Data Management in a Community-Based Birth Cohort: What the SEMILLA Study Teaches Us

PRACTICE PAPER

NATALY CADENA
FADYA OROZCO
STEPHANIE MONTENEGRO
FABIÁN MUÑOZ
ALEXIS J. HANDAL

*Author affiliations can be found in the back matter of this article

ABSTRACT

In cohort studies, systematic information management often receives limited attention in study protocols, resulting in delays, quality issues, and threats to data validity. This paper describes the data management process of a community-based cohort study, using the SEMILLA (Study of Environmental Exposure of Mothers and Infants Impacted by Large-Scale Agriculture) study conducted in Cayambe, Ecuador, as a case example, and highlights the challenges, adaptations, and lessons learned, with the aim of informing similar studies.

The SEMILLA data management process was structured around three key responsibility areas: strategic management, technical coordination, and data administration. The process unfolded in two main stages: Preparatory, which involved iterative refinement of data collection instruments, definition of coding rules, platform adjustments and migration, continuous team training, and implementation of security and anonymization protocols; and Organization, which included the assignment of interviewers for field data entry, primary data cleaning, creation of additional variables to better describe the sample composition and operational conditions, and the production of technical documentation.

This approach contributed to improving data entry consistency, reducing recurrent errors, and strengthening record traceability throughout the follow-up by means of operational monitoring procedures. Key lessons include the importance of establishing a data management protocol and involving a data manager from the study design phase, maintaining flexibility in selecting data collection platforms, ensuring proper assignment of interviewers for each instrument, automating quality control processes, and continuously generating technical and operational documentation. Collectively, these practices help preserve data quality and promote operational efficiency in longitudinal studies conducted in similar contexts.

CORRESPONDING AUTHORS:

Nataly Cadena

Centro de Transferencia de Tecnología CTT-USFQ, Cumbayá, Quito, Pichincha, 170901, Ecuador

natalycadena92@hotmail.es

Alexis J. Handal

Department of Epidemiology, University of Michigan School of Public Health, Ann Arbor, Michigan, United States

ahandal@umich.edu

KEYWORDS:

data management process; lessons learned; cohort studies

TO CITE THIS ARTICLE:

Cadena, N., Orozco, F., Montenegro, S., Muñoz, F. and Handal, A.J. 2026 Data Management in a Community-Based Birth Cohort: What the SEMILLA Study Teaches Us. *Data Science Journal*, 25: 4, pp. 1–12. DOI: https://doi.org/10.5334/dsj-2026-004

INTRODUCTION

Longitudinal and community cohort studies in maternal and child health enable the evaluation of well-being trajectories from pregnancy through early childhood, taking into account both risk factors and social determinants that shape outcomes in resource-limited settings (Feng *et al*., 2024). In practice, the operational and logistical planning of such studies often prioritize data collection and adherence to sampling design, while systematic data management receives limited attention (Young, Powers and Wheway, 2007). This oversight frequently leads to processing delays, information loss, and compromised internal validity, which in turn hinders timely decision-making for the project (Graves, Ball and Fraser, 2007; Adamson and Graves, 2007; Tavakoli *et al*., 2006).

In longitudinal studies, data management is a critical strategic component that must be addressed from the earliest stages of planning. It encompasses not only data collection, but also the organization, storage, and preservation of data to ensure its integrity and long-term utility. This is especially relevant for studies requiring extended follow-up, such as antenatal cohort studies, where data are collected across multiple stages (pregnancy, birth, and childhood) and must remain consistent and traceable throughout the process (Graves, Ball and Fraser, 2007; Young, Powers and Wheway, 2007; Dhudasia, Grundmeier and Mukhopadhyay, 2023; Adamson and Graves, 2007; Tavakoli *et al*., 2006).

Tavakoli *et al*. (2006) proposed a three-stage data management structure: instrument design, programming of data collection systems, and definitive archiving of information. Developed within the framework of the Rural Women's Health Project, this model has proven particularly effective in longitudinal contexts by enabling proactive planning of information flow. Building on this foundation, Dhudasia, Grundmeier and Mukhopadhyay (2023) advocate for an approach that begins with the creation of standardized forms and thorough documentation of variable contexts, complemented by real-time monitoring mechanisms during data collection (Dhudasia, Grundmeier and Mukhopadhyay, 2023; Tavakoli *et al*., 2006).

Both approaches emphasize the need to establish automated quality controls, document every modification, assign clear responsibilities, and ensure regular data backups and key actions to safeguard the validity of the process. Additionally, other authors highlight the importance of ensuring that data management in longitudinal studies is flexible, secure, and thoroughly documented. For example, Adamson and Graves (2007) stress the importance of a management system that reflects the dynamic status of participants, thereby facilitating monitoring and maintaining data integrity. Young, Powers and Wheway (2007) underscore the necessity of well-defined data management protocols, while Graves, Ball and Fraser (2007) emphasize the importance of primary data cleaning to optimize analysis and reduce systematic errors (Graves, Ball and Fraser, 2007; Young, Powers and Wheway, 2007; Adamson and Graves, 2007).

Integrating these activities enhances the consistency and utility of the data, minimizes reporting errors, and improves internal validity in cohort studies. In this context, the objective of this essay is to describe the data management process of a community-based cohort study, using the SEMILLA (Study of Environmental Exposure of Mothers and Infants Impacted by Large-Scale Agriculture) study conducted in Cayambe, Ecuador, as a case example, and to highlight the challenges, adaptations, and lessons learned, with the aim of informing similar studies.

