U.S. DEPARTMENT OF AGRICULTURE

Development of a WIC Participant and Program Characteristics Longitudinal Data Set

U.S. Department of Agriculture | Food and Nutrition Service | April 2025

Development of a WIC Participant and Program Characteristics Longitudinal Data Set

April 2025

Authors

Jake Beckerman-Hsu, Nicole Huret, and Polina Zvavitch

Submitted to

Food and Nutrition Service
U.S. Department of Agriculture
1320 Braddock Place
Alexandria, VA 22314

Project Officer

Amanda Reat

Submitted by

Insight Policy Research, Inc.
1310 North Courthouse Road
Suite 880
Arlington, VA 22201

Project Director

Nicole Huret

This study was conducted by Insight Policy Research, Inc., under Contract GS-10F-0136X, Order Number 12319820F0078, supported by the U.S. Department of Agriculture, Food and Nutrition Service.

Suggested Citation

Beckerman-Hsu, J., Huret, N., & Zvavitch, P. (2025). *Development of a WIC participant and program characteristics longitudinal data set.* U.S. Department of Agriculture, Food and Nutrition Service.

Acknowledgments

This report was prepared by Insight Policy Research (Insight) under Contract GS-10F-0136X from the U.S. Department of Agriculture, Food and Nutrition Service. It represents a team effort with many individuals making important contributions in support of the authors. We gratefully acknowledge their assistance. In particular, we recognize Amanda Reat, Project Officer, and Kelley Scanlon for their leadership and guidance. We thank Doris Kuehn and Sara Beckwith, from DC WIC, for their useful feedback. We also thank the numerous FNS staff who provided support, including Ruth Morgan, for their expert advice, and encouragement to the Insight team.

The authors also express appreciation to the dedicated individuals in each of the eight State agencies that participated in this study, with special thanks to the State agency that provided the pilot data set. Their time and effort made this research possible.

The authors gratefully acknowledge the numerous individuals from Insight who provided valuable assistance throughout this study. In particular, we thank Kevin Meyers Mathieu and Elaine Wilcox-Cook, who provided support during the planning and initial phases. We also thank Josh Rozen and Craig Ray for their expertise in SAS coding, probabilistic matching techniques, and other analyses. We thank Betsy Thorn, Jennifer Pooler, Kathy Wroblewska, and Jonathan Blitstein for their useful suggestions on the content of this report. We thank Kim Kerson for graphic and formatting support on this document.

Contents

Tables

Figures

Executive Summary

The U.S. Department of Agriculture's Food and Nutrition Service collects cross-sectional data every two years on Special Supplemental Nutrition Program for Women, Infants, and Children (WIC) Participant and Program Characteristics (PC). Currently, there is no way to link records longitudinally, which would allow analysts to follow individuals over time. Three main factors pose potential challenges to creating a longitudinal WIC PC data set: (1) collecting longitudinal data from State agencies, (2) distinguishing multiple WIC participants with the same WIC identification numbers (IDs), and (3) identifying WIC participants with more than one ID. The study team investigated these challenges by collecting longitudinal WIC PC data on infants and children from one State agency and developing a longitudinal WIC PC data set.

Potential Longitudinal Data Challenges Explored

- Collecting longitudinal data from State agencies
- Distinguishing multiple WIC participants with the same ID
- Identifying WIC participants with more than one ID

To select a pilot State agency, the study team surveyed eight State agencies about their capacity to provide the data; three had ideal capabilities, and one was chosen to participate in the study. Next, the study team worked with the selected State agency and its management information system (MIS) contractor to design a longitudinal data request, which included one record per infant/child per week from January 2014 to December 2019. Because the State agency was able to work with its MIS contractor to modify its standard WIC PC data reporting procedures, the State agency reported no challenges providing the longitudinal data. After reviewing the data, the study team used deterministic matching to link records longitudinally by participant ID; 99.9 percent of infants and children had only one ID, and all their records were linked deterministically.

Matching Approaches

Deterministic matching uses participant ID to link records belonging to the same person. It ***cannot*** link records of participants with more than one ID.

Probabilistic matching uses other variables (e.g., date of birth, first name, last name) to link records likely to belong to the same person. It ***can*** link records of participants with more than one ID.

To link the records of participants with more than one ID, the study team used probabilistic matching; that is, the team linked records based on a similarity score calculated using first name, last name, date of birth, sex, race, and ethnicity. In testing, the team's similarity score correctly distinguished which records belonged to the same person for 99.97 percent of potential matches. The matching error rate for the final data set was minimized by (1) manually reviewing a small subset of records identified as similar through the probabilistic matching procedure and (2) manually reviewing records identified as potentially problematic during a final meeting with the State agency.

The study team prepared a high-quality longitudinal WIC PC data set containing records for all infants and children over a six-year period for one State agency (one row per participant per week for more than 43,000 WIC participants, totaling over 700,000 records). The study findings suggest similar results could be achieved in other State agencies, although future research with a representative sample of State agencies is needed to better understand the feasibility of replicating this process nationally. Longitudinal WIC PC data are extremely valuable because they allow for many analyses not possible with cross-sectional data. Examples discussed in this paper include analyzing patterns of enrollment and retention in WIC, tracking children's weight status over time, and tracking anemia screening results over time.

Chapter 1. Introduction

The U.S. Department of Agriculture's (USDA) Food and Nutrition Service (FNS) currently has no participant-level longitudinal data sets containing Special Supplemental Nutrition Program for Women, Infants, and Children (WIC) Participant and Program Characteristics (PC) data. Collected biennially, WIC PC data contain participant-level information on demographics, income, nutritional risks, anthropometrics, hematology, breastfeeding status, and food package prescriptions for WIC participants during the month of April for each reference year. These data have been used since 1992 to provide updated cross-sectional analyses of participant characteristics. While these reports identify population trends over time (e.g., changes in the distribution of participant race, ethnicity, age), individual-level longitudinal data analyses (e.g., change in a participant's anemia status, participant retention) cannot be performed because the data lack identifiers to link participants over time.

Three main factors pose potential challenges to creating a longitudinal data set to facilitate individual-level longitudinal analyses. First, State agencies may encounter various challenges to providing longitudinal data. Second, more than one WIC participant may be given the same identification number (ID), meaning IDs could not be reliably used to link records[1] over time. While unlikely, this scenario would make longitudinal record linkage more challenging. Third, some WIC participants are likely to be assigned more than one ID, especially if they leave the WIC program for a period of time and then re-enroll. Reliably linking records belonging to the same person could be difficult when those records do not share an ID.

Study Goals

1. Evaluate State agencies' capacity to provide raw data to create a longitudinal WIC PC data set.
2. Develop a process for linking WIC PC data longitudinally and evaluate the accuracy of the linkages.
3. Identify potential longitudinal analyses.

This report describes how the study team investigated these three potential challenges by collecting longitudinal data from one State agency (referred to here as the "pilot State agency") and developing a longitudinal WIC PC data set for infants and children. The results of this process were used to consider potential analyses with longitudinal WIC PC data.

[1] For this report, a record is defined as a snapshot of all variables for a WIC participant at a given point in time.

Chapter 2. Data Collection and Preparation

This chapter describes the steps to select one State agency to pilot the longitudinal WIC PC data set, obtain participant data, and process and prepare the data for longitudinal linking.

A. State Agency Selection

The study team used a two-stage process to identify one State agency for this pilot. First, the team identified nine State agencies[2] to survey about their management information systems (MIS) and longitudinal data provision capabilities (section 1). Second, based on its responses, one of the eight State agencies that completed the survey was selected for the pilot (section 2).

1. State Agency Survey Sample

Figure 2.1 summarizes the selection process the study team used to identify State agencies to survey about their data provision capabilities.

Figure 2.1. Sequence for Selecting Nine State Agencies to Complete Preliminary Survey

Note: ITO = Indian Tribal Organization; MDS = Minimum Data Set; MIS = management information system; SDS = Supplemental Data Set

The selection process consisted of seven steps:

1. The study team considered the 50 States, District of Columbia, and Puerto Rico for this data collection because the burden of producing and submitting participant data is often high for small territories and Indian Tribal Organizations.
2. The study team considered only the 35 State agencies that submitted Supplemental Data Set (SDS) variables for WIC PC2020 because data items in the SDS may be helpful in evaluating data linking procedures.
3. The study team excluded State agencies that had significant data quality issues in their WIC PC2020 data submissions (e.g., duplicate case IDs).
4. The study team excluded State agencies that changed food package reporting methods in their WIC PC submissions two or more times since 2015, as the timeframe for the study was initially planned to be 2015–2019.

[2] Because of competing priorities, staff shortages, and lack of information technology (IT) resources, one of the nine selected State agencies was unable to respond to the survey.

5. The study team identified State agencies that used different MIS platforms. It was unclear how the ability to provide longitudinal data would vary by MIS platform. This criterion ensured all nine selected State agencies would not be limited in the same ways by their platform.
6. The study team considered the potential burden on State agencies to provide data for additional analyses under the WIC PC2020 contract and excluded State agencies participating in multiple study tasks.
7. The team selected State agencies within each of the 7 FNS Regions, yielding a final sample of 9 State agencies.

2. State Agency Survey

The study team developed a brief survey to assess each of the nine State agencies' capabilities to provide longitudinal data. Table 2.1 presents the criteria the team used to evaluate the survey responses and select one State agency to pilot the longitudinal data set. Appendix A provides detailed descriptions of each criterion.

