A Framework for Linking PRAMS with Administrative Data

Enhancing PRAMS through Data Linkages

Report Summary: This report provides a framework for integrating the Pregnancy Risk Assessment Monitoring System (PRAMS) with administrative data sources. It includes a series of tools and resources that PRAMS teams can use to support their linkage efforts. This Framework can help PRAMS jurisdictions create a robust systematic approach to linkages, leading to an increased understanding of maternal and child health outcomes.

- **Jared Parrish,** Ph.D., MS

  Contractor, *Parrish Analytics and Epidemiology Consulting*
- **Stephany Strahle**, MPH

  Contractor, *ASTHO*
- **Shannon Vance**, MPH

  Assistant Director, *ASTHO*

This work was funded by the Office of the Secretary Patient-Centered Outcomes Research Trust Fund (OS-PCORTF). The OS-PCORTF funding was made available to Centers for Disease Control and Prevention by the Office of the Assistant Secretary for Planning and Evaluation (ASPE) through Interagency Agreement (IAA) 750120PE090052. This report was funded by the Centers for Disease Control and Prevention under the OT18-1802 Cooperative Agreement, Grant #6NU38OT000290 The findings and conclusions in this document are those of the authors and do not necessarily represent the official position of Centers for Disease Control and Prevention, or the other organizations involved, nor does the mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government.

Contributors

This Framework was developed through the contributions of the PRAMS Data Linkage Learning Community and an interagency workgroup convened as part of the learning community. Specific contributions by:

- Brenda Bauman, MSPH
- Maria "Paz" Carlos, PhD, MBA
- Nahida Chakhtoura, MD
- Juanita Chinn, PhD
- Cheryl Clark, DrPH, RHIA
- Shanna Cox, MSPH
- Ann Davis, PhD
- Khaleel Hussaini, PhD
- Russel Kirby, PhD, MS, FACE
- Milton Kotelchuck, PhD, MPH
- Joyce Martin, MPH
- Ekwutosi Okoroh, MD, MPH
- John Prindle, PhD
- Emily Putnam-Hornstein, PhD, MSW
- Holly Shulman, MA
- Shae Sutton, PhD
- Megan Toe, MSW
- Keriann Uesugi, PhD, MPH
- Kristen Zycherman, RN, BSN

Special thanks to the following state health departments for their engagement and input through the PRAMS Data Linkage Learning Community:

- Alaska Department of Health
- Georgia Department of Public Health
- Massachusetts Department of Public Health
- Montana Department of Public Health and Human Services
- Nebraska Department of Health and Human Services
- New Mexico Department of Health
- Rhode Island Department of Health
- South Dakota Department of Health
- Tennessee Department of Health
- Texas Department of State Health Services
- Virginia Department of Health
- Washington State Department of Health

Suggested Citation

Parrish JW, Strahle S, Vance S. A Framework for Linking PRAMS with Administrative Data; 2024. Available online: https://www.astho.org/topic/report/framework-linking-prams-with-administrative-data/

Table of Contents

Tables and Figures

Executive Summary

The Pregnancy Risk Assessment and Monitoring System (PRAMS) is a cornerstone in evaluating maternal and child health issues, with its efficacy further bolstered through linkage with jurisdiction-wide administrative sources. By integrating PRAMS data with these sources, researchers can gain deeper insights into various factors (e.g., social determinants of health) directly impacting maternal and child health outcomes and intricate research inquiries that administrative datasets alone cannot fully address.

Supported by the CDC, ASTHO convened the PRAMS Data Linkage Learning Community, comprising of twelve jurisdiction teams, to conduct PRAMS data linkage projects. From the experiences of these participants, insights from national experts, and a review of published research, a five-phase framework for linking PRAMS data and a suite of supporting tools described below were developed to guide jurisdictions and external researchers in the preparation, execution, and analysis of linked PRAMS data and the establishment of a sustainable linkage environment:

- **Phase I: Linkage Preparation** underscores having a clear and defined PRAMS data linkage purpose to guide the process and create a compelling narrative to foster partner engagement. It involves evaluating internal capacity to perform linkages and securing all necessary data sharing agreements and approvals. Given its critical importance, four tools were developed to aid in linkage preparation:
    - **Data Linkage Readiness Assessment (Appendix B):** A guide for jurisdictions to assess their capacity and infrastructure for conducting PRAMS data linkage before starting a project.
    - **Data Linkage Process Map (Appendix C):** A visual representation of each phase of the framework and its main components.
    - **Data Linkage Process List (Appendix D):** A synthesized overview of the five phases, detailing various processes that may need to be conducted within each phase.
    - **Template Data Use Agreement for Public Health Linkages (Appendix E):** An example of a generalized sharing agreement specific to public health data linkage projects.
- **Phase II: Data Preparation** involves identifying administrative sources to link with PRAMS, conducting exploratory analysis, establishing common identifiers, and harmonizing the sources through cleaning, standardizing, and aligning processes.
- **Phase III: Data Linkage** focuses on selecting appropriate linkage methods and tools based on the project's needs and capacity, setting acceptance thresholds, and conducting validation. Resources such as shell tables are provided to support this phase.
- **Phase IV: Research Dataset Creation and Analysis** covers creating a linked research dataset from a recommended data structure and developing an analysis plan for focused variables and outcomes of interest.
- **Phase V: Sustainability** This final phase covers documenting processes and ensuring the secure storage of linked data for future use and replication.

By following this Framework, public health agencies and researchers can optimize integrated data to address pressing public health challenges and improve health outcomes effectively.

Introduction

Public health practices, programs, and policies designed to improve maternal and child health outcomes can be enhanced using data linkage, a process that combines data from different sources. PRAMS, a population-based survey capturing the lived experiences of persons with a recent live birth, can be integrated with administrative sources (e.g., hospital utilization records, Medicaid claims, and child welfare data) to leverage a broad spectrum of data sources. This framework was informed by the experiences of twelve PRAMS jurisdictions that conducted data linkage projects within an ASTHO learning community, input from an interagency workgroup, and a review of existing peer-reviewed studies using linked PRAMS data. This report presents the framework's rationale, followed by a detailed description of its five phases and a series of tools created to support a systematic approach for PRAMS data linkage efforts.

Why link PRAMS data

Linking PRAMS with statewide administrative data has the potential to significantly improve maternal and child health research, policymaking, and program implementation. By building on the strong foundation of the PRAMS survey, linked data is an efficient way to improve data quality and completeness, facilitate longitudinal analysis, support subpopulation analysis, assess health outcomes and associated risk factors, and support programmatic evaluation. Linked administrative data (e.g., birth records linked with hospital discharges) supports population-based descriptive epidemiology but is limited in providing insights into behavioral, structural, and other elements essential for analytical epidemiological investigations. Augmenting administrative linkages with epidemiological survey data, such as PRAMS, can further identify factors that are amenable to intervention.

The use of the PRAMS sample for population-based linkage projects provides several unique benefits to researchers, including:

1. PRAMS is directly derived from the statewide births. Because PRAMS is weighted to reflect the underlying distribution of the birth records, PRAMS sample-derived estimates can be validated directly with those observed from the full birth certificate.
2. Standardized PRAMS data collection provides an opportunity for comparisons and collaborations between jurisdictions.
3. PRAMS is collected near the time of birth and records a wide variety of indicators during the pre-birth, delivery, and postpartum periods. The timing of data collection and reporting periods facilitates comprehensive investigations on factors associated with multiple maternal, infant, and child health outcomes.

While PRAMS offers many benefits, a few considerations should be made when deciding if linked PRAMS data will be useful for answering a particular research question.

1. Linked PRAMS data is susceptible to imprecision due to small numbers and may not be appropriate for rare outcomes.
2. Any analysis of linked PRAMS data may contain response bias due to the self-reported nature of PRAMS questions.
3. Sampling schema can impact linkage coverage and may result in skewed population estimates.
4. Real-time data linkages can be contraindicated because PRAMS weights are applied to annual

cohorts.
5. PRAMS is limited to population-based assessments and cannot be used for individual tracking.

Several studies have utilized linked PRAMS data to evaluate patient-centered outcomes and program utilization. A few examples of studies using linked PRAMS data include the Enterprise Community Healthy Start (ECHS) program data in Georgia, the Oklahoma Toddler Survey, Medicaid claims in Wisconsin, and child welfare and the Childhood Understanding Behaviors Survey (CUBS) in Alaska. From these linkages, researchers examined a wide range of maternal and child health outcomes, such as child maltreatment, pregnancy intentions, and postpartum care utilization among Medicaid users.

Some jurisdictions currently have or previously had well-established linkage environments integrating maternal and child health data with many other data sources. The Alaska Department of Health established the Alaska Longitudinal Child Abuse and Neglect Linkage Project (ALCANLink) to integrate PRAMS with child welfare data to examine adverse childhood experiences (ACEs) across the state. The Office of Health Informatics at the Wisconsin Department of Health Services created the Linked Birth Outcomes Surveillance System (LBOSS), which integrated vital records, Medicaid claims, inpatient hospital discharge data, and PRAMS and the Division of Maternal and Child Health Research Analysis at the Massachusetts Department of Public Health conducts ongoing linkage of vital records to hospital discharge records as part of their Pregnancy to Early Life Longitudinal Data System (PELL), with potential opportunities to link other sources within this environment.

Linking PRAMS and Clinical Outcomes Data Multi-Jurisdiction Learning Community

With funding from CDC's Division of Reproductive Health and HHS' Office of the Assistant Secretary for Planning and Evaluation, between 2021-2023 ASTHO supported the efforts of twelve jurisdictions in two cohorts of the *Linking PRAMS and Clinical Outcomes Data Multi-Jurisdiction Learning Community* (hereinafter referred to as the PRAMS Data Linkage Learning Community). Cohort 1 included Alaska, New Mexico, and Washington as funded jurisdictions, while Texas participated but did not accept funding. Cohort 2, all of whom accepted funding, included Georgia, Massachusetts, Montana, Nebraska, Rhode Island, South Dakota, Tennessee, and Virginia. Within the PRAMS Data Linkage Learning Community, participating sites embarked on data linkage activities to support Patient-Centered Outcomes Research (PCOR) in maternal and child health. State teams sought to link PRAMS with clinical and administrative datasets and perform analyses for research topics of their interest.