DATA MANAGEMENT IN SEMILLA

In the SEMILLA community-based cohort study (Handal *et al*., 2024), an organizational structure with three areas of responsibility was defined: 1) Strategic Direction, represented by the principal investigators (FO and AJH), who were responsible for overall decision-making; 2) Technical Coordination, responsible for managing field data collection (SM); and 3) Data Administration, in charge of organizing, processing, and storing the information (NC and FM). These areas worked in an integrated and dynamic manner, establishing a continuous feedback loop that enhanced both operational efficiency and data reliability (Figure 1).

DOI: 10.5334/dsj-2026-004

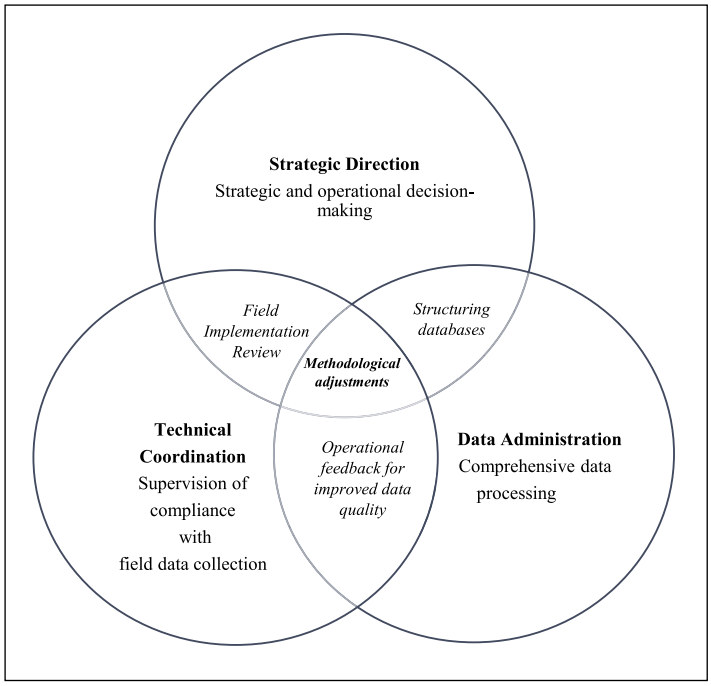

Figure 1 Organic structure of data management in the SEMILLA research study (2018–2024).

To enhance the clarity and usability of the manuscript, we provide a schematic representation of the SEMILLA data management workflow (Figure 2). The figure summarizes the sequential stages implemented during the study, from the preparatory phase to final data archiving, offering a concise overview of the process.

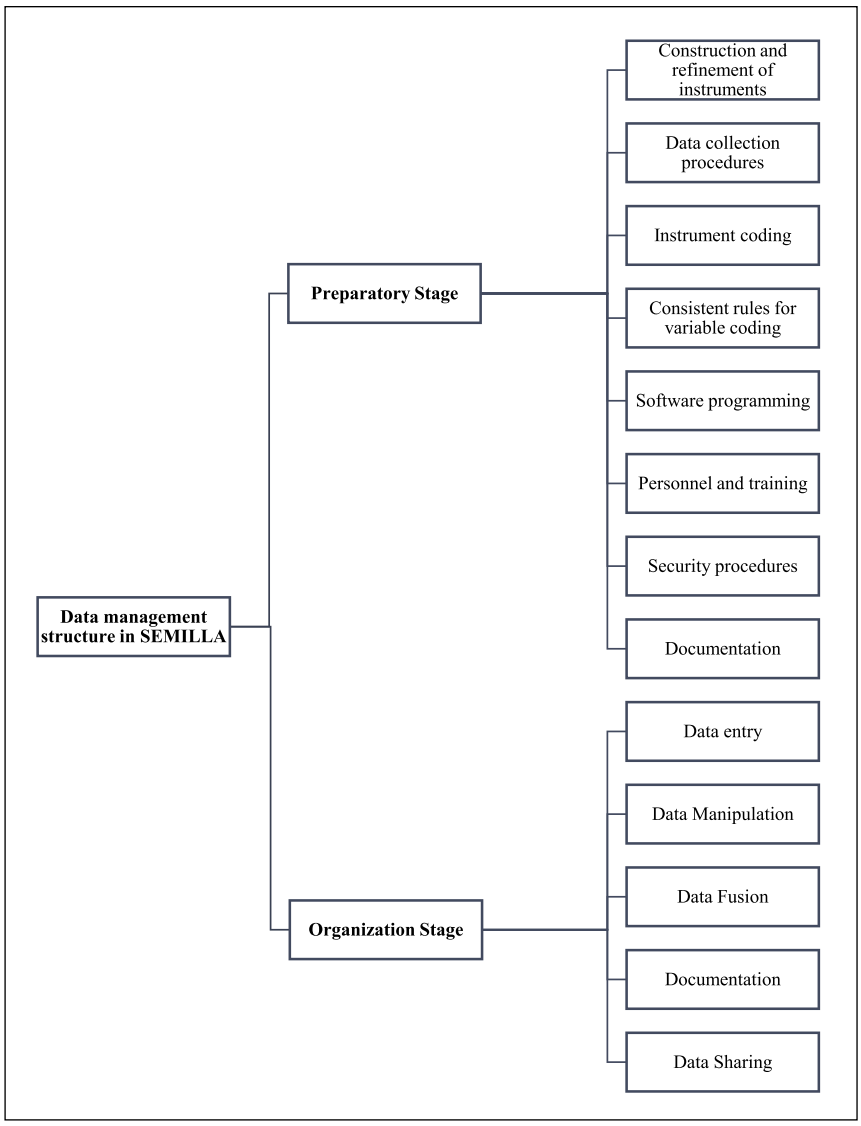

Figure 2 Workflow of data management in SEMILLA research study (2018–2024).