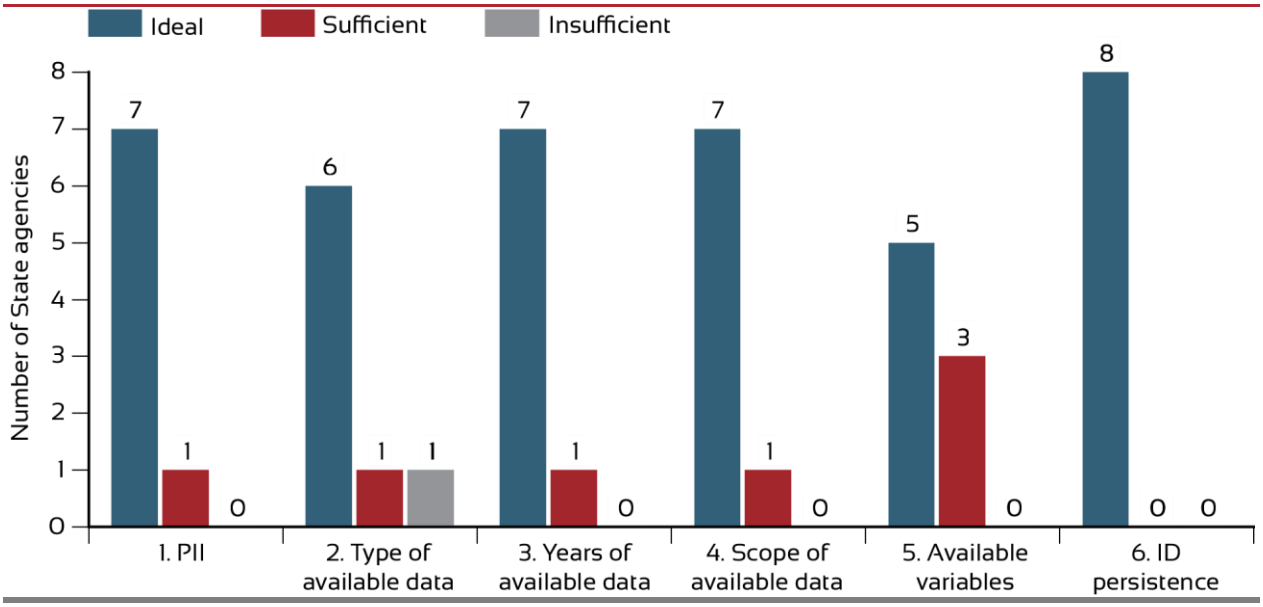
Table 2.1. Summary of MIS Criteria for State Agency Selection

MIS Criteria	MIS Capabilities		
	Ideal	**Sufficient**	**Insufficient**
1. PII	Consistent participant ID, household ID, and first and last name	Consistent participant ID only	No PII
2. Type of available data	All instances of updates and changes to each participant's record	All certification and recertification visits but no records from other visit types	Inconsistent or uncertain frequency of records across caseload
3. Years of available data	5 years of retrospective data	Between 4 and 5 years of retrospective data and capability to provide periodic prospective submissions	Fewer than 5 years of data
4. Scope of available data	All infants and children enrolled at any time during previous 5 years with an indicator for currently enrolled	All infants and children enrolled at any time during the previous 5 years without an indicator for currently enrolled or records for currently enrolled infants and children only	N/A
5. Available variables	Key SDS variables available throughout date range	Key SDS variables not available throughout date range	N/A
6. ID persistence	ID retained for infants and children who re-enroll in WIC after a period of nonparticipation	New ID assigned for infants and children who re-enroll in WIC	N/A

Note: MIS = management information system; PII = personally identifiable information; SDS = Supplemental Data Set

Figure 2.2 summarizes the State agency survey results. Three State agencies were determined to have ideal capabilities for all six criteria.

Figure 2.2. State Agency Survey Results



Note: Because of competing priorities, staff shortages, and lack of IT resources, one of the nine selected State agencies was unable to respond to the survey.
PII = personally identifiable information

B. Data Collection

One of the State agencies with ideal capabilities for all six criteria was invited to pilot the longitudinal data set for this study. The study team held a meeting with this State agency to discuss the best way to obtain data for all infants and children from January 1, 2014, to December 31, 2019. This timeframe was chosen to capture all 5 years of categorical eligibility for infants and children born in 2014 and ensure no interruptions as a result of COVID-19. During this meeting, the team discussed two main components of the data set: (1) variables included and (2) frequency of records per participant.

1. Variable Selection

The WIC PC Minimum Data Set (MDS) and SDS formed the basis of the data request for this study. The study team identified additional variables likely to be available in all State agencies' MIS that could assist in the longitudinal linking process (e.g., first and last name, participant ID). After discussing these options with the pilot State agency, the team developed a list of requested data elements (see appendix B).

2. Record Frequency

A comprehensive longitudinal WIC PC data set should have a record for every time new information is entered into the MIS (e.g., a new height and weight measurement, a new food package prescription, a new anemia screening). The study team and the pilot State agency staff determined the most efficient way to accomplish this task was to extract data for each participant at regular time intervals and remove

identical records as needed. Pilot State agency staff advised pulling one record for each infant/child per week because changes seldom occur more than once a week.

To reduce the volume of data transferred to the study team, the pilot State agency removed duplicate records. For instance, if an infant/child had no change for 5 weeks, only the first week of data was retained (and the other 4 weeks were removed from the data set). This step would not be required for any State agency.

3. Data security

The study team took the following measures to protect the participant data used in this study:

- Data were transferred using a secure file transfer protocol site.
  - Only Insight staff members directly working with these data were given access to the folder on the private server. For all other users, access was prohibited.
- Data were saved on a password-protected private server.
- The final data set shared with FNS was de-identified.

C. Data Processing Procedures

The study team conducted two data processing steps prior to longitudinal linking: (1) data diagnostics and cleaning and (2) verification of assumptions about participant ID.

1. Data Diagnostics and Cleaning

The study team began the diagnostic process by adapting data evaluation processes used for the biennial WIC PC data collection. During the diagnostics phase of this study, the team checked the number of records and unique IDs, ensured minimal missing data, and confirmed the data were internally consistent and in valid ranges. The study team then cleaned the data to meet the same standards used for WIC PC data, created analytic variables for analysis, standardized food codes, and conducted final accuracy checks.

2. Verification of Assumptions About Participant ID

The approach to creating the longitudinal data set assumes records with the same ID truly belong to the same participant. In other words, two different participants would never share an ID. To check this assumption, the study team ensured all records with the same ID had the same first name, last name, date of birth, and sex.

Chapter 3. Data Linking

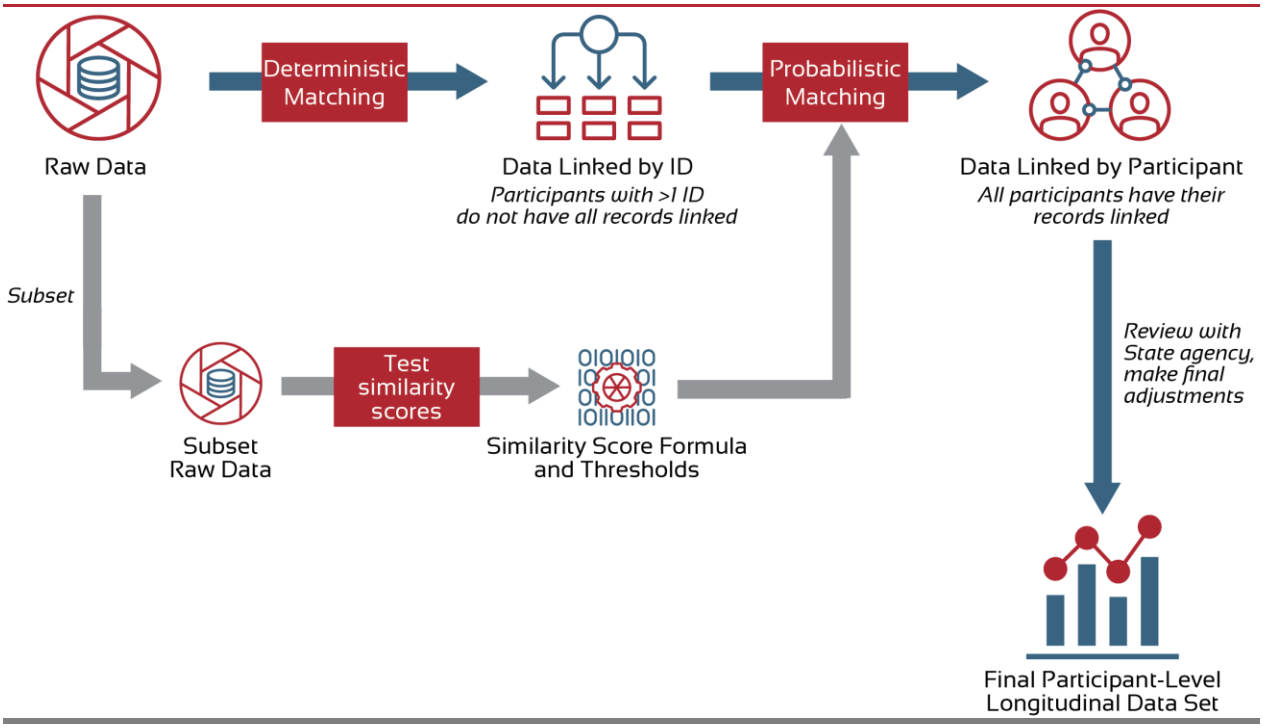
This chapter describes the process for linking the data over time, illustrated in figure 3.1. The study team first linked records deterministically by ID (section A) and then used probabilistic matching to link records belonging to the same infants/children with different IDs (section B). Section C provides sample longitudinal analyses.