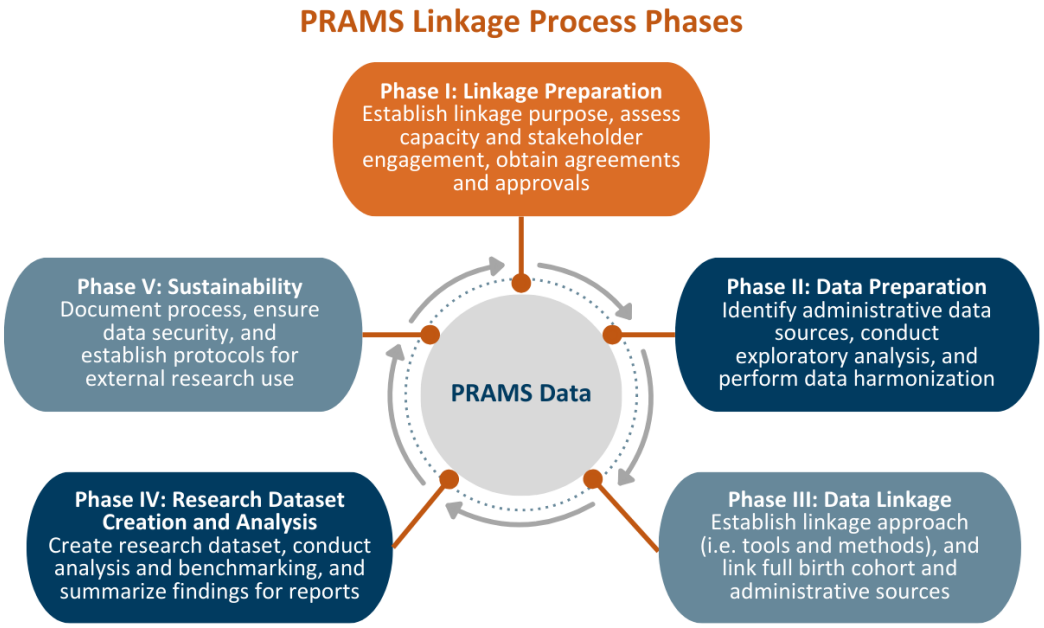
The primary lessons learned, and themes identified during the PRAMS Data Linkage Learning Community largely informed the construction of the Framework. Additional information about the learning community can be found in the PRAMS Data Linkage Learning Community Final Report. High-level project details for each jurisdiction are listed in Appendix A. Jurisdictions applied different approaches to linkage and resolving linkage issues. Each jurisdiction also had varying levels of knowledge, capacity, existing partnerships, and sources to be linked. This Framework presents strategies for addressing common issues related to designing and implementing a linkage approach with PRAMS data tailored to the project's purpose.

Linkage Framework

Data linkage projects are essential for integrating and analyzing data from various sources to inform public health policy, research, and decision-making. These projects enable researchers and policymakers to examine population health trends, track disease outbreaks, and evaluate the effectiveness of public health interventions. Multiple frameworks and toolkits[1] have been developed to support and standardize linkage efforts. For decades, the National Center for Health Statistics has linked survey data with administrative sources (e.g., Medicaid, National Death Index) to support an increased ability to analyze and explore complex questions that impact health.

This report outlines a comprehensive five-phase framework (Figure 1), associated tools, and additional resources for PRAMS data linkage projects. It addresses some of the unique challenges to consider when linking complex survey-weighted data and the benefits of linking a survey dataset based on a well-defined population (births). By following this Framework, health agencies and researchers can optimize integrated data to address pressing public health challenges and improve health outcomes effectively.

Figure 1. Synthesized visual of PRAMS Linkage Process Phases.



- [1] CSTE Injury Data Linkage Toolkit
- CSTE Tribal Epidemiology Toolkit
- Race and Ethnicity Data Improvement Toolkit
- NAHDO Data Enhancement and Linkage
- Dasy Center Data Linking Toolkit
- Quality and Complexity Measures for Data Linkage and Deduplication
- Linking Data for Health Services Research: A Framework and Instructional Guide
- The Linkage of the National Center for Health Statistics (NCHS) Survey Data to U.S. Department of Housing and Urban Development (HUD) Administrative Data: Linkage Methodology and Analytic Considerations

PHASE I: Linkage Preparation

The Linkage Preparation phase is pivotal for the success of all data linkage projects. Investing time upfront to discover, plan, and clarify the purpose of the linkage establishes the project's foundation and determines if linkages are necessary, feasible, and appropriate. Jurisdictions intending to link PRAMS data should start by defining a clear purpose and collaborating with community partners who stand to benefit or have an interest in the research findings. Once the purpose is established, a structured project plan should be developed to address the technical infrastructure and expertise needed for successful project completion. The Data Linkage Readiness Assessment (Appendix B) can aid in this preparation. This assessment tool consists of ten questions designed to help state and other jurisdictional PRAMS teams evaluate their readiness for a linkage project. The questions focus on existing infrastructure, capacity, leadership support, expertise, and data access. Based on the responses, users can score overall readiness and identify areas that may need development. The assessment underscores critical components of successful linkage projects, offering a way to evaluate these components, identify potential gaps, and address areas needing attention before initiating data linkage.

Establishing a Purpose

One of the first steps in setting up a linkage project is clearly defining its purpose and articulating what public health issue it will address, usually in the form of a research question(s). Teams should consider what specific questions or challenges the project aims to address and how the linkage (and information derived from it) will benefit society, public health, or specific communities. Most successful linkage projects start with one or two specific, narrowly defined research questions and then build upon the lessons learned to establish a robust linkage infrastructure. When a comprehensive linked data infrastructure doesn't already exist (such as the PELL project in Massachusetts, Wisconsin's LBOSS, and Alaska's ALCANLink), starting with specific research questions for the initial pilot project offers three key benefits:

- It encourages using the linkages as a means to an end as opposed to the end itself. Large linkage projects can get bogged down quickly, taking years to establish a strong infrastructure. A pilot project will help establish the utility of the linked data.
- Starting with a focused research question for a pilot project can reduce the technical burden, facilitate identifying and addressing linkage errors and challenges, and work out efficiencies to support establishing a robust infrastructure and ongoing linkages.
- A focused pilot project will help engage partners and communicate results from the linked data in a timely manner. Sharing results from the linked data will encourage hypothesis generation, expanded linkages, and new collaborations.

Table 1 provides examples of general purposes for PRAMS data linkages and corresponding research questions that could help focus a pilot project. Once teams establish a purpose, they should determine if a linkage is necessary and how or if the PRAMS sampling design may affect the ability to explore the research question or purpose (e.g., studying the impact of home visiting on infant health if not universally provided may not provide a large enough sample size).

Table 1. Examples of PRAMS linkage purposes and research questions.

Examples of General Purposes for PRAMS Data Linkage	Sample Research Questions for PRAMS Data Linkage
Understand factors influencing maternal and child health trends	• How does access to prenatal healthcare impact child hospitalization trends within one year of delivery? • What are the factors that influence trends in child immunization rates?
Identify maternal and child health populations experiencing adverse clinical health outcomes	• What are the rates and risk factors of gestational diabetes among different socioeconomic and racial/ethnic groups? • How do pre-birth familial stressors impact the frequency and severity of emergency department use among children within two years of birth?
Improve PRAMS data completeness and accuracy	• What is the accuracy of self-reported receipt of the flu vaccine during the 12 months before delivery? • What is the accuracy of self-reported WIC program participation?
Enhance maternal and child health program evaluation	• What is the impact of participation in Parents as Teachers (PAT) on child maltreatment incidence? • What is the impact of WIC program participation on early child education performance?

Ultimately, the design of the linkage project will be influenced by its breadth and scope. Appendix C provides a visual representation of the data linkage process, while Appendix D offers a detailed list of steps that may be required for each phase. These tools are valuable for considering the project's scope.

Initiating a linkage project around a specific research question as a proof-of-concept or pilot project can help establish the utility of the linkage. Once the utility is demonstrated, expanding the project's purpose can be advantageous, allowing for multiple research questions to be addressed using a single linked data source.

Establishing a Narrative and Forming Relationships

Establishing a compelling narrative around the purpose and goals of data linkage is essential for gaining support and building relationships with partners. This is especially true for external partners such as those within clinical, hospital, social service, or law enforcement settings. A well-crafted narrative helps convey the importance, benefits, and ethical considerations of data linkage while engaging and

educating your target audience. The narrative should center on the established purpose and goals of the project.

Forming relationships with internal partners from different departments or teams involves identifying the data steward and initiating open communication. During these initial conversations, it is crucial to understand their goals and constraints and to demonstrate the benefits of collaboration. Usually, it is best to start by scheduling a meeting with the data steward, any authorizer or approver, and, if necessary, an information technology team member to discuss project objectives, share expertise, and establish mutual trust. Emphasize the potential value of the partnership in improving data accuracy, efficiency, and, ultimately, public health outcomes. Foster ongoing collaboration through regular updates, feedback sessions, and clear delineation of roles and responsibilities, ensuring a harmonious and productive working relationship.

When forming relationships with partners outside of public health (e.g., education, child welfare, transportation), it helps to start by identifying key individuals within the agency who clearly understand their data structure, legal requirements, and approval process. This usually begins by reaching out to a connection within the agency and clearly explaining the project's objectives and potential benefits for the organization. It is best to start small by identifying common goals and areas of mutual interest, emphasizing how data linkage can enhance their work and outcomes. Schedule follow-up meetings to discuss collaboration opportunities, listen to their needs and concerns, and adapt the project plan to accommodate their requirements. Reducing burden and maximizing potential is key.