Following Tavakoli *et al*. (2006) as a reference framework, the SEMILLA data management process was organized into sequential stages adapted to our operational context. In the following sections, each stage of the workflow is described in detail, including specific challenges and adaptations.

1. PREPARATORY STAGE

Construction and refinement of instruments:

The instruments used in SEMILLA consisted of a main questionnaire tailored to each data collection point (baseline, pregnancy, delivery/post-partum, mother/baby follow-up), incorporating standardized scales (e.g., job stress, pregnancy stress), as well as standardized psychometric scales (e.g., depression) and laboratory data (Handal *et al*., 2024). The main questionnaires were based on a 2008 pilot study (Handal *et al*., 2015, 2016) and later adapted and expanded upon for the SEMILLA study, which began in 2018. Other instruments, such as the psychometric scales, were adapted from available standardized scales. Tavakoli *et al*. (2006) highlight the importance of well-structured instruments, clear interviewer instructions, and systematic documentation of missing responses. While developed independently, SEMILLA's instruments can be compared to these principles.

This phase presented several main challenges. The first major challenge was linguistic and content adaptation and refinement. This process required clarifying instructions and rewording complex items to ensure both clarity and cultural appropriateness. These adjustments were especially important for the main questionnaire for several reasons. First, translation into Spanish and adaptation to the local context was challenging due to the complexity of the main questionnaire, which for the baseline time point contained approximately 1,000 items. Second, the questionnaires often included multiple standardized scales originally developed in English and for the US or European populations. For example, during the baseline and pregnancy data collection time points, standardized scales assessing job stress and pregnancy stress were included in the main questionnaire. For the delivery/post-partum questionnaire, a standardized scale assessing obstetric violence was adapted and included. As such, not only did the questionnaires require translation and adaptation of content from the earlier pilot study, but the standardized scales embedded in the questionnaires also needed to be translated and culturally adapted to ensure their appropriateness for the local population. Applying Tavakoli's principles showed that adjustments had to be conceptual as well as linguistic. Using the baseline questionnaire as an example, an item originally phrased 'During the last 3 months, in the last week, did you enter a recently fumigated greenhouse?' combined overlapping time references that participants could not interpret. During piloting, many women asked whether they should report the last three months or just the last week, generating inconsistent responses. The question was revised to refer only to the past week, consistent with the quarterly schedule. Marital status was also reframed: many women declared themselves 'separated' during conflicts but later reported being 'married' again. To reflect this instability, 'conjugal status' was adopted, and culturally familiar terms like 'husband/partner/companion' were used. Household composition was clarified by instructing interviewers to define members as 'those who share the same pot,' a locally recognized expression. Employment constructs required extensive revision, as many women initially reported 'not working' when helping in family businesses or selling products occasionally. To avoid underestimation, interviewers probed with 'any activity that generates income for you or your household,' clarifying that even occasional activities counted as work. Interviewers noted standardized overtime (>8 hours/day) and distinguished formal floriculture jobs from autonomous activities. These culturally tailored adaptations improved consistency, reduced recurrent errors, and aligned with debates on the decolonization of research.

The second challenge was the development of complex skip patterns within the main questionnaire used at the various data collection time points. The SEMILLA study design included three distinct participant groups: workers in the floriculture/agricultural sector, workers in other sectors, and women not engaged in paid work (Handal *et al*., 2024). This design required specific sections and questions to be administered only to the relevant subgroup. For example, questions related to occupational exposures or workplace stressors were relevant only for participants currently employed outside of the home for pay, and in some cases, only for those

in the agricultural sector or even more specifically, only for those in floricultural work. Moreover, skip patterns were designed to separate self-employment/informal work from formal wage employment. Self-employed women were asked about type of activity, income generation, and business characteristics, including reliance on unpaid family labor. By contrast, women in formal jobs received modules on contracts, social security, supervision, maternity leave, lactation rooms, permissions for medical appointments, and overtime pay. These differentiated patterns prevented non-comparable responses and allowed SEMILLA to capture the diversity and precariousness of women's labor experiences in the study area. Ensuring that the skip patterns were accurately followed necessitated a meticulous review of the questionnaire's internal logic to maintain the coherence and relevance of each thematic block.

Finally, the onset of the COVID-19 pandemic forced further adjustments and refinements of study instruments. Shifts in employment conditions affected eligibility criteria, requiring changes in questionnaire content, skip logic, and administration modes. Instruments were adapted for hybrid or phone-based collection, ensuring continuity despite restrictions.

Data collection procedures: The instruments were administered both in person and by telephone, due to contextual constraints imposed by the COVID-19 pandemic (Handal *et al.*, 2024). Several strategies were implemented to ensure the quality of the data. First, consistency was prioritized in both the formulation of the questions and the recording of responses, following the scripting guidelines of Tavakoli *et al.* (2006). Instructions for interviewers were highlighted in bold, while prompts to be read aloud to participants were presented in italics, ensuring the interview scripts were clear and easy to follow. For open-ended questions, interviewers transcribed participants' responses verbatim into the designated response fields.