Matching Approaches

Deterministic matching uses participant ID to link records belonging to the same person. It ***cannot*** link records of participants with more than one ID.

Probabilistic matching uses other variables (e.g., date of birth, first name, last name) to link records likely to belong to the same person. It ***can*** link records of participants with more than one ID.

Figure 3.1. Summary of Data Linking Approach



A. Deterministic Matching

The study team linked all records with the same ID over time. Because the study team verified at the end of data processing that all records with the same ID had the same first name, last name, date of birth, and sex, the team was confident these deterministic linkages were correctly matching records. For infants/children with only one ID in the data set, this step linked all their records. It was not known at this stage how many infants/children had more than one ID and therefore how much more work would be needed to finalize the longitudinal data set.

As is the case for all State agencies, the pilot State agency implements procedures to minimize the number of participants who receive more than one ID in its MIS over time. Before assigning a new ID to a participant, clinic staff check if any existing records have the same first initial, last name, and date of birth. While this practice is sound, it is not infallible; data entry mistakes could result in the infant or

child receiving more than one ID. It was therefore necessary to conduct probabilistic matching to finalize the data set.

B. Probabilistic Matching

To identify infants and children with more than one ID, the study team designed a probabilistic matching process. The team first tested similarity scores on a subset of the data with three steps:

1. **Select a similarity score formula.** To inform decisions about whether records with different IDs belonged to the same infant or child, the team used variables that should not change over time and are available in all State agencies (i.e., first name, last name, date of birth, sex, race, and ethnicity). The team then selected a similarity score formula that could use those variables to quantify the similarity between records based on two basic concepts:
   - If records share a common value (e.g., female), they are more likely to belong to the same participant and should have a higher similarity score.
   - If records differ on a variable, they are less likely to belong to the same participant and should have a lower similarity score.

2. **Create preliminary probabilistic linkages**. The study team took a subset of the data with one observation per ID from 2014 and one observation per ID from 2015. Each 2014 observation was paired to the most similar observation from 2015 using the similarity score from step 1, creating a preliminary probabilistic linkage.

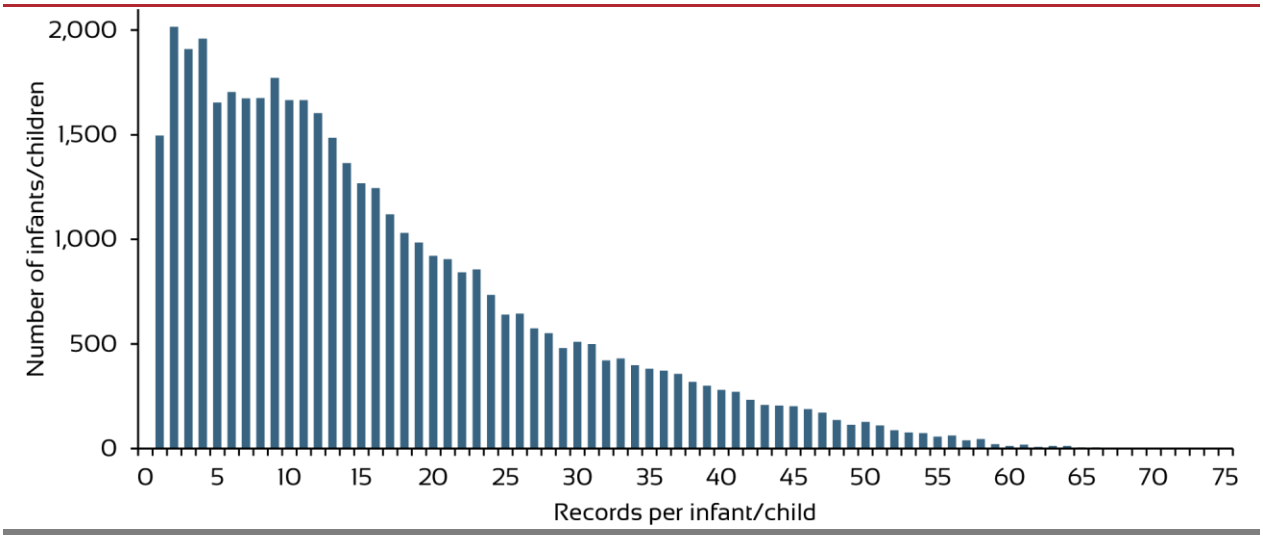
3. **Establish upper and lower similarity score thresholds**. Not all preliminary probabilistic linkages were correct. For example, participants present in 2014 who were not present in 2015 were linked to another person (an incorrect linkage that should not be retained for the final data set). The team selected similarity score thresholds to use on the preliminary probabilistic linkages that would differentiate between correct and incorrect linkages. The team set a lower threshold such that preliminary probabilistic linkages with similarity scores below that threshold could safely be assumed to be incorrect (i.e., the records belong to different infants/children); these preliminary probabilistic linkages would not be retained in the final data set. The team set an upper threshold such that preliminary probabilistic linkages with similarity scores above that threshold could safely be assumed to be correct; these preliminary probabilistic linkages would be retained in the final data set. Preliminary probabilistic linkages with similarity scores between the two thresholds would be reviewed manually to determine whether the records belonged to the same infant/child. For the pilot State agency, about 150 preliminary probabilistic linkages were reviewed manually during this process.

After selecting a similarity score formula and thresholds that could reliably determine when records should be linked, the study team returned to the deterministically matched data to link the records of participants with more than one ID. First, the study team identified when each ID was first and last observed. Then, the study team compared each ID with all other IDs that could belong to the same infant/child (e.g., an ID last observed in January 2015 was compared with all IDs first observed in February 2015 or later) and calculated a similarity score for each comparison. The study team used the similarity thresholds to determine whether to link records with different IDs, reviewing pairs of records manually as needed.

About 30 participating infants and children had 2 IDs; their records were linked probabilistically. The study team manually reviewed about 25 additional IDs after the final meeting with the pilot State

agency. About 10 of these additional IDs were deemed to belong to the same infant/child as another ID, and the records were linked accordingly. The remaining additional IDs reviewed were deemed to belong to different participants and were not linked. Overall, about 0.1 percent of infants/children in the final data set had more than 1 ID. The final data set, which had one row per infant/child per week, included over 43,000 infants/children and a total of over 700,000 records. Figure 3.2 shows the number of records per infant/child, which had a median of 13. About 3 percent of infants/children had only one observation; among the others, the most recent record in the data set was 90 weeks after the first record, on average. Appendix C provides more details about the team's probabilistic matching process.

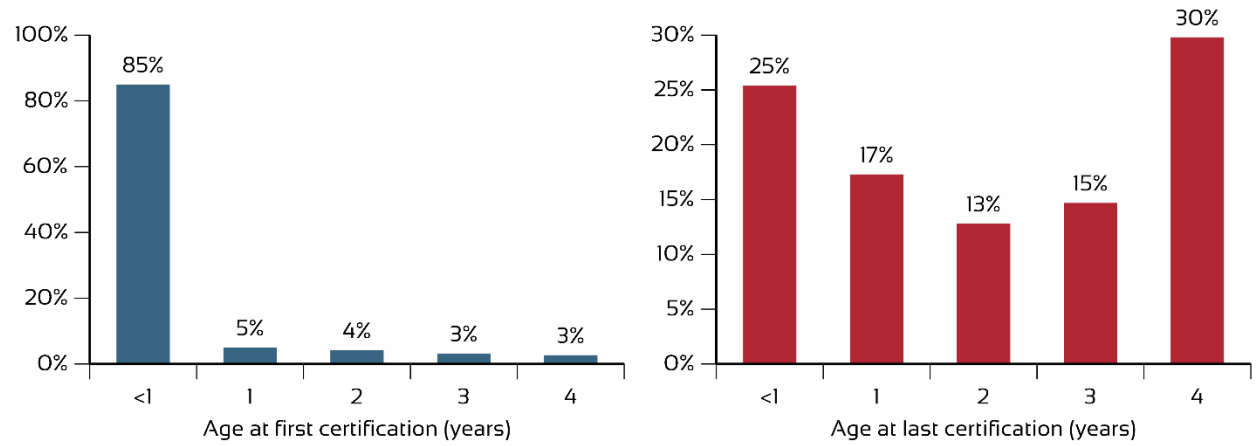
Figure 3.2. Records per Infant/Child in Final Longitudinal Data Set



C. Sample Longitudinal Analyses

The final longitudinal data set facilitates a multitude of analyses not possible with standard WIC PC data, such as tracking infants/children from their first certification through their last (see figure 3.3). Among those born in 2014 who participated in WIC in the pilot State agency, 85 percent first enrolled before turning 1 year old. More than half were last certified before turning 3 years old (56 percent). Future analysis could build on this work to investigate factors associated with first enrolling in WIC later (e.g., after turning 1) and dropping out of WIC earlier (e.g., certifying for the last time by age 3). WIC recruitment and retention efforts could then be targeted to improve participation rates among the eligible infants and children least likely to participate.