As indicated above, with a PRAMS Data Linkage project, it is helpful to first conduct a proof-of-concept or pilot project instead of trying to link multiple data systems. During this initial project, trust can be built, and the utility of future linkages can be demonstrated. Keeping the initial linkage small, with only a few elements, keeps the project focused and can reduce hesitation in participation. Establishing trust and understanding a partner’s organizational culture critical, along with maintaining open communication channels throughout the partnership to address any challenges and ensure successful collaboration.

Forming relationships with external partners for a data linkage project requires careful navigation of legal and privacy considerations, in addition to building trust and mutual understanding. Establishing formal agreements or Memoranda of Understanding (MOUs) regarding data sharing protocols, confidentiality, and security measures is essential to address legal and privacy concerns. The data use agreement template in Appendix E can provide guidance on some essential components that should be included in the agreements. Regular communication and transparency regarding data usage, protection, and project progress are vital for maintaining trust and ensuring compliance with regulatory requirements. Additionally, fostering personal relationships through face-to-face meetings or virtual interactions can help build rapport and facilitate effective collaboration over time.

Throughout project development, it is important to ensure that external partners are included when appropriate. Tribal entities and epidemiology centers should be engaged to ensure that their perspective, needs, and/or issues can be addressed and included during project development. Other key partners may be topic or content area experts, individuals with lived experience, medical providers,

Key considerations when linking with PRAMS

- PRAMS is annually weighted, represents a portion of the population, and does not require real-time data integration.
- Linking PRAMS with large data integration projects should be approached with a clear purpose and always retain the full PRAMS respondent cohort to ensure weights are applied correctly.
- PRAMS operates under IRB approval, with participants providing implicit consent when responding. Be sure the linkage of the PRAMS respondents is consistent with the consenting procedure and/or has received further IRB approval.
- Be aware of other birth record linkage projects that may be leveraged (e.g., if the birth records have been previously linked with Medicaid, hospital discharges, or other sources).
- Rare outcomes or low-occurrence events captured in a data source can make it challenging to identify enough cases in the PRAMS sample for meaningful analysis.
- Sources that only represent a portion of the birth population may need to be reweighted if they are not representative of the entire birth population from which the weights were estimated and may only measure outcomes on a portion of the population.
- Sources/outcomes related to any of the sampling strata may result in either over- or under-detection from what is observed.

and/or non-governmental agencies working on addressing aligned topics. Including partners during the development phase helps strengthen relationships and improves the likelihood that the data will be used, meaningful, and relevant to those addressing a particular health issue.

While developing relationships may come naturally to some, it can be challenging for others. Make sure to formalize the relationship whenever you can to ensure the project's long-term success. Informal agreements built on existing relationships may help initiate projects but are problematic with staff turnover.

Data Infrastructure

Existing data infrastructure, available tools, and dedicated resources are crucial to the success of the linkage project. The recommended structure for linked PRAMS data is a centralized relational data repository. The PRAMS respondent data must first be merged with the birth records to obtain identifiers to facilitate linkages. To ensure patient confidentiality and data protection, these merged PRAMS data should adhere to established data governance policies and comply with relevant state and federal regulations, such as HIPAA for covered entities. It is essential for teams to thoroughly understand the limitations and strengths of their existing infrastructure before initiating a linkage project.

The PRAMS-linked infrastructure should also include data integration and harmonization mechanisms, which involve cleaning, standardizing, and aligning different datasets to ensure they can be combined accurately. Administrative data to be linked with PRAMS may come from various sources with differing formats, making robust tools for data cleaning, transformation, and normalization—structuring data into consistent formats—essential. The data infrastructure should provide secure and controlled access to authorized users, implementing role-based access control mechanisms to regulate data access based on user roles and permissions. To ensure data quality and integrity, ongoing monitoring and evaluation processes should be established, including regular audits, data validation checks, and quality assurance procedures to identify and address any issues or discrepancies in the data.

It is important to note that jurisdictions are likely engaged in various data linkage and modernization activities, with differing types of infrastructures that can either support or hinder data linkage projects. Some jurisdictions may have a "data lake" or other centralized data repository that facilitates resource connections. Others may operate with a centralized team performing linkages within a federated system, or they may have completely siloed systems with limited connections. Understanding the data governance, required approvals, information-sharing processes, and technical resource requirements specific to each jurisdiction is essential before embarking on a linkage project.

Identifying Administrative Source(s) for Linkage with PRAMS

It is highly recommended to start by linking a single administrative source (e.g., hospital discharge records) with PRAMS. By starting with a single source, the process of creating metadata, establishing linkages, validation, and use of the linked resource will be simplified. A pilot project with a single source will also enable researchers to establish the best solution for scaling in the future.

Various types of data with different structures may be encountered. For example, hospital discharge or Medicaid data may be used to obtain information on hospital stays, diagnoses, procedures, or immunization registries to obtain data on vaccination status. Common sources linked with PRAMS via the birth record include vital death records, child welfare, education records, hospital discharges, census, and emergency records, Medicaid, and social service records such as WIC. Each source should be assessed to clearly understand its structure before integration. The following questions may help jurisdictions understand the data structure of a source to inform their data linkage strategy:

- Are the source records event- and/or individual-based?
- Does the source contain a unique identifier for individuals?
- Does it contain multiple rows (records) per individual?
- Does it have identifiers that can be used to link with the PRAMS/birth file?
- Is the data longitudinal?
- Does the source have legal restrictions prohibiting/restricting linkages?

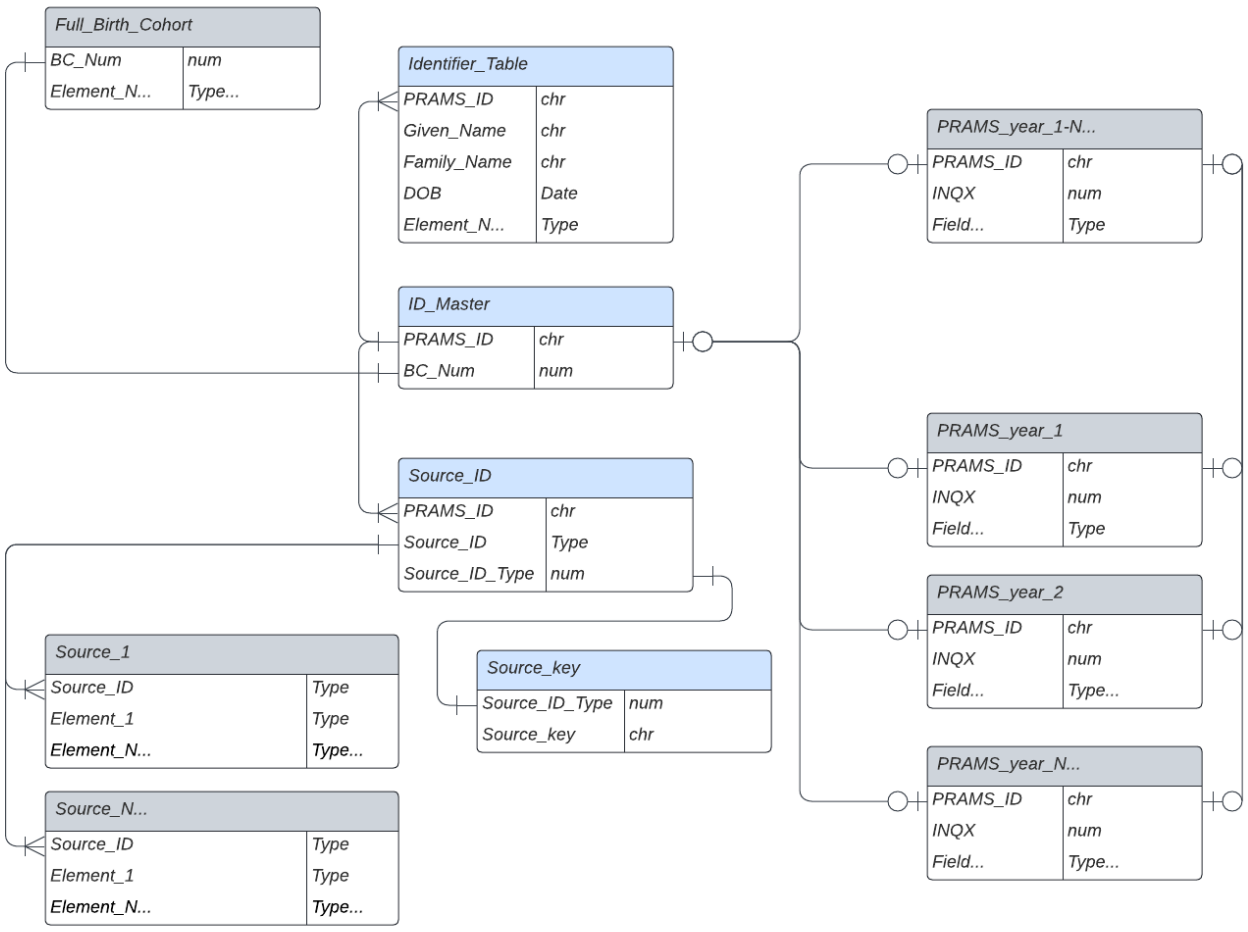
Important Note: When linking records, especially over time, non-linkage misclassification due to emigration and migration could bias estimates derived from linked PRAMS data. For example, linking the PRAMS child with child welfare records over time would not capture maltreatment events that occur when a child moves out of the jurisdiction. Misclassification can also occur when individuals move in or out of the data source's system during the observation period, such as those with changes in health insurance status or discontinuation of participation in programs such as Healthy Start.