This approach aligns with the principles described by Tavakoli *et al.* (2006), who emphasized that consistency in question wording and recording responses is critical to minimizing interviewer bias. However, SEMILLA had to go beyond these recommendations by adapting procedures to unexpected conditions. For example, questionnaires were piloted with the field team using paper versions to support training on skip patterns, which were more difficult to visualize on the digital platform. This strategy allowed the team to identify and correct navigation issues in real time before fully implementing the electronic system. In addition, telephone administration required adaptations not discussed in Tavakoli's framework, such as ensuring that rapport could be established without face-to-face interaction and reinforcing instructions to avoid misinterpretation over the phone.

Instrument coding: To facilitate the organization and traceability of the data, particularly given the longitudinal design and repeated application of the same instruments over time, a standardized coding system was implemented. Each variable was assigned an alphabetical prefix indicating both the time point and the instrument used. For example, the prefix 'bl' denoted the baseline questionnaire administered at the start of the study. This coding strategy helped standardize file names, variable labels, and other key elements within the data management system.

Tavakoli *et al.* (2006) similarly highlight the importance of consistent naming conventions and coding rules, noting that variable names must reflect different time points and that standardized codes should be defined for missing responses. SEMILLA's approach shared this principle but expanded it further. In addition to the coding prefixes, a comprehensive flow chart was developed for each time period when the questionnaire was administered (e.g., baseline, pregnancy follow-up, etc.), allowing for a clear visualization of skip patterns and the variables associated with each stage of data collection. This tool was essential for maintaining the logical structure of the instruments, optimizing data coding, minimizing errors, and facilitating training. It contributed significantly to preserving methodological consistency throughout the follow-up period and reinforcing the quality of the coding process.

Consistent rules for variable coding: To ensure uniform coding, we adopted standards widely recommended in the literature (Dhudasia, Grundmeier and Mukhopadhyay, 2023; Tavakoli *et al.*, 2006). For instance, we defined specific codes to differentiate types of non-response: 'don't know' = 7, 'not applicable' = 8, and 'refused to answer' = 9. For variables using two-digit formats, these were converted to 97, 98, and 99, respectively. For longer numeric fields, we used three-digit codes (997, 998, and 999) to avoid confusion with actual values, such as distinguishing 99 from a valid response like 99 hours of overtime worked.

Software programming: Initial data collection was conducted using the Open Data Kit (ODK) platform, as outlined in the study protocol (Handal *et al*., 2024). However, due to technical challenges, including the complexity of skip patterns, the length of the questionnaires, and limitations in the data export formats, we decided to migrate to the Census and Survey Processing System (CSPro) (Census Bureau United States, n.d).

While Tavakoli *et al*. (2006) focus on the importance of programming steps and labeling for accurate data handling and reproducibility, SEMILLA faced a different but related challenge: ensuring that the data capture software itself could support the complexity of the study instruments. The migration from ODK to CSPro was necessary to implement advanced skip logic and maintain consistency across waves.

During the transition period to the new digital platform, data collection continued using Word documents while all instruments were redesigned in CSPro. The design, testing, and progressive consolidation of the new data structures in CSPro took nearly a year. During this period, the questionnaires completed in Word format were systematically digitized into the new platform. This process illustrates the methodological flexibility required in community-based cohort studies: whereas Tavakoli emphasizes programming transparency, SEMILLA demonstrates the importance of platform adaptability to sustain data collection under demanding field conditions.

Personnel and training: A rigorous training process was developed for interviewers to ensure a thorough understanding of each question, with particular emphasis on sensitive topics, to ensure high quality data collection through confidentiality, trust, and respect. Tavakoli *et al*. (2006) emphasize manuals and close supervision to establish consistency among multiple data collectors; SEMILLA followed these principles but also required additional, system-specific training during the migration from ODK to CSPro, complemented by ongoing supervision to prevent confusion and ensure accurate and complete data collection.

Security procedures: Details regarding the Institutional Review Board Statement and the informed consent process are described elsewhere (Handal *et al*., 2024) and summarized here. The SEMILLA study received ethical approval from institutional review boards in the United States (University of New Mexico and University of Michigan), Ecuador (Universidad San Francisco de Quito), and the Ecuadorian Ministry of Public Health, with additional authorization from the District Health Directorate 17D10 (Cayambe-Pedro Moncayo). Written informed consent was obtained from each woman, and, after delivery, also for her infant.

Tavakoli *et al*. (2006) emphasize physical safeguards such as locked storage, controlled access to completed instruments, and regular backups. SEMILLA addressed security in a digital environment, combining ethical oversight with data protection measures adapted to electronic platforms. Personal information was under the exclusive responsibility of the Technical Coordinator (SM), who safeguarded the nominal lists required for follow-up logistics. For data analysis and database management, each woman was anonymized with a unique four-digit identifier (idmadre), jointly assigned by the Technical Coordinator (SM) and the data administration team (NC and FM). This identifier was used consistently across all instruments, allowing traceability of records throughout the study waves without exposing participants' identities. Access to identifying information was restricted to the Technical Coordinator, while anonymized data files (in ODK, Word, and CSPro formats) were managed solely by the data administration team (NC and FM). Furthermore, the data collection software was installed exclusively on field office equipment, and downloads to personal devices were not permitted.