Figure 3.3. Age at First and Last WIC Certification Among Infants/Children Born in 2014



Another type of analysis made possible by longitudinal data is tracking health indicators over time (e.g., anthropometric data, hematological data). Figure 3.4 shows the weight status of children born in 2014 at ages 2, 3, and 4. The thickness of the grey lines connecting the weight statuses is proportional to the number of children who experienced each type of change over time. Figure 3.5 shows the proportion of children born in 2014 who were ever anemic and the amount of time until they had a normal hemoglobin level after first being identified as anemic. Further analysis can help WIC agencies better understand which infants and children are most likely to experience poor health outcomes (e.g., unhealthy weight gain, anemia) and when those changes tend to happen to develop targeted approaches to supporting participants' health.

Figure 3.4. Weight Status Trajectories Among Children Born in 2014

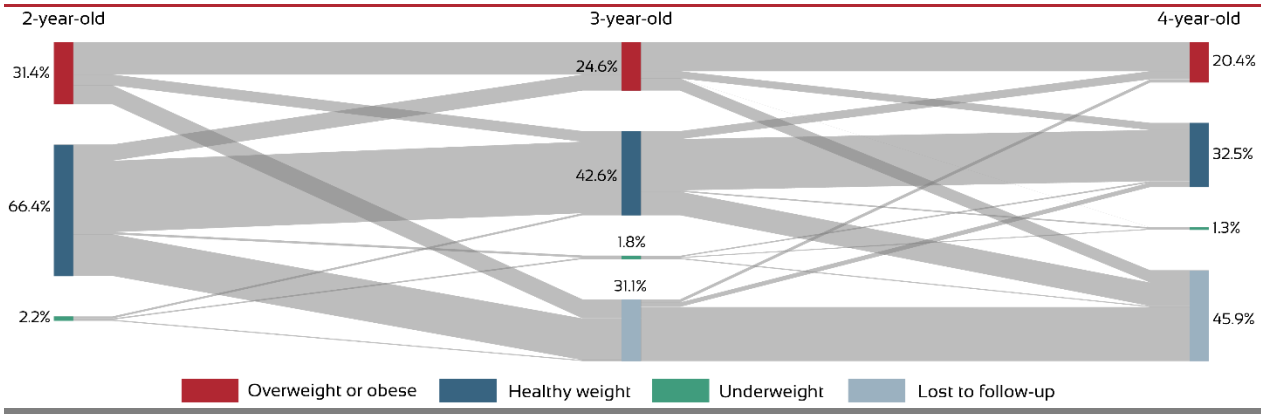
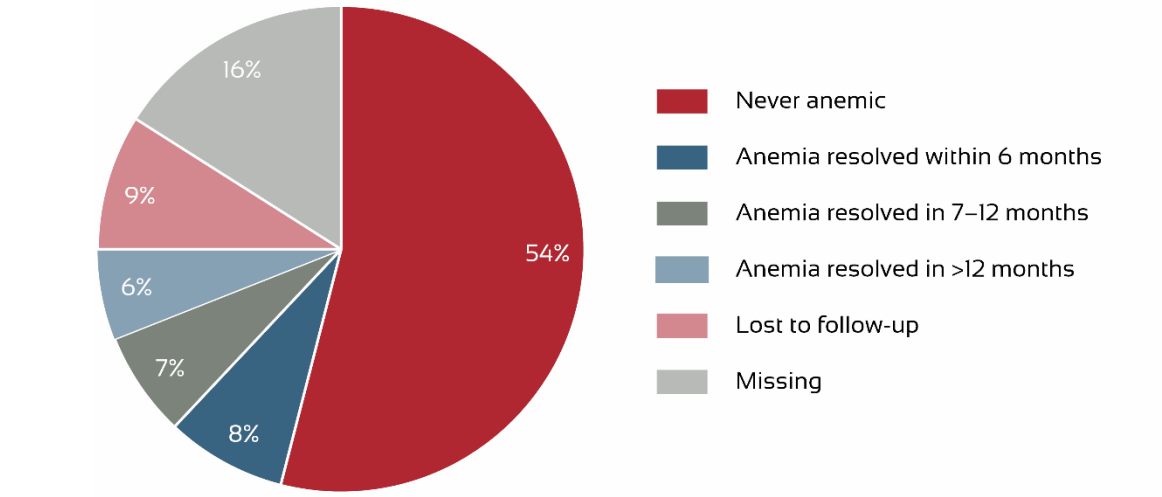
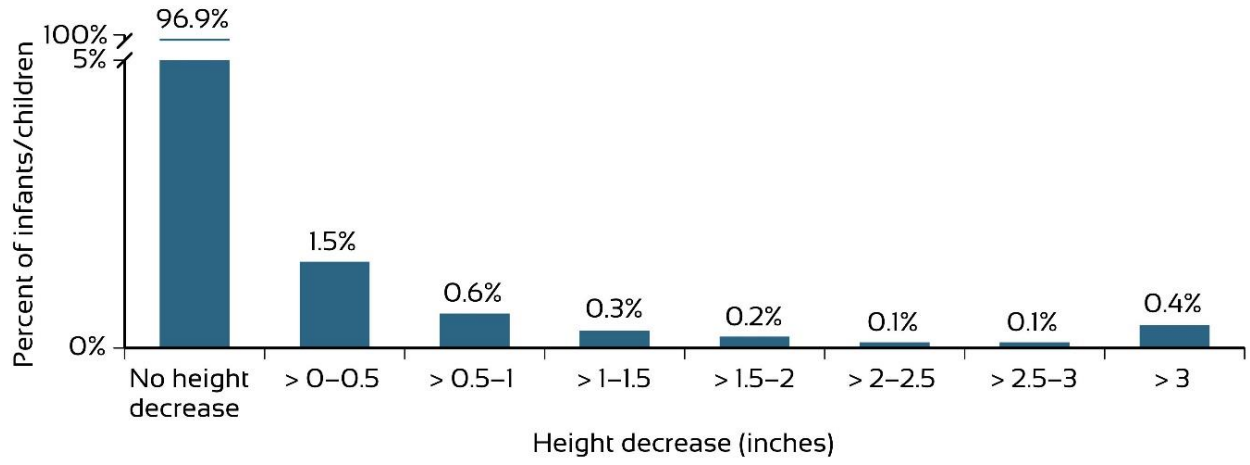


Figure 3.5. Anemia Status and Resolution Among Infants/Children Born in 2014 Aged 9–59 Months



Longitudinal data can also be used for data cleaning. For example, 3.1 percent of infants/children had a decrease in height between at least one consecutive pair of height measurements (figure 3.6).[3] In similar work on child longitudinal data, thorough longitudinal cleaning of anthropometric data resulted in identifying 6.2 percent of body mass index measurements as unreliable; these measurements were removed from the sample prior to analysis.[4] For WIC PC, longitudinal data cleaning could improve the validity of cross-sectional and longitudinal analyses alike for many variables.

Figure 3.6. Decrease in Height Between Consecutive Measurements Among Infants/Children Participating in WIC in 2014–2019



[3] Longitudinal data cleaning was not conducted on anthropometric measurements for the purpose of the example BMI analysis in figure 3.4.
[4] Beckerman-Hsu, J., Aftosmes-Tobio, A., Lansburg, K., Leonard, J., Torrico, M., Kenney, E. L., Subramanian, S. V., Haneuse, S., & Davison, K. K. (2021). Summer weight gain among preschool-aged children with obesity: An observational study in Head Start. *Preventing Chronic Disease, 18,* E25.

Chapter 4. Discussion

The study team produced a high-quality longitudinal WIC PC data set for infants and children in the pilot State agency. This chapter discusses the lessons learned and the feasibility of preparing longitudinal data for other State agencies. Limitations and considerations for future research follow.

A. Lessons Learned

The study team identified areas for improvement in three main activities: (1) assessing State agency capacity to provide longitudinal data, (2) data collection, and (3) data linking.

1. State Agency Capacity to Provide Longitudinal Data

The survey administered to eight State agencies helped identify those with higher capacity for providing ideal data for this study (i.e., 5 years of data for infants and children that could be reliably linked longitudinally). Table 2.1 summarizes the criteria used to evaluate State agencies' responses to the survey. During the study, the team identified other factors that could inform the level of effort required to produce a longitudinal data set. Table 4.1 presents updated criteria for assessing State agencies' capacity to provide longitudinal data. Differences between tables 2.1 and 4.1 are described below.