Furthermore, birthing parents or other partners listed on birth records linked with historical administrative records (pre-birth event) may exhibit different linkage rates among subpopulations with varying migration patterns, potentially resulting in undetected individuals in administrative records. For example, linking PRAMS mothers with juvenile justice records would exclusively capture those mothers who were juveniles within the jurisdiction of the PRAMS survey.

Recommended Data Structure

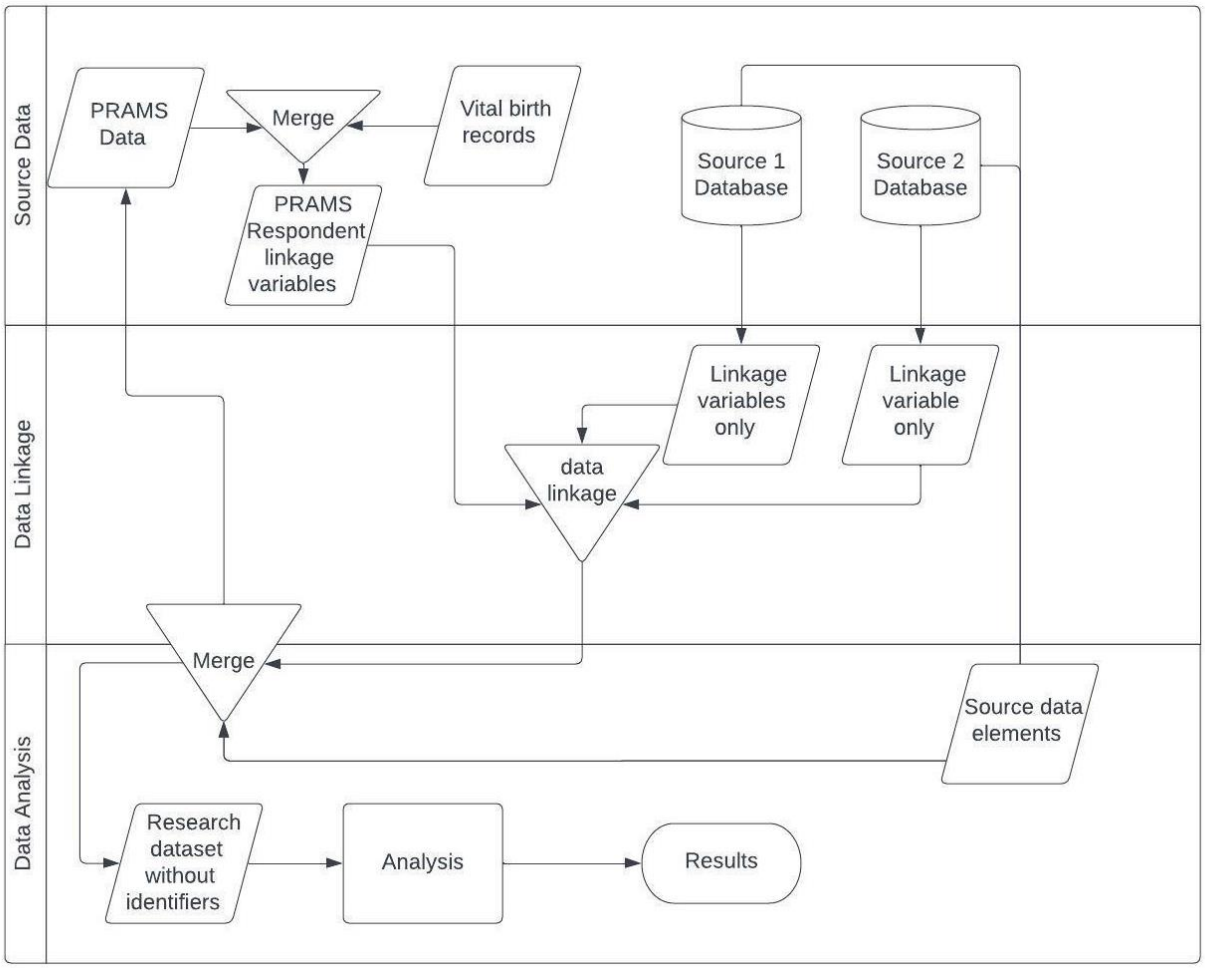
It is important to develop a plan for the final data structure that will be used to store the linked data before beginning. We recommend a simple relational data structure (Figure 2), which balances normalization and utility and provides flexibility for future and ongoing linkages and different research questions. This structure is adaptable for one-time data linkage between PRAMS and a single source with one record per individual (e.g., mortality records) to more complicated linkages (e.g., linking birthing parents with all hospitalizations and emergency room visits for the year preceding and 5 years after the birth event). Regardless of whether the recommended data structure is used, teams should plan the final data structure, visualize it with a data map (often hand-drawn to start) to depict how the sources will be connected and stored (Figure 3), and document the process.

Figure 2. Recommended data structure for PRAMS data linkages with administrative data sources.



The recommended data structure connects each PRAMS phase to the ID_Master table. A combined PRAMS multi-year data set (PRAMS_year_1-N…) should also be created to facilitate efficient future analyses. This is especially useful when multiple PRAMS phases are used. The ID_Master table is also connected to a full birth cohort file and each of the administrative sources through the Source_ID table. Having a Source_ID table enables multiple IDs from a single source and multiple IDs from different sources. It reduces redundancy and the need to store repeated information that would be required in a single table.

Figure 3. Example data map depicting how the data are connected and stored.



Agreements, Approvals, and IRB

Each jurisdiction will have unique processes that determine the types of agreements (if any) that need to be signed and approved. Typically, at a minimum, a data use agreement is required. Appendix E is a data use agreement template that jurisdictions can use. Although most jurisdictions have standard language and formats for agreements, this template may still help ensure that all critical elements are included. Having a formalized agreement helps institutionalize the project, describe expectations and responsibilities, provide legal protection for data sharing, and, among other things, pave the way for future linkage projects. When establishing these agreements, an often-overlooked part is future use, reporting, and data sharing of the linked data resource created. These components should be clearly defined and documented (even when an agreement is not required). The process of establishing agreements should not be underestimated and can take multiple months to complete. It is important to note that depending on the source, different agreements and approvals may be required.

A brief description of common agreements or other documents that may be relevant to the linkage project is outlined below.

1. Data Use Agreements (DUAs): DUAs are legal agreements that specify the terms and conditions under which PRAMS data can be accessed, used, and shared. These agreements typically outline data security measures, confidentiality requirements, permitted uses of the data, and restrictions on data sharing.

2. Institutional Review Board (IRB) Approval: IRB approval is often required for research involving human subjects, including data linkage projects with PRAMS. Researchers must submit their study protocols to the IRB for review to ensure that the study meets ethical guidelines and safeguards participants' rights and welfare.

3. Data Sharing Agreements: Data sharing agreements govern the sharing of PRAMS data with external partners or collaborators. These agreements define the terms of data sharing, including data security protocols, restrictions on data use, and liability provisions.

4. HIPAA Authorization: PRAMS response data combined with identifiers from the birth records may be considered individually identifiable health information by some health departments that are covered entities. A Health Insurance Portability and Accountability Act (HIPAA) authorization or exemption may be required for linked data use and disclosure for covered entities.

5. Confidentiality Agreements: Confidentiality agreements may be necessary when sharing PRAMS-linked data with external entities or individuals. These agreements outline obligations to protect the confidentiality and security of the data and may include provisions for data destruction or return after the project is completed.

6. Data Security Plans: Data security plans outline measures to protect PRAMS data from unauthorized access, disclosure, or misuse. These plans should address physical, technical, and administrative safeguards, such as encryption, access controls, and employee training, to ensure data security throughout the project lifecycle.

7. Consent Forms: If data linkage between PRAMS and other datasets is not a function already covered in current consent procedures, additional consent from the participants may be required to proceed with a data linkage project. The reviewing IRB or other scientific review committee can provide guidance on whether this is required. Typically, most IRBs will waive this requirement when it is not feasible to obtain consent and the data is being used for research or public health purposes.

Documenting Study Protocol/Plan

A documented study protocol or plan will be your roadmap to success and a way to keep your project focused. This document should summarize all the key steps during your planning phase and serve as the foundation for the project documentation to expand as the project develops. Appendix F outlines a

general template for the recommended minimal project plan elements. This outline includes the key sections with short descriptions about what information to include, example figures, and shell tables.

Recommended minimal documentation:

- Short background that outlines the reasoning for the PRAMS data linkage project.
- A clear purpose statement that describes the overarching goal of the linkage project.
- Short, concise objectives in the form of research questions. The more specific the objectives, the better.
- Description of the data infrastructure that articulates existing or prior linkage work, the population that will be included, years of data, data sources to be linked, and planned data structure of the linked data.
- Short overview of the existing or needed agreements or approvals needed.
- A brief description of the linkage plan that outlines the general process, tools and methods to be used, and anticipated validation procedures.
- A brief analysis plan and description of key products or methods for communicating results.
- Short description of how the project will be sustained.

The analysis and sustainability plans will likely be vague in the initial phases, but as the project develops this protocol document should be updated and refined.

The Linkage Preparation phase is designed to aid jurisdictions through the critical steps and considerations necessary for establishing a strong foundation for their PRAMS linkage project to build from. By defining a clear project purpose, engaging key partners, assessing their data infrastructure, and constructing a documented study plan before linking, jurisdictions can use these steps to reduce the risk of unintended challenges later in the process. Using a guided tool such as the Data Linkage Readiness Assessment (Appendix B) encourages jurisdictions to make careful considerations that could impact the feasibility and quality of a linkage project.

PHASE II: Data Preparation

The quality of the data to be linked will directly impact linkage results. It is therefore critical to harmonize the data sources (i.e., applying the same data cleaning/standardization rules across all data sets). This step involves assessing missingness in identifiers, removing invalid data, and aligning string types, date formats, special characters, and other elements. The amount of data cleaning before conducting the linkage will depend on the data, what PRAMS phases are included, and any changes that have occurred with state vital records data. The majority of the time spent linking data will occur in this data preparation stage to ensure high-quality linkages and analyses.