Documentation: To monitor activity compliance at each data collection time period and record participant-reported reasons for withdrawal, we created an Excel spreadsheet called the Tracking Planner (Montenegro, Handal and Orozco, 2025), which the field team updated daily. This allowed for a continuous, detailed record of cohort progression, enabling timely identification of dropouts and ensuring traceability for each participant. Tavakoli *et al*. (2006) emphasize that the heart of effective data management lies in documentation, including logs of participant progression, coding decisions, missing data, and dropouts. The Tracking Planner represented SEMILLA's practical adaptation of this principle: it served as both a monitoring and documentation tool, bridging field operations with data management. Unlike the general tracking logs described by Tavakoli, the Tracking Planner was integrated into daily field routines, creating a systematic record that could be directly linked to data quality and follow-up logistics.

DOI: 10.5334/dsj-2026-004

2. DATA ORGANIZATION STAGE

Data entry: Tavakoli *et al*. (2006) note that consistency in data entry is often best achieved by a single individual, as the probability of transcription errors increases with multiple operators. However, they also acknowledge that centralization may introduce systematic bias and therefore recommend complementary quality control procedures such as double entry or random checks. Based on this framework, SEMILLA's experience provides a useful comparison for balancing efficiency and reliability in a large, community-based cohort study. In SEMILLA, three interviewers were assigned to enter data from the main questionnaires for the same group of participants, four additional interviewers were responsible for psychometric scales, and one interviewer managed the laboratory data.

A central tool in ensuring quality was the Tracking Planner previously described, which was used in SEMILLA to cross-check weekly data entry with fieldwork compliance. By linking each participant's unique identifier to completed or pending questionnaires, the Tracking Planner facilitated the identification of missing instruments and required interviewers to provide justifications when a questionnaire was not administered (e.g., participant refusal, scheduling conflict, interviewer oversight). In this way, it functioned as an accountability mechanism that directly supported data verification and minimized unexplained gaps across study waves.

In addition, the data administration team (NC and FM) conducted weekly database reviews, a process described in greater detail in the following section. These reviews were complemented by the configuration of CSPro with internal validation rules and skip pattern checks to reduce entry errors at the point of digitization. Together, these procedures created an iterative verification and correction cycle that substantially strengthened the reliability and consistency of the data entry process across SEMILLA.

Data Manipulation: As Tavakoli *et al*. (2006) note, data manipulation is an essential step before analysis, involving recoding, creating new variables, and developing consistent categories. In SEMILLA, this stage began with primary data cleaning, followed by a series of actions tailored to the complexities of a longitudinal, community-based cohort. The following procedures were implemented:

a) At the end of each monitoring wave, the Technical Coordinator (SM) informed the Strategic Direction (FO and AJH) and Data Administration team (NC and FM) that the collected data on the digital platforms were ready for review, accompanied by a report of any field-related updates. The cleaning process began by exporting the CSPro databases into Stata statistical software (StataCorp, 18). Variable formats were then standardized and quality checks were performed.

Standardization involved harmonizing open-text responses to reduce semantic variability introduced by different interviewers. Across monitoring waves, each principal questionnaire database included approximately 100 open-text variables, which substantially increased heterogeneity in wording and classification. For example, in the section on medication use, participants frequently explained the indication for prenatal supplements using different expressions such as 'for my baby to grow,' 'for the baby to develop,' or 'for the baby to be healthy.' Although all referred to the same purpose, the variability complicated aggregation and interpretation. To ensure consistency, these descriptions were standardized into a unified phrase: 'to support fetal growth, development, and health.'

A similar procedure was applied to other qualitative fields. In the section on medication use, participants frequently listed drug names with spelling variations, inconsistent capitalization, or commercial denominations. These entries were standardized to ensure correct spelling, internal coherence, and compatibility with reported symptoms or diagnoses. Likewise, in the question on beverage consumption during pregnancy, several commercial brand names were reported and were subsequently harmonized into broader analytical categories such as 'carbonated drinks' or 'sugar-sweetened beverages,' improving consistency across records. Additionally, in open-ended questions exploring maternal concerns about their baby, respondents expressed similar ideas using different phrases such as 'that the baby falls and gets hurt,' 'that something bad happens,' 'that the baby is injured,' or 'that the baby suffers an accident.' Although semantically equivalent, this variability complicated aggregation. These descriptions were therefore standardized under a unified thematic label reflecting infant physical safety concerns, preserving the original meaning while enabling clearer longitudinal comparison.

Additionally, orthographic, lexical, and semantic inconsistencies were corrected across all open-text variables to maintain internal coherence within the dataset, ensuring consistency across all qualitative fields. Quality control procedures ensured that standardization preserved the original meaning of participant responses, prevented the artificial creation or duplication of categories, and remained fully aligned with the pre-established coding framework for qualitative variables. This systematic process enhanced semantic clarity across the longitudinal dataset, improved interpretability, and facilitated reliable aggregation and analysis, while respecting the substantive content and intent of the information provided by participants.

Once textual standardization and semantic harmonization were completed, a critical verification step was the review of the idmadre identifier, which uniquely linked each participant across study waves. Although the identifier itself did not change, typographical errors by field interviewers occasionally led to inconsistencies. In some cases, the idmadre was mistyped (e.g., a single digit inverted or omitted), resulting in codes that did not correspond to any participant. In other cases, confusion between similar identifiers (e.g., 1034 vs. 1043) led to apparent duplicates. To resolve these discrepancies, we cross-referenced the interview dates and times automatically recorded in CSPro with the entries in the Tracking Planner, which documented the interviewer responsible for each case. By combining these sources, the correct idmadre could be identified. When inconsistencies persisted, the issue was discussed in dedicated meetings with the field team to confirm the valid record. This process ensured that all identifiers remained accurate and traceable throughout the follow-up period.

