRB) Package Submission (PUF + Disclosure Plan)																													
	NCHS RSS team reviews DRB package																													
	DRB Reviews and Approves Public-Use Data File																													
Reporting	Develop Reporting Plan																													
	Program and Develop Reports																													

Building the Actual Schedule

- From end of data collection to the final PUF we generally try to target 3 months for all of our post processing tasks.
- To achieve this goal, our project schedule focuses on:
  - Significant planning in advance of the end of data collection
  - Doing tasks traditionally done in series are done in parallel (accepting the data may change)
  - Staffing multiple teams to ensure we have resources to run tasks concurrently

Actual Schedule

- From end of data collection to the final PUF we try to target 3 months for all of our post processing tasks.



Organizational Structure

- Parallelization means multiple teams are required. Our core teams are:
  - Data Quality
  - Data processing/RUF development
  - Weighting
  - Disclosure Review/PUF Development
  - Reporting
- Communication across teams is essential, with focus on updating teams of the status of the data:
  1. Receive raw data from panel providers (Data Quality and Data Processing start)
  2. Raw datasets merged into one working dataset (Reporting and Weighting start)
  3. Harmonization of data is complete and no variable changes (Disclosure Review starts)

Planned Approach

Technical Consideration

Preparing for Round 1

- Setting expectations for panel providers:
  - Data Standardization Guidelines
    - Defined completes, missing value ranges, eligible/ineligible cases
    - Outlined requested files and supporting documentation (SAS7BDAT format, with formats)
  - Data File Layout
    - Lists all expected variables in the dataset we receive from the panel providers.
    - Noting any deviations from the questionnaire programming

Preparing for Round 1

- Preparing for post-processing:
  - Post-Processing Data Editing Plan
    - Handling inconsistences and data editing
    - Standardizing labeling of variables and descriptions
    - Outlining new variables that needed to be created
  - Early Examples of the Data
    - “Synthetic” data – RNG values produced based on the programmed survey
    - “Soft Launch” data – Real data soon after data collection started.

Example: Data File Layout

- Layout is a table that outlines each variable we expect to receive from the panel providers
- Development of the layout requires working from the questionnaire before we see any real data

NCHS Rapid Survey - Round # File Layout									
TYPE	**DCC**	**Variable Name**	**Variable Label**	**Variable Type**	**Format Values**	**Notes**	**Universe**	**Comments from Panel #1**	**Comments from Panel #2**
QUEX	BOTH	HIS_GENERAL	*Self-reported health status*	numeric	1 = 'Excellent' 2 = 'Very good' 3 = 'Good' 4 = 'Fair' 5 = 'Poor';		ALL		
…	…	…	…	…	…	…	…	…	…

Data Quality Review

- Three main reviews that inform data harmonization:
  - SAS Formats Comparisons Review
    - Align each variable format from the two panel providers and our file layout and identify inconsistences in coding.
  - Frequency comparisons between the two panel providers
    - Both datasets should be weighted to the same population and should produce similar estimates.
    - Need to consider the methodological differences and where outcomes may vary.
  - Logic Skip Check Review
    - Based on the programming specifications, we evaluate if the data matched the possible range of responses.
- Additionally, we review paradata (e.g., drop offs by questions, section timing) and key estimates by each panel provider and against known benchmarks.

Results

Results After Round 1

- Even with planning, it still was tough!
- Timelines were still very tight.
- A lot of the original code needed to be developed which affected how quickly we could complete each task.
- We still had discrepancies from the panel providers despite our planning.

Results - Planning Stage

- Version control across contractors and NCHS (4 different organizations)
  - Originally, we traded comments and file versions via email, where RTI consolidated and resolved feedback.
  - This was time consuming and led to multiple versions and required time to resolve and incorporate comments and edits.
  - Specifically, the file layout went through many early iterations in Round 1.
- Due to tight timelines for the panel providers, early snapshots of the data ("*Synthetic*" data and "*Soft Launch*" data) did not reflect the final data to review for quality or build programs from.

Results - Data Processing Stage

- Teams were then stretched by the amount of work required in the first round. Many decisions and processes needed to be developed as we were implementing them.
- Tasks still need discussion with NCHS and panel providers to complete even under tight timelines.

Results - Data Processing Stage (Cont.)

- Data file layout does not reflect the questionnaire and data collection complexity perfectly.
  - Different interpretation of the questionnaire and instructions across RTI and the panel providers.
  - Resulted in many small differences, especially around missing values (e.g., *System missing, Don't know, Refused, Skipped, etc.*)

Lessons Learned

Lessons Learned

- Version control across organizations:
  - We started using collaboration software (e.g., “Share” link on MS Excel, Smartsheet) to allow multiple organizations view/edit a single document.
- Early data of limited use:
  - Ask panel providers to provide us with a “near final” preliminary dataset later in the fielding period before the “final” dataset is received.
  - Less time to work with the early data but enhances utility for reviewing and programming

Lessons Learned (Cont.)

- Data file layout inconsistencies:
  - Clarified expectations for variables that are the same round to round
  - Engage in early conversations when questionnaire utilizes complex logic that may need more clarification
- Ongoing discussions with NCHS
  - We push the discussion earlier in the timeline to allow for as much time as possible
  - Establish procedures in advance on what to do to minimize delays in the work

What is Next?

- We are wrapping up our 5th round of this study with potentially more rounds in the future.
- We continue to refine our approach to support the rapid release of data.
  - Review the schedule to find more efficiencies
  - Expand data documentation for panel providers to minimize inconsistences from round to round
  - Automate data processing code for our post processing tasks increase speed in data processing.

Thank you

Contact: Adam Lee | email: atlee@rti.org

Note: The findings and conclusions in this report are those of the authors and do not necessarily represent the official position of the Centers for Disease Control and Prevention