Pre-Linkage Assessment

PRAMS and birth records

To obtain identifiers and facilitate validation assessments of the linked data, it is recommended to merge the entire PRAMS sample (INQX == 0 and 1) with statewide birth records. In this case, the PRAMS sample can be merged with the entire statewide records using the state file or birth certificate number.

Outlined in Figure 4, the ideal process is to use statewide birth records (limited to PRAMS residency inclusion criteria) to link with other administrative data sources to support validation efforts of PRAMS weighted estimates. However, it may not be feasible to link all birth records for every year (e.g., the size of the birth population may be too large based on capacity). In this case, it is highly recommended to link at least one annual statewide birth cohort for a PRAMS year as outlined in Figure 4, and then link all subsequent PRAMS years as outlined in Figure 5. Regardless of approach, the steps outlined below describe how to maximize efficiency and security. Each jurisdiction, however, may need to implement individual adaptations.

Example: Let's consider a jurisdiction that between 2009-2011 has 10,000 births each year meeting eligibility criteria (N = 30,000 births for all three years). For each year, 1,000 PRAMS responses are received (n = 3,000 for all three years). The jurisdiction links the 2009-2011 PRAMS respondents with hospital discharge records but only has the capacity to link the entire birth cohort for a single year. To assess the linkages, the jurisdiction linked the entire birth cohort for 2010 (the middle year of the three cohorts), but only the PRAMS respondents in 2009 and 2011. The hypothetical results of the linkages are outlined below:

2009 Cohort: 975 PRAMS respondents linked to a hospital discharge recorded, representing a weighted % of 98.7 (95%CI 97.9, 99.5) births.

2010 Cohort: 9,900 (99.0%) births linked to a hospital discharge record. 980 PRAMS respondents linked to a hospital discharge record, representing a weighted % of 98.4 (95%CI 97.6, 99.2).

2011 Cohort: 969 PRAMS respondents linked to a hospital discharge record, representing a weighted % of 97.5 (95%CI 96.6, 98.4).

General approach for preparing PRAMS records for linkage

- Determine what PRAMS years are going to be used.
  - If the selected PRAMS years include multiple PRAMS phases, develop a crosswalk of the variables if one does not already exist.

- Determine if the infant, birthing parent, non-birthing parent, or multiple individuals will be linked.
- Create a PRAMS data set limited to the following data elements:
    - inqx: (INQX is a dichotomized variable (0,1), where 1 represents those who responded to the PRAMS survey, and 0, those that did not)
    - nest_yr: (This variable represents the PRAMS sample year and is equivalent to the birth year)
    - prams_id: (This is the unique PRAMS ID)
    - birth_certificate_number: (The state file number of the sampled infant)

  (Note: the PRAMS ID and birth certificate number may be labeled differently in each jurisdiction; for example, ID and BC may be used.)

- Create a limited Birth Record file for the PRAMS years selected.
    - Include all statewide births consistent with the PRAMS residence/location eligibility requirements.
    - Limit to data elements to facilitate linkages. Each jurisdiction may have different variable names and elements available. The list below is an example:
        - birth_certificate_number
        - infant_sex
        - infant_birth_weight
        - infant_date_of_birth
        - infant_first_name
        - infant_middle_name
        - infant_last_name
        - infant_ssn
        - maternal_first_name
        - maternal_middle_name
        - maternal_last_name
        - maternal_maiden_name
        - maternal_residence_city
        - maternal_residence_zip
        - maternal_race
        - maternal_ssn
        - plurality (multiple births)

      Note: If the years selected cover a year where there was a change in how the identifiers were collected on the birth record (e.g., rather than allowing for respondents to choose only one option for race, the form now allows for respondents to select multiple races), then the birth record elements will need to be standardized. If linking the non-birthing parent, those elements will need to be included.

    - Right join the PRAMS limited file to the limited birth record file. Make sure you retain all records from the birth file and those that merge from the PRAMS file.
        - Verify all PRAMS records merged. Although uncommon, an infant may occasionally be assigned multiple state file numbers or be incorrectly sampled (e.g., actually a stillbirth). In these instances, they may not merge with an

updated multi-year birth file. Manually review any PRAMS records that do not merge and reconcile.

- Check for duplicates and reconcile.

Figure 4. Linkage based on full birth cohort.

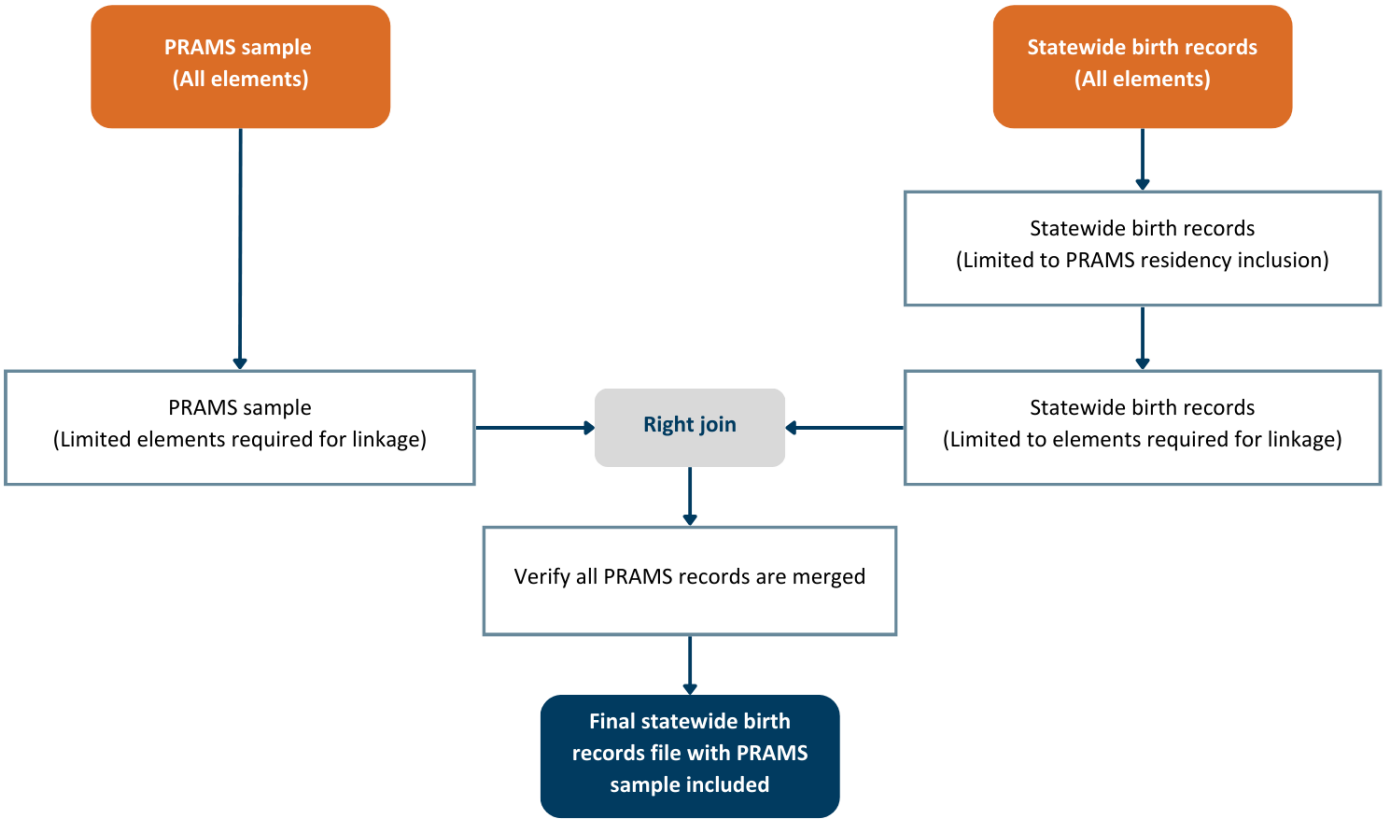
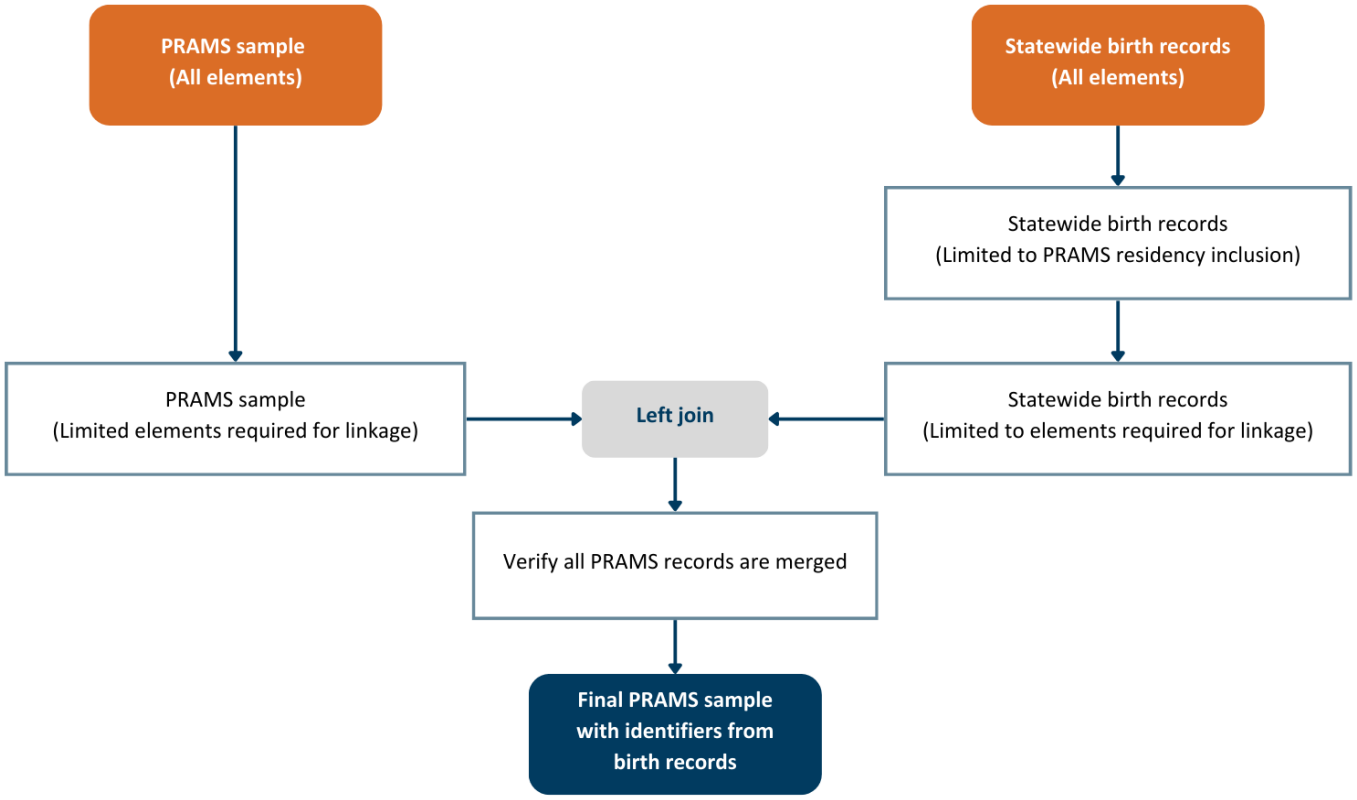