After resolving identifier inconsistencies, the next verification step focused on detecting potential extreme values. Apparent outliers primarily arose in the laboratory datasets, such as thyroid hormone or urinary iodine results provided by an accredited external laboratory in Ecuador. Because these values had to be manually entered into CSPro from external laboratory reports, typographical errors occasionally produced implausible results outside physiologically expected ranges. Each potential outlier was re-verified against the original laboratory report. When the value matched the official report, it was retained even if extreme; when a typographical error was identified, it was corrected. In this way, corrections were applied only when clear evidence of entry error existed, while true extreme values remained unaltered.

To ensure a transparent and verifiable distinction between transcription errors and genuine extreme values, we applied several criteria: physiological plausibility was assessed by comparing suspected outliers against the laboratory's reference ranges, so values outside the reference range but within physiologically plausible limits were treated as potentially valid. Every flagged value was checked directly against the official laboratory printout, and values that matched the report were retained, regardless of extremity. Implausible numbers typically resulted from common data-entry issues such as misplaced decimals, extra digits, or incorrect units, and if the original report confirmed a different value, the CSPro entry was corrected. When applicable, internal consistency checks were performed by comparing repeated measurements or related biomarkers to assess coherence. Using this systematic approach, only values with clear, documentable evidence of transcription error were corrected, whereas physiologically plausible extreme values remained unchanged.

This multi-step review process led by NC and supported by SM combined CSPro time stamps, the Tracking Planner, and structured cross-checks to ensure that identifiers were accurate, duplicates were resolved transparently, and laboratory data remained faithful to their original source.

b) We reviewed the open-ended categories of qualitative variables to recode them into existing categories or, when needed, to consolidate them into new, consistent response options. For instance, the variable 'labor sector' included 14 predefined categories and one 'other' option. Entries marked as 'other' were reviewed and either reassigned to existing categories or grouped into new options, thereby improving consistency and reducing variability in responses.

c) Due to the complexity of the study's longitudinal design, complementary variables were created. These variables allowed for a more detailed description of the sample composition and operational conditions, thereby facilitating methodological adjustments to the frequency of visits.

1) The total number of participants did not always match the number of surveys completed in each data collection phase (Cadena, Handal, Muñoz, Orozco (2025)). To address this, we created three participation variables: *Permanence,* which indicates whether a participant remained in the study at a given time; *Assessment*, which indicates whether the corresponding evaluation was completed; and *Observation*, which provides qualitative information on the reasons for non-evaluation, when applicable. The combinations of these variables are presented in Table 1.

Table 1 Possible combinations of Permanence and compliance with Assessments.

		ASSESSMENT	
		YES	**NO**
PERMANENCE	Yes	Participant continued in the study and complied with study activities	Participant continued in the study but did not comply with the activities, the reason is detailed in Observation
	No	⦸	Participant did not continue in the study; the reason is detailed in Observation

2) Socio-structural and political challenges in the implementation area impacted the study protocol, necessitating adjustments to the participant follow-up schedule. To account for these changes, we created the following variables: final wave of follow-up (12 or 18 months), pilot participant (yes/no), type of follow-up (monthly or quarterly), and data collection phase (I or II).

d) Finally, we re-labeled all variables to comply with Stata character limits and to provide clear, standardized descriptions. We harmonized numeric and text formats and developed a variable dictionary for each dataset, detailing the variable name, type, coding, and operational definition.

Data Fusion: As Tavakoli *et al*. (2006) note, in longitudinal studies, separate files from each wave are later merged using participant identifiers. Prior to merging databases across study waves, we conducted a pre-verification of the idmadre identifier. This step was distinct from the review described in the data manipulation section, where typographical errors and duplicates were corrected. In the fusion stage, the goal was to confirm that each participant's identifier was consistently linked to the correct wave. Since wave assignment was determined by the woman's gestational age at enrollment, we cross-checked the Tracking Planner with the variable 'gestational age at entry' to ensure proper alignment. The detailed procedures for defining waves according to gestational age have been reported in a separate publication (Handal *et al*., 2024); here, we highlight their role as a verification step prior to data fusion. This process safeguarded longitudinal integrity and prevented misclassification of follow-up status.

Documentation: We developed a SEMILLA Data Management Manual, which details the study design, database structure, and variable definitions. These efforts ensure that the SEMILLA datasets are clean, consistent, technically compatible, and readily usable by other researchers. The SEMILLA Data Management Manual is available upon request from the Principal Investigators (AJH, FO).

Data Archiving: The final repository includes the raw data in CSPro, processed databases in Stata, complete documentation, and operational protocols. All electronic files are stored in the study's Dropbox folder, with synchronized backup copies maintained on two external hard drives held by each principal investigator, ensuring long-term availability for future analyses.

Data Sharing: Once the primary scientific aims of SEMILLA have been fully analyzed and the main findings published, data sharing will be considered for qualified researchers. At this time, neither the study data nor the associated methodological materials—including questionnaires, flowcharts, and the Data Management Manual—are publicly available. These materials are integral to the construction and interpretation of the final datasets and are part of an ongoing binational research project conducted in a sensitive context involving vulnerable populations.