Table 4.1. Updated Criteria for Assessing Capacity to Provide Longitudinal Data

MIS Criteria	MIS Capabilities		
	Ideal	**Sufficient**	**Insufficient**
1. PII	Consistent participant ID, first name, last name, and other potential linking variables (e.g., household ID, date of first visit, home address, authorized representative first and last name)	Consistent participant ID, first name, and last name	No PII
2. Type of available data	All instances of updates and changes to each participant's record OR Records at any date in the past (e.g., information about each participant as of the first of each month)	Depends on analyses of interest Example: to analyze certification patterns, all certification and recertification records	Inconsistent or uncertain frequency of records across caseload
3. Years of available data	Depends on analyses of interest Example: to analyze infant/child WIC retention, more than five years of data (more data will increase the sample size)	Depends on analyses of interest Example: to analyze infant/child WIC retention, five years of data	Depends on analyses of interest Example: to analyze infant/child WIC retention, less than five years of data

MIS Criteria	MIS Capabilities		
	Ideal	Sufficient	Insufficient
4. Scope of available data	All participants enrolled at any time during period of interest	Depends on analyses of interest Example: infants/children only	Records for currently enrolled participants only
5. Non-PII variables	Key SDS variables, date of each record, certification start and end dates, date of most recent benefit issuance, benefit issuance frequency, and date of most recent benefit redemption[a] available throughout date range	Date of each record available throughout date range	Date of each record not available throughout date range
6. ID persistence	ID retained for participants who re-enroll in WIC after a period of nonparticipation and agency can provide detailed description of procedures for assigning new IDs	New ID assigned for participants who re-enroll in WIC and agency cannot describe procedures for assigning new IDs	N/A
7. Process for addressing participants with multiple IDs	Agency can provide a detailed description of how it determines a single participant has been assigned more than one ID and procedures for rectifying such occurrences	Agency lacks standard procedures for identifying participants with more than one ID and/or addressing this issue	N/A
8. Data quality	Agency has strong protocols in place to minimize errors in data	Data quality issues in a limited number of variables	Data quality issues in many or most variables
9. Date of variable measurement	Can provide the date of measurement for all variables	Can provide the date of measurement for some variables	N/A

Note: MIS = management information system; PII = personally identifiable information; SDS = Supplemental Data Set
[a] The certification start and end dates, date of most recent benefit issuance, benefit issuance frequency, and date of most recent benefit redemption for each record would not be used to prepare the longitudinal data set. They would, however, be useful for many longitudinal analyses.

- **PII.** This study demonstrated that records from the pilot State agency can be reliably linked using participant ID for deterministic matching and first name, last name, and variables from the MDS for probabilistic matching. Therefore, participant ID, first name, and last name are considered requirements for sufficient capabilities. While participant ID may not actually be required to produce a longitudinal data set (i.e., it may be possible to produce a high-quality longitudinal data set with only probabilistic matching), it is not clear how well the probabilistic matching results from the pilot State agency would generalize to other State agencies. Therefore, participant ID has been retained as a requirement for sufficient capabilities. Similarly, while other linking variables such as household ID were not used in this study, they could be useful for other State agencies, so additional potential linking variables were added to the criteria for ideal capabilities.

- **Type of available data.** A comprehensive longitudinal WIC PC data set should have a record for every time new information is entered into the MIS. Some State agencies may have a variable on each record in their MIS that indicates the last time the record was updated, which could be used to ensure all needed records are pulled for such a comprehensive longitudinal data set. For the pilot State agency, the easiest approach was to pull records based on date (i.e., one record per participant per week) and deduplicate, rather than pulling records based on whether new information had been added about the participant. The ability to pull records based on date was added as an option for ideal capabilities. Sufficient capabilities for the type of available data depend on the analyses of interest; some analyses may require only one record per participant per year, while others may require more frequent updates.
- **Years of available data.** Providing at least 5 years of data was a specific requirement for this study. The criteria for ideal, sufficient, and insufficient years of available data for other longitudinal data efforts depend on the analyses of interest.
- **Scope of available data.** This study focused exclusively on infants and children, but all WIC participants should be considered when evaluating the general capability to provide longitudinal data. Originally, part of the scope of available data criterion was the ability to provide a variable indicating current enrollment in WIC, but this variable was deemed unnecessary for preparing a longitudinal data set, so it has been removed from the criteria presented in table 4.1. Sufficient scope of available data depends on the analyses of interest; some analyses may only require data on a subset of participants. The study team also updated the criteria for insufficient scope of available data. While "records for currently enrolled infants and children only" was initially considered sufficient, the study team moved it to the insufficient category because this type of sample would fail to include all records for participants with more than one ID. As an example, consider a four-year-old child currently enrolled in WIC who first enrolled at birth. If this child was given a second ID at age 1, a sample of currently enrolled children may only have records with this child's most recent ID, meaning the dataset would be missing all information about the child's first year in WIC.
- **Non-PII variables.** The only non-PII variable required for longitudinal data linkage is the date of each record. This requirement was not explicitly stated in the criteria from table 2.1; it was added for table 4.1.

  During the study, it became evident it would have been helpful to have variables to precisely define when participants were actively enrolled. Although preparation of longitudinal data does not require indicators of active enrollment, many analyses of interest could be strengthened by knowing with certainty when participants were actively enrolled. To address this issue, five variables were added to the ideal criterion: certification start and end dates, date of most recent benefit issuance, benefit issuance frequency, and date of most recent benefit redemption. Certification start and end dates define enrollment in WIC, and State agencies use date of food benefit issuance for reporting on the monthly participation caseload to FNS on the FNS-798 form. Benefit issuance frequency and date of most recent benefit redemption can be used to show how long each participant actively made use of their benefits. FNS and State agencies should consider including any other variables that would help define active enrollment. If the State agency cannot provide such variables, the agency should indicate if there is another way for analysts to know when an infant/child was actively enrolled.

- **ID persistence.** To achieve ideal ID persistence, participants should retain their ID after a period of nonparticipation and the State agency must also be able to provide a detailed description of its procedures for assigning new IDs. The description of the procedure for assigning new IDs helps with two issues related to creating a longitudinal data set. First, it provides information about how each State agency ensures multiple people never have the same ID. Cases of more than one participant sharing a given ID need to be identified and resolved prior to carrying out the longitudinal linkage methods used in this study. Second, it provides information about the State agency's process for ensuring individuals only have one ID. Understanding this process can help analysts identify individuals with more than one ID and link their records appropriately.
- **State agency process for handling participants with more than one ID (new).** To properly design a longitudinal matching procedure, it is important to understand (1) how State agencies determine a single participant has been assigned more than one ID and (2) the procedures for rectifying this issue when it occurs. The first item, a detailed description of how State agencies determine a single participant has been assigned more than one ID, will help analysts design an accurate probabilistic matching process, including which variables to include in the similarity score and the protocols for manual review of records with high similarity scores. It is particularly helpful for agencies to describe if there are certain variables they check to determine if two records with different IDs belong to the same participant and any variables they would ignore or consider to be less important. The second item, the procedures for handling participants assigned more than one ID would also help inform the probabilistic matching procedure. Both of these items should be readily available, as State agencies conduct regular data deduplication to identify and address any individuals with more than one ID.
- **Data quality (new).** Any errors in the variables used for probabilistic matching can reduce the accuracy of the procedure. Ideally, State agencies have very high quality data with strong protocols in place to prevent data errors and catch and correct them when they do occur. An agency can have sufficient capacity to provide longitudinal data despite a limited number of known data quality issues. Analysts could simply make sure to use only highly reliable variables when linking records (i.e., avoid variables with known issues).
- **Date of variable measurement (new).** It would be ideal if, for each variable to be analyzed longitudinally, State agencies could provide a second variable for the date of measurement (e.g., a variable with the date SNAP participation information was updated in addition to the variable on SNAP participation). These dates of variable measurement/update generally play a limited role in preparing a longitudinal data set, but they are important in many longitudinal analyses. Consider a participant with the same breastfeeding status in January and February. It is important to know if breastfeeding was assessed in both months and there was truly no change, or if breastfeeding was only assessed in January and the February record is simply displaying the most recent breastfeeding data collected in January. Thus, the date of variable measurement criterion was added for ideal capacity.

2. Data Collection

The pilot State agency emphasized that while the process of preparing the requested data was not difficult, it can be challenging to manage during particularly busy times of the year. For this study, the timing of the original data request conflicted with other major projects and had to be postponed. Involvement in this study required 5 hours of State agency staff time and 80 hours of MIS contractor staff time, more than is typical for the biennial cross-sectional WIC PC data pull. All work for this study was completed under the pilot State agency's existing MIS contract; a longitudinal data pull may be more burdensome in State agencies where the contract does not cover this work.

A primary consideration for data collection is whether State agencies should include records at all time points for each participant, even if there has been no change from the previous time point. The study team asked the pilot State agency to remove records with no change to reduce file size. The requested deduplication resulted in a missed opportunity to identify active enrollment; even if there was no change in a record since the previous time point, that newer record still added information to the data set because it indicated the participant was still enrolled. For example, consider a child with a record update on January 1, 2015, who dropped out of the program on February 28, 2015. If nothing on this child's record changed after January 1, 2015, all records from February 2015 would have been removed from the data set during the deduplication step. Even though the child was enrolled for the entire month of February, there would have been no records showing this enrollment. Had the study team asked the State agency to include one or more variables that indicated active enrollment (e.g., date of most recent benefit issuance, date of most recent benefit redemption), deduplicating would not have led to the loss of any information from the data set. The best approach to take in the future would depend on the availability of variables that indicate active enrollment and the importance of enrollment status in the analyses planned.

3. Data Linking

Deterministic matching is simple to implement. For the pilot State agency, 99.9 percent of infants and children had only 1 ID and therefore had all their records linked using deterministic matching. The proportion of infants and children with all their records linked deterministically was not known until probabilistic matching was conducted; it is only through the probabilistic matching process that participants with more than one record can be identified, and the total number of unique participants can be determined. Working with any State agency, it would not be possible to decide if probabilistic matching is worthwhile after having completed only deterministic matching. The proportion of infants and children that could have all their records linked with only deterministic matching would depend on the State agency's processes for assigning IDs.