Figure 5. Linkage limited to PRAMS sample.



After the PRAMS sample is merged with the birth records, the next step is to evaluate the final file to be used for linkage. This begins with a detailed examination of the identifiers intended for use in the linkage process.

Names: If names are chosen as identifiers, it is essential to understand the variations in string structures and differences by race and ethnicity. For example, one might need to understand how names with multiple middle or last names are formatted (e.g., Mary Ann Elizabeth Smith), or if suffixes such as Jr. or III are included (e.g., John Smith Jr.). Naming traditions among different racial/ethnic groups may result in long names (long character length) that could be subject to character truncation, inclusion of special characters (e.g., accent symbols, apostrophes, hyphens, etc.), or use of phonetic or written spellings. Naming conventions might also present the family surname first or include multiple surnames. Being familiar with the demographic composition of the datasets is key in navigating any missed linkages resulting from algorithms not recognizing these variations in naming conventions.

Dates: When dates are utilized as identifiers, it is crucial to ensure they are consistently formatted across datasets. For instance, dates should follow a standardized format (e.g., MM/DD/YYYY or YYYY-MM-DD), and any variations or inconsistencies must be addressed during the linkage process. For example, the dataset may have month, day, and year as three separate fields that may be combined into a single standardized date format.

Special characters and string restrictions: Attention should be given to any special characters or string restrictions present in the identifiers. For instance, certain datasets may have limitations on character width, resulting in truncation for long names or other identifiers. An example could involve identifying and addressing any truncation issues that may occur due to character restrictions, ensuring no loss of information during the linkage process.

Data type and units: Each variable used for linkage should be formatted as the same data type (i.e., numeric, integer, factor, character, etc.). For example, if social security numbers are used, they may be stored as character or numeric values and should be standardized across sources to be linked. Variables such as birth weight may be presented in grams or pounds and will need to be standardized to ensure correct linkages are made.

By meticulously reviewing and understanding the nuances of identifiers such as names, dates, and any special characters or string restrictions, researchers can ensure the accuracy and reliability of the linkage process between PRAMS and birth records.

Data Harmonization

Data harmonization involves standardizing, cleaning, and aligning data from disparate sources to ensure consistency, compatibility, and interoperability. Datasets collected from different sources often have unique formats, terminology, and quality; data harmonization begins by addressing these differences. One aspect involves standardizing variables across datasets, ensuring that units of measurement, date formats, and coding schemes are consistent. For instance, dates may need to be transformed into a standardized format to enable accurate comparison.

Inconsistencies in the data, such as missing values, outliers, or conflicting information, need to be resolved using data cleaning techniques such as imputation or validation procedures to enhance data quality and reliability. By

Considerations for multiple births or adoptions

Multiple births often result in high similarity matches with siblings. By including indicators of plurality and birth order, improved linkages can be conducted. It is important to consider the impact of the PRAMS sampling design as well. If your PRAMS program over samples on low birth weight, then it is plausible that a higher-than-expected proportion of multiple births will be included. Thus, having a strategy to support linkages will be helpful. Some jurisdictions have found success with taking an iterative approach with twins. This approach usually first identifies twin births from the birth record and links these records using exact deterministic matching, followed by probabilistic linkages with emphasis on the first name. Remaining unlinked cases then undergo manual review or an applied rule-based strategy to classify linkages.

In many jurisdictions, adoption records are “sealed” and unavailable to the researcher. If possible, include a flag for adoptions. Another strategy is to compare the original birth file used for the sampling with a current birth file and include any name changes associated with the state file number.

Using iterative strategies that include and exclude different partial identifiers are typically applied, with a check for duplicates followed by a comparison with prior matches conducted. Often projects start with multiple partial identifiers and then step out identifiers to increase the sensitivity. This process is usually more involved with the first few linkages until the process is refined.

normalizing the data, the process of organizing data in a database to improve data integrity and reduce redundancy, you transform it to a common scale or distribution, facilitating comparison and analysis across datasets. This normalization process might involve scaling numerical variables or transforming categorical ones.

In cases where unique identifiers are necessary for linkage, data harmonization involves creating common identifiers that uniquely identify records across datasets, ensuring accurate linkage between related records from different sources. Furthermore, aligning the structures and schemas of datasets is essential, which may include matching field names, ensuring consistent data types, or reshaping data to a common format for seamless integration and analysis.

Documenting the harmonization process is key for transparency and reproducibility and it should include details of data transformations, standardization procedures, and any assumptions or decisions made during the harmonization process. Additionally, metadata describing the characteristics and lineage of the data are important for understanding and interpreting the harmonized dataset. Data harmonization is a fundamental step in the data linkage process. By standardizing, cleaning, and aligning data, the dataset overcomes heterogeneity and promotes interoperability across disparate datasets, enabling meaningful analysis and decision-making.

PHASE III: Data Linkage

Generating matches generally involves deterministic, probabilistic, or machine learning techniques. These techniques work by evaluating comparison vectors for each record pair to identify similarities or differences between pairs and then determine a match based on a set threshold. Record linkage mechanics have been described in detail elsewhere and are beyond the scope of the Framework.

A non-exhaustive list of helpful resources describing linkage details is provided below. Additional resources can be found in Appendix G.

- CSTE Injury Data Linkage Toolkit
- Quality and Complexity Measures for Data Linkage and Deduplication
- Linking Data for Health Services Research: A Framework and Instructional Guide
- The Linkage of the National Center for Health Statistics (NCHS) Survey Data to U.S. Department of Housing and Urban Development (HUD) Administrative Data: Linkage Methodology and Analytic Considerations

Training and Staff Resources

Successfully linking data most often requires a team of individuals with varied knowledge and skill sets. Finding a single individual with all the skills to complete the entire linkage project may be difficult. A foundation of the skills and knowledge needed include:

1. Strong data management skills and ability to handle large datasets, clean the data to ensure its quality, and transform it for analysis;
2. A solid understanding of statistical analysis and ability to apply appropriate methods and techniques to discern patterns within the data;
3. Understanding of linkage methods and the ability to review the data to select appropriate deterministic and probabilistic techniques and quantify the error based on the choice.
4. Strong background in Python, R, SAS, or other programming languages to implement linkage algorithms, conduct analysis, and automate repetitive tasks, and familiarity with machine learning. Even if using a specific data linkage tool (e.g., Link Plus), this ability adds a powerful dimension to data linkage efforts, especially when tackling intricate or large datasets or unstructured data sources.
5. Domain expertise and understanding of the population structure and underlying demographics are required to guide the analysts in defining linkage criteria, selecting pertinent variables, and interpreting linkage results within the context of the population and subject matter.
6. Awareness of data privacy and ethics, ensuring compliance with regulations, and safeguarding the confidentiality of individual data.

With critical thinking and problem-solving abilities, linkage teams can navigate challenges inherent in the linkage process, while effective communication skills enable them to collaborate seamlessly with team members, stakeholders, and external partners. Even when a single individual is capable of conducting all parts of the linkage process, it is important for the team to be familiar with the process and have shared knowledge. This will ensure that the project's success does not hinge solely on a single individual.

Linkage methods overview

Choosing the appropriate data linkage method depends on various factors, including the characteristics of the data to be linked with PRAMS, the linkage task at hand, available resources, and the desired level of accuracy. The process of identifying and selecting the best linkage method requires a clear understanding of the data structure and quality, types of identifiers available, and size of the datasets to estimate the computational resources required. The primary methods used are broadly described below. For a list of what jurisdictions in the PRAMS Data Linkage Learning Community used, please see Appendix A.

As has been described previously, typically the process of identifying a method or combination of methods involves an iterative process. Initially, researchers define the research question and assess the quality and compatibility of the datasets to be linked. Subsequently, they explore various methods, including deterministic rule-based, probabilistic, and machine-learning approaches. Deterministic linkage relies on predefined rules or criteria to match records based on exact or highly similar identifiers, such as names and dates of birth. Probabilistic linkage, on the other hand, assigns weights to different variables and calculates the likelihood of a match based on the similarity of these variables across records. Machine learning methods leverage algorithms to automatically learn patterns and relationships in the data, enabling more flexible and adaptive linkage strategies. Pilot testing is often conducted to evaluate the performance of these methods, refining and iteratively improving them based on the results. Validation ensures the accuracy and reliability of the chosen linkage method(s), with documentation and reporting providing transparency and accountability in the process. Continuous monitoring and evaluation facilitate ongoing improvement and optimization of the linkage process, ensuring the integrity of linked data for research, analysis, and decision-making purposes.