Data and related methodological materials will be released together at a future date under a formal data-sharing agreement, ensuring compliance with ethical guidelines and institutional regulations. Such agreements will require users to: (1) utilize the data strictly for predefined and approved research purposes, (2) obtain prior approval from their Institutional Review Board (IRB), (3) ensure secure storage and handling of the data, (4) acknowledge and cite the

SEMILLA study and its principal investigators in all outputs, and (5) delete or return the data upon completion of the agreed analyses. Requests for access will be reviewed on a case-by-case basis by the principal investigators.

WHAT THE SEMILLA STUDY TEACHES US

Throughout the implementation of the data management system in the SEMILLA study, we encountered a range of technical, logistical, and methodological challenges that offered valuable lessons. These learnings, presented in Table 2, reflect both the strengths of the procedures adopted and the areas where we identified opportunities for improvement. They serve as a practical guide for optimizing data management in future cohort studies conducted in similar contexts:

Table 2 Key lessons for future cohort studies, based on challenges identified in SEMILLA.

COMPONENT	CHALLENGES IDENTIFIED	LESSON
Planning	The absence of a data management protocol in the early stages of data collection required redesigning instruments and training interviewers while fieldwork was already in progress. This led to initial data entry errors and made it difficult to validate entries promptly due to the initial choice of ODK as the capture system.	-Develop a data management protocol to define resources, timelines, and effective procedures for data generation. -Involve the data manager from the instrument design stage to anticipate critical requirements for data collection, such as the appropriate software based on instrument complexity and the workflow needed to guarantee data quality.
Instrument construction and refinement	Long questionnaires caused participant fatigue; some participants memorized the questions and responded mechanically. Additionally, certain concepts were misunderstood (e.g., paid work, marital status, 'household members'), which required rewording and additional field instructions.	-Validate each instrument not only for content but also for length and usability, evaluating the degree of fatigue of both the participant and the interviewer. -Avoid redundant items and adjust the language to the participants' sociocultural context and the interviewers' training level to ensure that each question yields high-quality responses.
Data collection procedures	Omissions of questionnaires, incomplete activities, and typographical errors in idmadre were observed. The Tracking Planner, implemented from the start, allowed weekly monitoring and required clear justifications. Later, migration to CSPro further strengthened this control by incorporating automatic validations and skip checks during data entry.	Implement a monitoring protocol with periodic data entry validations for each instrument to ensure timely correction of inconsistencies and improve data accuracy.
Staff and training	Some interviewers struggled to build rapport and to correctly apply skip patterns or specialized activities.	-Develop a training manual and a checklist of best practices for interviewers, complemented by continuous feedback. -In some cases, interviewers required additional support to establish rapport with respondents and to apply skip patterns or specialized activities accurately.
Instrument coding	At the beginning of the study, both the field and data management teams were still becoming familiar with the coding rules. This learning phase required ongoing supervision to ensure consistent application of the criteria, which initially resulted in some inconsistencies in variable naming and delays in data cleaning. Once the rules were fully standardized and consolidated, errors could be identified and resolved more efficiently.	-Share coding rules with the field team to streamline data cleaning and, if necessary, to facilitate re-interviews. -Automate double-entry procedures wherever possible and run checks for an early detection of systematic errors.

(Contd.)

COMPONENT	CHALLENGES IDENTIFIED	LESSON
Software programming	The initial use of ODK generated multiple technical limitations (e.g., handling complex skip patterns, ensuring longitudinal follow-up). Detecting these problems and migrating to CSPro was a key decision that improved data quality without affecting the fieldwork calendar.	Maintain flexibility in choosing data collection platform; be open to system migration, even mid-operation, if technical limitations arise. In SEMILLA, transitioning from ODK to CSPro allowed us to resolve operational challenges without disrupting the fieldwork schedule.
Documentation	Technical documentation was prepared at the end of data collection, which made it impossible to identify in time problems such as recall bias in information about pesticide application or last pregnancies, which could not be rectified retrospectively.	Prepare as many manuals, protocols, and field reports as possible before starting the data collection, as each record is an essential resource to reproduce the workflow, guarantee traceability, and facilitate the reuse of the data by other researchers.

The SEMILLA experience demonstrates that having a data management team based in the local implementation context is essential to ensuring the validity, coherence, and traceability of the process. In settings involving binational or multinational academic collaborations, this approach enables dynamic and iterative feedback during the entire data collection process, facilitating the meaningful and valid use of data in future research.

FUNDING INFORMATION

Support for this research was provided by a grant (R01ES026603) from the National Institute of Environmental Health Sciences, National Institutes of Health, USA.

COMPETING INTERESTS

The authors have no competing interests to declare.

AUTHOR CONTRIBUTIONS

NC, FO, and SM jointly conceptualized the scope and objectives of this report, drawing on established theoretical frameworks and best practices in cohort data management. These authors conducted a thorough review of the SEMILLA study data and pertinent literature, systematically organizing and synthesizing the results into the report's thematic structure. NC, FO, and SM also drafted the initial manuscript, comprising the abstract, the description of the study's organic structure, the data management stages, and the lessons learned. FO, as the lead investigator in Ecuador, oversaw the Ecuadorian research team and led the development of the data management structure, with final responsibility over strategic decisions. She also provided critical review and substantive edits to enhance analytical rigor, coherence, and academic style. FM played a key technical leadership role during the preparatory phase, particularly in the development of the data infrastructure. He served as the data manager, developed the variable flow for each SEMILLA questionnaire, and was responsible for creating the data repository. AJH, as lead investigator of the SEMILLA study, provided substantive edits and performed a critical review of all manuscript drafts and the final version. All authors contributed to the development of the manuscript and have read and approved its final form.