For the pilot State agency, probabilistic matching was also simple to implement. The only challenge that arose stemmed from an incorrect assumption the study team made. To make the probabilistic matching process more efficient, the study team assumed that if an infant or child was assigned a second ID, records with the second ID would appear only after the final record with the first ID. The study team learned this assumption was not valid during the final meeting with the pilot State agency (i.e., the State agency would archive records assigned a second ID and return to using the first ID). To avoid making incorrect assumptions in the future, the study team added the process for addressing participants with multiple IDs to table 4.1 as one of the criteria used to assess agencies' capacity to provide longitudinal data.

Probabilistic matching could be more complex for other State agencies. First, virtually endless possibilities exist for defining similarity scores. The study team's first approach worked exceedingly well to differentiate records that did and did not belong to the same participant in the pilot State agency (see figure C.2), but that may not be the case for all State agencies, particularly larger State agencies and State agencies with lower data quality. It may be necessary to test different similarity score approaches in other State agencies to find one that performs well, thereby reducing the manual review required to properly link all records.

A second aspect of probabilistic matching that could be more complex for other State agencies is the process to assess the performance of the similarity scores and establish thresholds (shown with grey arrows in figure 3.1). Evaluating the performance of a similarity score requires a data set with a variable that indicates all records that should be matched longitudinally (see table C.2 for a simplified example).

In the pilot State agency, constructing this data set required little effort. Because this State agency has a robust system for ensuring each participant is assigned only one ID, only about 10 participants were found to have a different ID in 2014 and 2015; the study team manually reviewed about 130 pairs of records to identify these participants.

In other State agencies, if the procedure for ensuring participants only have one ID is less robust, this stage could require more time and effort. The amount of time and effort could be limited by simply using fewer records to set the similarity score thresholds. However, if more records are used to set the similarity score thresholds, the thresholds will be less likely to lead to incorrect longitudinal linkages.

The third step in probabilistic matching that could require more time and effort in other State agencies is the manual review process during the preparation of the final longitudinal data set. This study required manual review for only about 150 pairs of records that did not match on ID. Depending on the performance of the similarity score in other State agencies and the size of those State agencies, it could be necessary to conduct more manual reviews to ensure the records of all participants with more than one ID are properly matched. If many records must be reviewed manually even after optimizing the similarity score, it would be possible to simply skip this step and accept a higher error rate in longitudinal linkage. Especially if the number of participants to potentially have their records linked after manual review is a very small proportion of the population, the added value of manual review may not be worth the cost.

B. Limitations

The pilot State agency was selected for this study because it had ideal capacity for providing longitudinal data (e.g., could provide key SDS variables throughout date range, could provide 5 years of retrospective data). Only two other State agencies of the eight that responded to the survey had the same level of capacity. These circumstances limit the generalizability of the study's findings. A full survey of all 89 State agencies is required to determine with greater certainty how the results from this study could be generalized to other State agencies.

It is not possible to know with complete confidence if all longitudinal matches were made correctly. When records shared an ID, the study team had high confidence of a correct longitudinal match. In all other cases, the study team relied on similarity scores. Participants with more than one ID may not have been correctly matched; some may have had similarity scores below the threshold for manual review, while others may have been reviewed manually and the wrong matching decision could have been made. It is also possible the study team linked records that were similar but did not belong to the same participant.

C. Areas for Future Longitudinal WIC PC Research

Before creating longitudinal WIC PC data sets for the remaining 88 State agencies, it may be helpful to identify analyses of interest on the longitudinal data. For example, relatively few records and variables are required if the main analysis of interest is the first and last WIC certification of each infant/child (see figure 3.3). However, if the goal is to conduct a longitudinal analysis on all variables in the MDS, a similar approach to the current study would be required.

The study team identified three areas for future research on longitudinal WIC PC data.

- **Include women in the longitudinal data sets.** Compared with infants and children, women are more likely to leave WIC for extended periods (e.g., between pregnancies) and then re-enroll. The proportion of women with more than one ID could therefore be greater than the proportion of infants and children with more than one ID, requiring a heavier emphasis on probabilistic matching to produce a final longitudinal data set for women. If the study period is long enough, completing a longitudinal data set with women could require linking their infant/child and adult records.
- **Link household members.** Members of the same household cannot be linked in current WIC PC data. Because eligibility and benefit accuracy depend strongly on having an accurate household ID, linking households may be as simple as asking State agencies to include a household ID. Cross-sectional and longitudinal logic checks similar to those conducted in the current study could identify issues with household IDs (e.g., ensure household size does not differ for members of the same household).
- **Link records across State agencies.** This study did not include a systematic assessment of the feasibility of linking records longitudinally across State agencies. However, the study does provide insight into the potential for doing so. Unlike the within-State linkage in this study, no ID would remain consistent for WIC participants who move to another State. Therefore, all cross-State matches would rely on a probabilistic matching process similar to that used to set the similarity score thresholds in this study (see appendix C). During this analysis, the study team observed that all IDs from 2014 that were also present in 2015 were linked in the preliminary probabilistic matching procedure (i.e., based on first name, last name, date of birth, sex, race, and ethnicity). These results suggest records could be linked across State agencies with a low rate of error. However, further work in this area is required because this process would present many challenges not involved in linking records longitudinally for a single State agency. For instance, it would likely be necessary to include different variables in the similarity score (e.g., verification of certification transfer nutritional risk code). In addition to creating cross-State longitudinal linkages, this process could identify individuals enrolled in more than one State at a given time.

Appendix A. Longitudinal Capability Criteria

This appendix provides a detailed description of the criteria the study team used to assess State agencies' longitudinal data capabilities.

A. Available Personally Identifiable Information

Personally identifiable information (PII), particularly a participant identification number (ID), is central to linking records longitudinally.

- **Ideal capabilities.** The State agency can provide a participant ID, household ID, and participant first and last name.
- **Sufficient capabilities.** The State agency can provide a participant ID only.
- **Insufficient capabilities.** The State agency cannot provide any participant identifiers other than case ID—provided in the Minimum Data Set—and therefore cannot link a participant's records over time deterministically.

B. Type of Available Data

The frequency with which infant and child participant records are updated can vary. Inconsistent frequency of repeated measures may limit the ability to capture all data.

- **Ideal capabilities.** The State agency can provide all instances of updates and changes to each participant's record. While a variable describing the reason for the record update (e.g., certification visit, measurement visit, other) would be preferred and was originally considered a requirement for having ideal capabilities, it was not deemed necessary to consider the State's capability "ideal."
- **Sufficient capabilities.** The State agency can provide records from all certification and recertification visits but no records from other visit types.
- **Insufficient capabilities.** The State agency cannot provide consistent frequency of measurement across the caseload (e.g., all visits for infants and only certification visits for children) or would not be able to provide data from all certification visits.

C. Years of Available Data

For this study, the objective was to produce a longitudinal data set that can connect an infant's initial certification records to all subsequent visits (e.g., recertification or measurement) until loss of categorical eligibility at age 5.

- **Ideal capabilities.** The State agency can provide 5 years of retrospective data.
- **Sufficient capabilities.** The State agency can provide between 4 and 5 years of retrospective data and prospective data until 5 total years of data are collected.
- **Insufficient capabilities.** The State agency cannot provide 5 years of total data.

D. Scope of Available Data

The objective of this study was to include all infants and children who participated in WIC in the pilot State agency.

- **Ideal capabilities.** The State agency can provide records for all infants and children over the available data period, with an indicator for currently enrolled participants.
- **Sufficient capabilities.** The State agency can provide records for all infants and children over the available data period without an indicator for currently enrolled participants, or the State agency can provide records only for currently enrolled infants and children.
- **Insufficient capabilities.** The State agency can provide records for only a subset of enrolled infants and children.

E. Available Variables

The ability to provide Supplemental Data Set variables may improve the longitudinal linking process. Specifically, the following five variables were considered: date of first WIC certification, education level of parent, number in household on WIC, birth weight, and birth length.

- **Ideal capabilities.** The State agency can provide all five variables for all available years of data.
- **Sufficient capabilities.** The State agency can provide fewer than five of the variables for all available years of data.

F. ID Persistence

Some State agencies may reassign the same ID to a participant upon re-enrollment, while others may assign a new ID. State agencies that can provide a persistent ID over time to participants who re-enroll are preferred over those that assign new IDs to participants upon re-enrollment. While not preferred, State agencies that lack a persistent ID are still considered to have sufficient capabilities, although incomplete longitudinal matching in the final analytic data set would likely be more common.

- **Ideal capabilities.** The State agency can provide a consistent identifying variable for participants who re-enroll in WIC after a period of nonenrollment. Each participant retains their original ID when re-enrolling in the program.
- **Sufficient capabilities.** The State agency cannot provide a consistent identifying variable for participants who re-enroll in WIC after a period of nonenrollment. Participants are assigned new IDs when re-enrolling in the program.