Deterministic (rule-based methods)

Deterministic record linkage is a method of linking records from different datasets based on exact matches or predefined rules without considering the probability of a match. Unlike probabilistic linkage, which assigns weights to matching variables and calculates the likelihood of two records belonging to the same individual, deterministic linkage relies on specific criteria to identify matches.

One common approach in deterministic linkage is exact matching, where records are linked if they have identical values in one or more key fields, such as a unique identifier like social security number or birth certificate number. This method is straightforward and efficient but requires high-quality and consistent data across datasets.

Another approach is rule-based deterministic linkage, where matching rules are predefined based on domain knowledge or specific criteria. For example, records might be linked if they have similar values in certain fields, such as names and addresses, allowing for variations due to typographical errors or differences in formatting.

Deterministic linkage methods are often used in scenarios where exact matches or specific criteria can reliably identify true matches between records, such as merging datasets with unique identifiers or linking records within a single database. While deterministic linkage can be faster and simpler than

probabilistic methods, it may overlook potential matches that do not meet the predefined criteria, leading to high specificity but lower sensitivity and potentially missing important relationships between records.

Probabilistic

Probabilistic record linkage is a method for linking records from different datasets by assessing the likelihood or probability that two records belong to the same individual or entity. Unlike deterministic linkage, which relies on exact matches or predefined rules, probabilistic linkage considers the similarity between records across multiple attributes and calculates the probability of a match based on these similarities. Similarities between attributes may use edit distance approaches (e.g., Levenshtein, Q-grams, JaroWinkler, SOUNDEX, date and address specific metrics). These metrics that identify similarities are then used to calculate match probabilities.

One of the most widely used probabilistic linkage methods is the Fellegi-Sunter method. This approach assigns weights to different matching variables based on their discriminatory power and calculates a probabilistic match score for each pair of records. The match score represents the likelihood that the records belong to the same entity, considering similarities and differences across multiple attributes. Thresholds can then be applied to determine which pairs of records are considered matches.

Probabilistic linkage methods are particularly useful when dealing with datasets that contain errors, inconsistencies, or missing information, and only have partial identifiers (e.g., name, date of birth, sex, age, race). By considering the probabilities of matches rather than relying solely on exact matches, probabilistic linkage can handle variations in data quality and account for potential errors or discrepancies. Additionally, probabilistic linkage allows for flexibility in the matching process, as weights can be adjusted and thresholds can be tailored to specific requirements or characteristics of the data. Overall, probabilistic record linkage offers a powerful and flexible approach to linking records from different datasets, enabling researchers to merge data accurately and effectively across diverse sources while accounting for uncertainties and variations in the data.

Machine Learning

Machine Learning Record Linkage (MLRL) is a modern and evolving approach to linking records from different datasets using machine learning techniques. Unlike traditional deterministic or probabilistic linkage methods, MLRL leverages algorithms and models to automate the linkage process, handle large-scale datasets efficiently, and reduce manual reviews.

The MLRL typically involves:

- *Data Preprocessing:* The datasets to be linked are preprocessed to standardize formats, clean inconsistencies, and handle missing values. This step is crucial for ensuring the quality and compatibility of the data for the machine learning models.
- *Feature Extraction:* Relevant features or attributes are extracted from the datasets to represent each record. These features may include demographic information, textual data (e.g., names, addresses), numerical data, or other characteristics that can help distinguish between records.

- *Model Training*: Machine learning models are trained on a labeled dataset, where pairs of records are labeled as matches or non-matches. Various supervised learning algorithms can be used for this task, such as logistic regression, decision trees, random forests, support vector machines, or deep learning models like neural networks. During training, the models learn patterns and relationships in the data to predict whether pairs of records represent the same entity or not.
- *Matching:* Once trained, the machine learning models are applied to unseen data to predict the likelihood of matches between pairs of records. The models assign a probability score or similarity measure to each pair, indicating the likelihood that they represent the same entity. Thresholds can then be applied to determine which pairs are considered matches based on these scores.
- *Evaluation and Refinement:* The performance of the MLRL models is evaluated using metrics such as accuracy, precision, recall, and F1 score. The models may be refined by adjusting hyperparameters, feature selection, or incorporating additional training data to improve performance.

MLRL offers several advantages over traditional linkage methods, including automation of the linkage process, scalability to large datasets, and the ability to handle complex data structures and relationships. However, MLRL also requires careful consideration of feature selection, model training, and evaluation to ensure accurate and reliable linkage results. It also requires increased staff training to ensure the process' sustainability over time.

Record Review

Record review can be a laborious process and subject to subjective interpretation. Developing guidance for classification and having multiple reviewers classify common cases to refine the guidance will improve consistency and reduce bias. Manually reviewing and reconciling possible matched pairs from data linkages involves a systematic process to validate the accuracy of the linkage and resolve any discrepancies or uncertainties. To ensure that manual review is replicable over time and between staff, a defined process and protocol should be established and followed. This may include:

- *Define Review Criteria:* Establish clear criteria for what constitutes a valid match between records. This may include criteria such as matching names, addresses, dates of birth, or other relevant variables. Define thresholds or rules for determining whether pairs are matches or non-matches.
- *Sample Selection:* Select a representative sample of linked pairs for manual review. The sample should be large enough to provide meaningful insights but manageable enough to review thoroughly. Consider stratifying the sample to ensure representation across different demographic or data characteristics.
- *Review Process:* Examine each pair of linked records manually, comparing the information in each record to assess whether they likely represent the same individual or entity. Verify the accuracy of matching attributes such as names, addresses, and other identifying information. Pay attention to discrepancies or inconsistencies between records that may indicate a potential mismatch. Consider additional contextual information or external sources to aid the review process, such as supplementary documentation or expert input.

- *Reconciliation:* For pairs that meet the criteria for a match, reconcile any discrepancies between the records to ensure data consistency. This may involve merging information from both records or selecting the most accurate data. For pairs that do not meet the criteria for a match, document the reasons for non-matching and any discrepancies or uncertainties encountered during the review process. Resolve any ambiguities or conflicting information by consulting additional sources or seeking input from domain experts.
- *Documentation:* Document the outcomes of the manual review process, including details of matched and non-matched pairs and any discrepancies or issues encountered. Record the reasons for match or non-match decisions, along with any relevant notes or observations. Maintain clear documentation of the reconciliation process for transparency and reproducibility.
- *Validation:* Validate the results of the manual review by comparing them to the original linkage results generated by automated methods. Assess the level of agreement between the manual review findings and the automated linkage results. Investigate any discrepancies including those between the manual and automated results to identify potential areas for improvement in the linkage process.
- *Iterative Improvement:* Use the findings from the manual review to refine and improve the linkage process. This may involve adjusting linkage criteria, revising matching algorithms, or addressing data quality issues identified during the review. Implement changes based on lessons learned from the manual review to enhance the accuracy and reliability of future linkage efforts.

When using a linkage software to classify linkage probabilities, potential matches are often presented to the reviewer to classify. The reviewer must consider the identifiers in each dataset and determine if they are the same individual. For some cases this may be a simple to identify typo like John Smith vs. Jonh Smith, but others may be more difficult like Jon Smith vs John Smith. In many cases having access to additional administrative records like birth records, medical records, child welfare records can help with reconciliation.

Establishing Review/Acceptance Thresholds

When establishing review and acceptance thresholds for probabilistic data linkages, several best practices can ensure accuracy and reliability. First, it's crucial to involve domain experts who understand the intricacies of the data and its context. They can provide valuable insights into setting appropriate thresholds tailored to the specific requirements of the linkage process. Second, a comprehensive understanding of the data quality is essential. This involves assessing the completeness, accuracy, and consistency of the data sources involved in the linkage process. Utilizing statistical techniques such as receiver operating characteristic analysis or precision-recall curves can aid in determining optimal thresholds that balance sensitivity and specificity, ensuring both high linkage accuracy and minimal false positives. Additionally, iterative testing and validation are key to refining thresholds over time as new data becomes available or as the linkage process evolves. Regular monitoring and adjustment of thresholds based on performance metrics can help maintain the integrity of the linkage process and ensure its effectiveness in various applications. Tools like the Pareto distribution can also be used to help establish linkage thresholds for manual review. Iterative manual review with multiple people can also help identify and establish thresholds.

Evaluate Linkages

This section emphasizes conducting linkage assessments to ensure accuracy. It introduces diverse evaluation metrics, performance optimization techniques, and strategies for error handling. By scrutinizing data linkages meticulously, organizations can enhance the reliability of their analyses and fully leverage their data resources. Methods for assessing linkage quality (e.g., comparison with a gold standard, conducting clerical review, etc.) are described in detail in Chapter 3 of the UK Office of National Statistics' Quality Assessment in Data Linkage Guidance, with a simple descriptive table of the key methods.

In general, when two data sets are being combined using multiple partial identifiers, a large number of high-quality matches can be made through deterministic matches. The remainder of the unmatched records are what ultimately defines the quality of the linkage. When starting, it is often prudent to use deterministic matches, check and reconcile duplicates, quantify the percent of records linked, then extend the linkage to incorporate uncertainty in matches for those remaining, reconcile and check for duplicates and again quantify the percent linked. This iterative process will help identify when a large degree of uncertainty results in poor quality matches. A few key points or linkage quality are described in detail below.

Linkage assessment

Specific criteria and processes may vary depending on the context and requirements of your data linkage project. Calculating linkage quality, in theory, is simply quantifying the proportion of false and missed links. Quantifying these proportions, however, is often challenging to calculate due to the lack of a clear gold standard. In nearly all circumstances, one must make a trade-off between precision and recall or attempt to balance these. Thus, linkage quality is predicated on the overriding assertion of the linkage project (i.e., does the project value false linkages or missed linkages?).