AUTHOR AFFILIATIONS

Nataly Cadena orcid.org/0000-0003-2004-6113
Universidad San Francisco de Quito USFQ, Centro de Transferencia de Tecnología, Quito, Ecuador

Fadya Orozco orcid.org/0000-0002-5743-6898
Universidad San Francisco de Quito USFQ, Centro de Transferencia de Tecnología, Quito, Ecuador

Stephanie Montenegro orcid.org/0000-0002-3037-4156
Universidad San Francisco de Quito USFQ, Centro de Transferencia de Tecnología, Quito, Ecuador

Fabián Muñoz orcid.org/0000-0003-2520-1602
Visor Análisis Estadístico Cia. Ltda, Quito, Ecuador

Alexis J. Handal orcid.org/0000-0001-8890-5813
Department of Epidemiology, University of Michigan School of Public Health, Ann Arbor, Michigan, United States

REFERENCES

Adamson, L. and **Graves, A.** (2007) 'Cohort management: Developing and maintaining participant databases in longitudinal studies', *International Journal of Multiple Research Approaches*, 1(2), pp. 147–155. Available at: https://doi.org/10.5172/mra.455.1.2.147

Cadena, N., Handal, A.J., Muñoz, F. and **Orozco F.** (2025) 'Sociodemographic characteristics and predictive factors of attrition: comparison in two final waves of a birth cohort study in Ecuador', *Front. Reprod. Health*, 7, p. 1605182. Available at: https://doi.org/10.3389/frph.2025.1605182

Dhudasia, M.B., Grundmeier, R.W. and **Mukhopadhyay, S.** (2023) 'Essentials of data management: An overview', *Pediatric Research*, 93, pp. 2–3. Available at: https://doi.org/10.1038/s41390-021-01389-7

Feng, Q., Ireland, G., Gilbert, R. and **Harron, K.** (2024) 'Data resource profile: A national linked mother-baby cohort of health, education and social care data in England (ECHILD-MB)', *International Journal of Epidemiology*, 53(3). Available at: https://doi.org/10.1093/ije/dyae065

Graves, A., Ball, J. and **Fraser, E.** (2007) 'Data management: The building blocks of clean, accurate and reliable longitudinal datasets', *International Journal of Multiple Research Approaches*, 1(2), pp. 156–174. Available at: https://doi.org/10.5172/mra.455.1.2.156

Handal, A.J., Hund, L., Páez, M., Bear, S., Greenberg, C., Fenske, R.A. and **Barr, D.B.** (2016) 'Characterization of pesticide exposure in a sample of pregnant women in Ecuador', *Archives of Environmental Contamination and Toxicology*, 70(4), pp. 627–639. Available at: https://doi.org/10.1007/s00244-015-0217-9

Handal, A.J., McGough-Maduena, A., Páez, M., Skipper, B., Rowland, A.S., Fenske, R.A. and **Harlow, S.D.** (2015) 'A pilot study comparing observational and questionnaire surrogate measures of pesticide exposure among residents impacted by the ecuadorian flower industry', *Archives of Environmental and Occupational Health*, 70(4), pp. 232–240. Available at: https://doi.org/10.1080/19338244.2013.879563

Handal, A.J., Orozco, F., Montenegro, S., Cadena, N., Muñoz, F., Ramírez del Rio, E. and **Kaciroti, N.** (2024) 'The Study of Environmental Exposure of Mothers and Infants Impacted by Large-Scale Agriculture (SEMILLA): Description of the aims and methods of a community-based birth cohort study', *Children*, 11(9), p. 1045. Available at: https://doi.org/10.3390/children11091045

Montenegro, S., Handal, A.J. and **Orozco, F.** (2025) 'Development of a structured tracking system to improve retention in a birth cohort in rural Ecuador', *Global Health Action*, 18(1). Available at: https://doi.org/10.1080/16549716.2025.2569207

Tavakoli, A.S., Jackson, K., Moneyham, L., Phillips, K.D., Murdaugh, C. and **Meding, G.** (2006) 'Data management plans: Stages, components, and activities', *Applications and Applied Mathematics*, 1(2). Available at: http://pvamu.edu/pages/398/asp

U.S. Census Bureau. (n.d.) *Census and Survey Processing System (CSPro)* [Software]. Available at: https://www.census.gov/data/software/cspro.html

Young, A., Powers, J. and **Wheway, V.** (2007) 'Working with longitudinal data: Attrition and *retention*, data quality, measures of change and other analytical issues', *International Journal of Multiple Research Approaches*, 1(2), pp. 175–186. Available at: https://doi.org/10.5172/mra.455.1.2.175

TO CITE THIS ARTICLE:
Cadena, N., Orozco, F., Montenegro, S., Muñoz, F. and Handal, A.J. 2026 Data Management in a Community-Based Birth Cohort: What the SEMILLA Study Teaches Us. *Data Science Journal*, 25: 4, pp. 1–12. DOI: https://doi.org/10.5334/dsj-2026-004

Submitted: 01 July 2025
Accepted: 13 January 2026
Published: 06 February 2026

Data Science Journal is a peer-reviewed open access journal published by Ubiquity Press.