Appendix B. Longitudinal Data Request

Table B.1. Requested Data Items for Longitudinal Data Extract

	Data Items
WIC PC Minimum Dataset (MDS) Variables	1. State Agency ID
	2a. Local Agency ID
	2b. Service Site ID
	3. Participant ID
	4. Date of Birth
	5. Race/Ethnicity
	6a. Certification Category
	6b. Expected Date of Delivery
	6c. Weeks Gestation
	7. Date of Certification
	8. Sex
	9. Risk Priority Code
	10a. Participation in TANF
	10b. Participation in SNAP
	10c. Participation in Medicaid
	11. Migrant Status
	12. Number in Family/Economic Unit
	13a. Family/Economic Unit Income
	13b. Income Period
	13c. Income Ranges
	14a. Nutritional Risk 1
	14b. Nutritional Risk 2
	14c. Nutritional Risk 3
	14d. Nutritional Risk 4
	14e. Nutritional Risk 5
	14f. Nutritional Risk 6
	14g. Nutritional Risk 7
	14h. Nutritional Risk 8
	14i. Nutritional Risk 9
	14j. Nutritional Risk 10
	15a. Hemoglobin
	15b. Hematocrit
	15c. Date of Blood Test
	16a(i). Participant's Weight in Pounds
	16a(ii). Nearest Quarter Pound of Participant's Weight
	16b. Participant's Weight in Grams
	17a(i). Participant's Height in Inches
	17a(ii). Nearest Eighth of an Inch of Participant's Height
	17b. Participant's Height in Centimeters
	18. Date of Height and Weight Measurement
	19a. Currently Breastfed
	19b. Ever Breastfed
	19c. Length of Time Breastfed
	19d. Date Breastfeeding Data Collected

	Data Items
WIC PC Minimum Dataset (MDS) Variables (continued)	20a. Food Code 1
	20b. Food Code 2
	20c. Food Code 3
	20d. Food Code 4
	20e. Food Code 5
	20f. Food Code 6
	20g. Food Code 7
	20h. Food Code 8
	20i. Food Code 9
	20j. Food Code 10
	20k. Food Code 11
	20l. Food Code 12
	20m. Food Code 13
	20n. Food Code 14
	20o. Food Package Type
WIC PC Supplemental Dataset (SDS) Variables	21. Date of First WIC Certification
	22. Education Level
	23. Number in Household in WIC
	24. Date Previous Pregnancy Ended
	25. Total Number of Pregnancies
	26. Total Number of Live Births
	27a(i). Prepregnancy Weight in Pounds
	27a(ii). Nearest Quarter Pound of Participant's Prepregnancy Weight
	27b. Participant's Prepregnancy Weight in Grams
	28a(i). Weight Gain During Pregnancy in Pounds
	28a(ii). Nearest Quarter Pound of Participant's Weight Gain During Pregnancy
	28b. Participant's Weight Gain During Pregnancy in Grams
	29a(i). Birth Weight in Pounds
	29a(ii). Ounces of Birth Weight
	29b. Birth Weight in Grams
	30a(i). Length at Birth in Inches
	30a(ii). Nearest Eighth of an Inch of Length at Birth
	30b. Length at Birth in Centimeters
	31. Participation in the Food Distribution Program on Indian Reservations
Additional Linking Variables	32. Record Date
	33. Participant First Name
	34a. Participant Middle Name
	34b. Participant Last Name
	34c. Family ID
	35. eWIC ID
	36a. Participant Address: Street Line 1
	36b. Participant Address: Street Line 2
	36c. Participant Address: City
	36d. Participant Address: Zip code
	37. Most recent visit date

Appendix C. Additional Probabilistic Matching Details

This appendix provides additional details on the steps for probabilistic matching the study team used to identify infants and children with more than one ID: define a similarity score (section A), assess similarity score performance and establish thresholds (section B), and apply probabilistic matching to create the longitudinal data set (section C).

A. Define a Similarity Score

To inform decisions about whether records with different IDs belonged to the same infant or child, the team first considered variables that should not change over time and are likely to be available in all State agencies: first name, last name, date of birth, sex, race (five indicator variables: American Indian/Alaska Native, Asian, Black/African American, Native Hawaiian/Pacific Islander, White), and ethnicity (Hispanic or non-Hispanic). These variables are strong candidates for successfully matching records longitudinally in the absence of a consistent ID variable. While time-varying factors can be used to match records probabilistically (e.g., certification category, certification date, issuance date), the study team decided to first assess the performance of the probabilistic matching using this simpler approach.

Next, the study team chose a method to calculate a similarity score to quantify the similarity between records based on the variables selected. The basic concepts of the similarity score follow:

1. If records share a common value (e.g., female) or are similar on a variable (e.g., Carole and Carol), they are more likely to belong to the same participant.

2. If records differ on a variable, they are less likely to belong to the same participant.

To operationalize the first concept (i.e., records sharing common values), the study team used the Fellegi-Sunter Model.[C1] The core principle of this method is that if two records share an uncommon feature (e.g., a unique last name), they are more likely to belong to the same infant/child than records that share a more common feature (e.g., being female). When two records have the same value of a given variable, the similarity score assigned is given by the following equation:

$$-\log_2(proportion\ of\ shared\ variable\ value)$$

For example, if half the records are male, two records that are male would have a similarity score based on sex equal to $-\log_2(0.5) = 1$. The overall similarity score for two records was calculated as the sum of the similarity scores for all variables (i.e., first name, last name, date of birth, sex, race, and ethnicity).

To operationalize the second concept (i.e., records with different values), the study team assigned a similarity score of zero to records with different values. Typically, negative similarity score points are assigned in this situation, and a score of zero is reserved for cases where one or both records are missing data (because missing data are assumed to provide no indication of the likelihood that two records belong to the same person). However, because there were no missing data for the variables selected in this study, a similarity score of zero was assigned when records differed on a variable.

[C1] Wright, G. (2011). *Probabilistic record linkage in SAS.* Kaiser Permanente. https://www.lexjansen.com/wuss/2011/data/Papers_Wright_G_76128.pdf

If variables used for the similarity score for other State agencies do have some missing data, zero points should be assigned when there are missing data, and negative similarity score points should be used when there are differences in a variable. The number of negative points assigned should depend on the reliability of each variable.

Consider a record that is male and a record that is female: The strength of the evidence that the records belong to different infants/children depends on the reliability of the sex variable. If the sex variable has close to no errors, a very large negative value should be assigned to this example pair of records because there would be a high degree of certainty that the records belong to different infants/children. If, on the other hand, errors are common in the sex variable, the difference in sex would be weak evidence that the records belong to different infants/children, and a smaller negative value should be assigned.

The similarity score value assigned when observations differ on a variable must also be appropriate for the range of scores assigned when observations have the same values for the variables included in the similarity score. For example, if two observations share a very rare first and last name, they might be assigned 30 similarity score points based on name. If those observations differ on date of birth, it would not be appropriate to then assign -1,000 points for this difference, especially given that it is easy to make mistakes when entering date of birth into a computer system.

B. Assess Similarity Score Performance and Establish Thresholds

Ideally, a similarity score would perfectly differentiate between records that do and do not belong to the same person (i.e., the score would be very low when records belong to different people and very high when they belong to the same person). In practice, similarity score values often fall between the extremes; it may not be clear based on the similarity score alone whether to link the records for the final data set. Therefore, the study team identified two similarity score thresholds. Similarity scores below the lower threshold would provide strong evidence that a pair of records were *unlikely* to belong to the same person; these pairs of records would not be linked. Similarity scores above the upper threshold would provide strong evidence that a pair of records were *likely* to belong to the same person; these pairs of records would be linked. Similarity scores between the two thresholds would indicate the need for manual review to determine if a pair of records belonged to the same person.

The study team took an empirical approach to establish these thresholds. The team created two subsets of the data from the pilot State agency: one with the first record for each ID from 2014 and another with the first record for each ID from 2015. Each record from 2014 was compared with each record from 2015, and a similarity score was calculated for the comparisons. Next, each 2014 record was paired with the 2015 record with which it had the highest similarity score to create a preliminary probabilistic linkage. This analysis was separate from the deterministic matching already done and was performed to inform the probabilistic matching process to be used to finalize the longitudinal data set.

Table C.1 illustrates a simplified example of this process for five children. The first two children, assigned IDs 1 and 2, were observed in 2014 and 2015. The third child, with last name “An” and ID 3 was observed in 2014 only. The fourth child, with last name “Example,” was observed in 2014 with ID 4 and in 2015 with ID 6. The fifth child, with last name “Another” and ID 5 was observed in 2015 only.

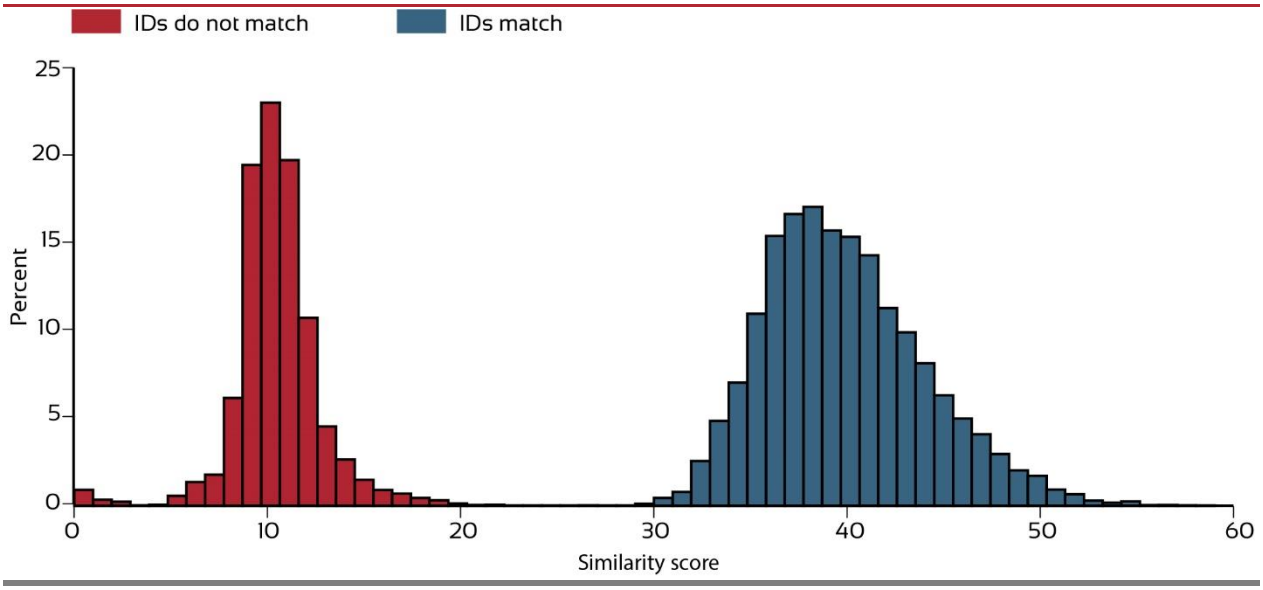
Table C.1. Simplified Example of Preliminary Probabilistic Linkages of 2014 and 2015 Data

2014 Example Records			2015 Example Records			Similarity Score
ID	Date of Birth	Last Name	ID	Date of Birth	Last Name	
1	01/01/2014	This	1	01/01/2014	This	50
2	01/10/2014	Is	2	01/10/2014	Is	50
3	01/20/2014	An	5	01/20/2014	Another	25
4	01/28/2014	Example	6	01/28/2014	Example	50

Figure C.1 shows the similarity scores for pairs of records that did and did not share an ID. The approximately 12,700 preliminary probabilistic linkages between records that shared an ID in 2014 and 2015—and therefore truly belonged to the same infant/child—all had a similarity score of at least 26.111. Based on this result, the lower threshold might be set around 26; pairs of records with lower similarity scores could be assumed to belong to different infants/children and therefore would not be linked longitudinally. However, this number does not account for infants/children with a different ID in 2014 and 2015; the records of those infants/children may have lower similarity scores, in which case the lower threshold should be set below 26.

The approximately 7,000 preliminary probabilistic linkages between records that did not share an ID in 2014 and 2015 had similarity scores as high as 43.488. Because some of these pairs of records may have belonged to the same infant/child in the event they were assigned a different ID in 2014 and 2015, it would be premature to conclude that records belonging to different infants/children could have similarity scores that high. Therefore, it was important to conduct further analysis before finalizing the similarity score thresholds to be used to produce the final longitudinal data set.

Figure C.1. Histogram of Similarity Scores for Preliminary Probabilistic Linkages Between 2014 and 2015 Records That Did and Did Not Match on ID



To finalize the similarity score thresholds, the study team manually reviewed about 130 preliminary probabilistic linkages of records with different IDs that had similarity scores greater than 20; pairs of records with lower similarity scores were assumed to belong to different infants/children. During this

manual review, about 10 pairs of records were determined to belong to the same infant/child. Table C.2 shows the result of this process for the simplified example from table C.1.

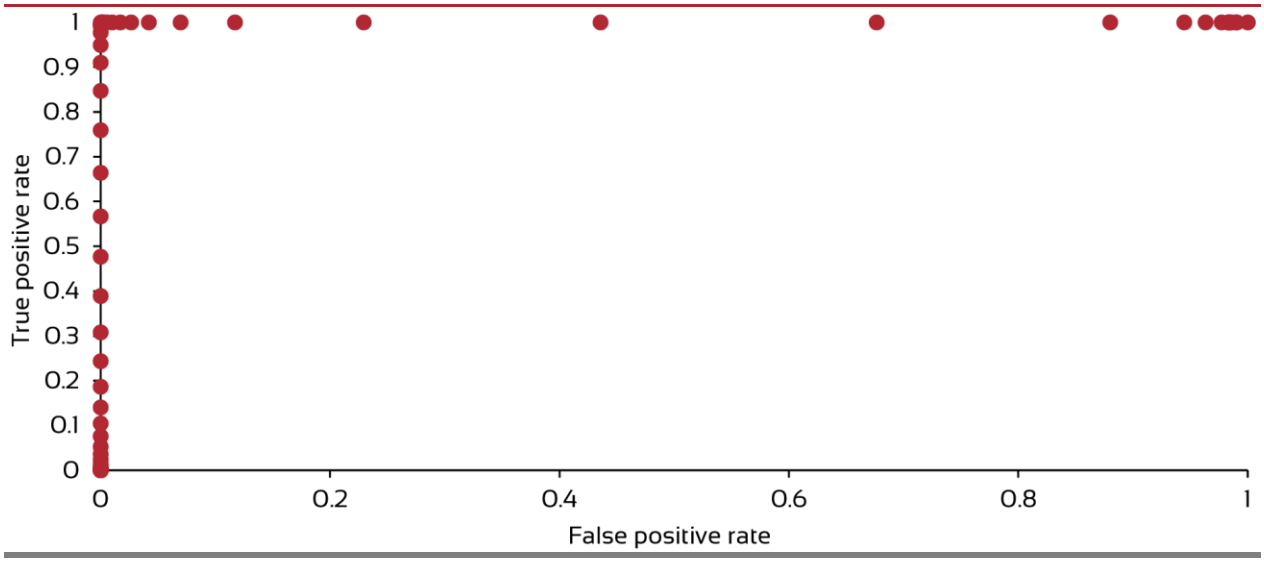
Table C.2. Simplified Example of Preliminary Probabilistic Linkages With Manual Review

2014 ID	Most Similar ID in 2015	Similarity Score	Same Infant/Child
1	1	50	Yes
2	2	50	Yes
3	5	25	No
4	6	50	Yes

The other check to conduct before finalizing the similarity score thresholds was to look for any pairs of records that did share an ID but were not linked during the preliminary probabilistic matching process. While unlikely, it was possible that a 2014 record was more similar to a 2015 record with a different ID than the 2015 record with the same ID. This scenario did not actually occur. All IDs with a 2014 and 2015 record were linked by the preliminary probabilistic matching.

After checking these two potential issues, the study team concluded about 12,700 of the about 19,600 IDs present in 2014 had a record belonging to the same infant/child in 2015.[C2] Figure C.2 shows the receiver operating characteristic curve for the similarity score. The area under the curve (AUC) is a measure of how accurately the similarity score can differentiate between records that do and do not belong to the same infant/child. An AUC above 0.9 is considered outstanding.[C3] The AUC for figure C.2 is over 0.999. If all pairs of records with a similarity score above 26 were considered to belong to the same infant/child, only 0.03 percent would be improperly matched.

Figure C.2. Receiver Operating Characteristic Curve



[C2] If any infants/children were assigned more than one ID in 2014, there would have been fewer infants/children than IDs present in the 2014 data. It was not necessary to resolve this issue for this analysis; the purpose was merely to observe the similarity score for pairs of records across 2014 and 2015 that did and did not belong to the same infant/child.

[C3] Mandrekar, J. N. (2010). Receiver operating characteristic curve in diagnostic test assessment. Journal of Thoracic Oncology, 5(9), 1315–1316. https://www.sciencedirect.com/science/article/pii/S1556086415306043#bib5

The study team used the results from the initial preliminary probabilistic matches and the manual review process to finalize the two similarity score thresholds. Because the approximately 7,000 pairs of records belonging to different infants/children had a maximum similarity score of 26.336, the study team decided to set the upper threshold at 32. The study team had a high degree of certainty that any records with a similarity score above this upper threshold truly belonged to the same infant/child.

Because the approximately 12,700 pairs of records belonging to the same infant/child had a minimum similarity score of 21.973, the study team decided to set the lower threshold at 20, with a high degree of certainty that any records with a similarity score below this lower threshold truly did not belong to the same infant/child.

The team decided manual review would be required for all pairs of records with similarity scores between 20 and 32. For these cases, the similarity score alone could not be used to determine with a high degree of certainty whether two records belonged to the same infant or child.

C. Apply Probabilistic Matching to Create the Final Longitudinal Data Set

To create the final longitudinal data set, the study team began with the data set already linked deterministically on ID. For each ID, the team searched for records from later dates that could have belonged to the same infant/child (e.g., an ID last observed in January 2015 was compared with all IDs first observed in February 2015 or later). The study team then used the similarity score thresholds to make decisions about which records to link probabilistically.

About five pairs of records with different IDs were linked without manual review; they had similarity scores of 32 or higher. The study team manually reviewed about 150 pairs of records with different IDs and similarity scores ranging from 20 to 32. Of these, about 30 were determined to belong to the same infant/child. Most frequently, these records had slight variations in the spelling of the first or last name.

Among those judged to not belong to the same infant/child, there were several examples of exact match on first and last name but a difference in date of birth. Similarly, there were several matches on last name and date of birth with a clear difference in first name, suggesting the records belonged to different individuals.

To finalize the longitudinal data set, a new variable, NewID, was created. All records belonging to a given infant/child had the same value for NewID, even when the infant/child had more than one ID assigned by the pilot State agency. The final data set had one row per infant/child per week.