Example: Let's consider a data linkage project where we're trying to match records to find individuals who might have been exposed to a contagious illness. In this case, we really want to make sure we catch as many potential cases as possible, even if it means including some people who weren't exposed. This is what we call 'sensitivity' in data linkage. It's similar to casting a wide net to capture all potential cases, even if it means some false matches are also captured.

Now imagine we're trying to match records to identify individuals involved in a car accident. Here it is crucial that we only include records of people who were genuinely involved in the accident, even if it means we might miss a few cases. This aspect of ensuring accuracy and relevance by selectively including relevant records while excluding those that are not related to the event is what we call 'specificity' in data linkage. It's about being precise in our selection process to accurately identify individuals who are truly involved in the event, while also minimizing the inclusion of unrelated records.

Linkage error results in the factors that influence the choices between record pairs. All linkage algorithms require specifications that can result in error. Measuring this error is often best accomplished through multiple methods (triangulation), including the development and testing of a training dataset with extensive validation through multiple sources.

To assess potential differential bias in linkages, linkage probability distributions for linked data should be assessed overall by demographics, geographic regions, economics, and key exposure and outcome variables. Table 2 is an example of the type of information that should be recorded and documented for linkage projects by source and Table 3 is an expanded assessment that evaluates linkages by demographics recorded on the birth record.

Table 2. General data linkage results: the % PRAMS records that were successfully linked with each administrative data source records, overall (all years of PRAMS linked), and by each PRAMS year.

Source Data	**# of PRAMS Records**	**# of Source Records**	**# of PRAMS Records Linked with Source Records***	**% of PRAMS Records Linked to Source Data**
Insert name of administrative data source linked with PRAMS]				
Overall (YYYY - YYYY)*				
[By PRAMS year - insert rows for each year of PRAMS linked]				
[Add rows for each additional source linked with PRAMS]				
Notes: *If relevant, please specify the total PRAMS records linked and include the breakdown of the number linked deterministically and those detected through probabilistic determination or manual review (e.g., If there are 1000 PRAMS respondents and 900 of them were linked using full name, DOB, and sex using a basic edit distance approach, one would report the number linked (900) and the number with exact matches (n1) and those detected through manual review (n2), where n1 + n2 = 900.)				

Table 3. Characteristics of the % of source data and PRAMS observations that were successfully linked, overall and by race/ethnicity, age, education, and payer sources*.

Source Data	**# of PRAMS Records**	**# of Source Records**	**# of PRAMS Records Linked with Source Records**	**% of PRAMS Records Linked to Source Data**
Overall				
Race/ethnicity				
[Racial/ethnic group 1]				
[Racial/ethnic group 2]				
Birthing parent age				
[Age group 1]				
[Age group 2]				

Birthing parent education				
[Education group 1]				
[Education group 2]				
Insurance type during pregnancy				
[Insurance type 1]				
[Insurance type 2]				
**These demographics are often available on the birth certificate and have been documented to be subject to differential linkage accuracy. If not available, please indicate NA in the table.*				

Below is a list of methods that can be used to help assess the accuracy of linkages.

Error Handling

- Error documentation: Document all errors, including missed links and incorrect links.
- Error correction: Develop a procedure for correcting errors and re-running the linkage.
- Error rates: Measure and document error rates, including missed links, incorrect links, and any other data quality issues that arise during linkage.

Data Quality Assessment

- Data completeness: Assess if linked data retains its integrity and completeness.
- Data consistency: Check if the linked data aligns with expectations and business rules.

Evaluation Metrics

Below is a list and description of commonly used metrics for assessing the quality of data linkages.

- Precision: The proportion of correctly linked records out of all linked records. High precision means fewer false positives (incorrect links).
- Recall: The proportion of correctly linked records out of all possible linked records. High recall means fewer false negatives (missed links).
- F1 Score: The F1 score is the harmonic mean of precision and recall. It provides a single metric indicating better data linkage quality.
- Confusion matrix: A confusion matrix provides a detailed breakdown of true positives, true negatives, false positives, and false negatives. It is a valuable tool for assessing the performance of the linkage process. When evaluating the linkage match quality, a simple two-by-two confusion matrix is useful for documenting and assessing match rates (Table 4).
- Record-level comparison: This method involves comparing linked records against a gold standard or a subset of manually reviewed records. Key metrics include precision and recall.
- Sample-based review: Randomly sample a subset of linked records and manually review them to assess data quality. This can help identify issues not caught by automated methods.

- Visualizations: Use visualizations like histograms, scatter plots, or linkage graphs to gain insights into the linkage process and identify patterns or anomalies.

Table 4. Two-by-two confusion matrix of match rates.

		Truth		
		Linkage	**Non-Linkage**	
Predicted	**Linkage**	True Link (a)	False Link (false positive, Type I error) (b)	Total predicted linkages (p1)
	Non-linkage	Missed Linkage (false negative, Type II error) (c)	True non-linkage (d)	Total predicted non-linkages (p2)
		Total True Linkages (t1)	Total True non-Linkages (t2)	Total records (M1)

- Sensitivity (aka Recall) = $\frac{a}{t_1}$
- Specificity = $\frac{d}{t_2}$
- Predictive value positive (aka Precision) = $\frac{a}{p_1}$
- Predictive value negative = $\frac{d}{p_2}$
- Accuracy = $\frac{(a+b)}{M_1}$
- F1 Score (aka harmonic mean of precision and recall) = $\frac{\left[\left(\frac{a}{p_1}\right)\cdot\left(\frac{a}{t_1}\right)\right]}{\left[\left(\frac{a}{p_1}\right)+\left(\frac{a}{t_1}\right)\right]}$

All evaluation metrics have limitations and may or may not be appropriate depending on the project's intent or linkage methods used. For example, a linkage project with a clearly specified question versus a general linkage project with multiple questions will establish linkage quality differently. Regardless, all methods should retain information related to linkage match probabilities (i.e., confidence) in the match to enable researchers to refine and interpret analyses.

Setting a Target

Determining what Is "good enough"

Evaluating data linkage quality is a critical step in assessing the reliability and accuracy of linked datasets. To determine what level of quality is "good enough," you must consider the specific context of your project, including the data's purpose and the potential consequences of errors. It is critical to remember

that the goal of the project is to combine the records and conduct analysis using these data. A lot of time can be spent tweaking and adjusting linkages with minimal gains in accuracy. The goal should be getting linkages completed. It is important to remember that while overall gains may be minimal, they may be substantial with some marginal populations.

The acceptability of data linkage quality varies based on several factors, and there is no one-size-fits-all answer. You should consider the following aspects when determining what level of quality is acceptable:

- **Context and Purpose:** The intended use of the linked data is a crucial factor. In some applications, such as healthcare or financial transactions, high-quality linkage is critical due to potential legal, ethical, or financial consequences.
- **Tolerance for Errors:** Consider the impact of false positives and false negatives on your analysis or decision-making. Some applications may have a low tolerance for errors, while others can accommodate a higher error rate.
- **Resource Constraints:** Assess the resources available for data linkage, including time, budget, and human resources.
- **Risk Assessment:** Conduct a risk assessment to understand the potential consequences of data linkage errors. This can help determine the required quality level. What is the impact on the data and inferences made with false positive and false negative matches? This should be both conceptual (e.g., scenario assessments) and analytical (e.g., sensitivity analysis).
- **Continuous Improvement:** Consider the feasibility of ongoing monitoring and improvement of data linkage quality. It may be acceptable to start with a lower-quality linkage process if there is a plan for continuous enhancement.
- **Benchmarking:** Like surveys, benchmarking is a critical step and can help build confidence in estimations derived from the linked data. Linked PRAMS data has the unique benefit of allowing for a comparison with the full birth cohort, including benchmarking linkage performance and quality.

What is "good enough" for data linkage quality depends on your project's specific needs and constraints and may require a balance between precision and recall. Regularly review and reassess the quality to ensure it meets your evolving needs and expectations.

- Tools like the Pareto distribution can be used to help establish linkage thresholds where successful matches are statistically relatively unlikely given all possible combinations.
- Each linkage project should conduct sensitivity or extreme case analysis to help determine when is good enough.
- Potentially most critical is to evaluate linkages by different sub-group analyses to quantify potential differential misclassification.

Choosing a Linkage Tool

Multiple tools are available to support data linkages. Choosing the best approach and tool depends on various factors, including your purpose and specific requirements, budget, technical expertise, scalability needs, and the nature of the data you're dealing with.

Here's a general guideline to help you select a linkage tool:

- **Define Requirements:** Consider factors such as the volume of data, types of data sources (structured, unstructured, semi-structured), data quality, types of identifiers available in each source, security and compliance requirements by your IT department, and any specific matching algorithms or rules you need.
- **Flexibility of Matching Algorithms:** Look for tools that offer a variety of matching algorithms such as deterministic, probabilistic, and machine learning-based approaches. Most linkage tools implement the Fellegi-Sunter model for calculating linkage probabilities, but other methods, such as those employed by Contiero et al. in the EpiLink Record Linkage, provide additional flexibility. Some tools enable different algorithms to be applied to each element and weighted or allow for other flexible refinements to be implemented. Depending on your data, you may need different algorithms to achieve accurate linkage.
- **Scalability:** Consider the scalability (e.g., processing demands) of the tool and your need for the ability to handle increasing volumes of data.
- **Integration Capabilities:** Determine whether the tool can integrate with your existing systems and data sources. It should support the data formats and protocols used in your environment.
- **Data Quality and Cleansing:** Check if the tool provides features for data quality assessment, cleansing, and standardization. Data quality issues can significantly impact the accuracy of linkage results. These processes are often referred to as extract transformation and loading (ETL).
- **Threshold establishment and manual record review:** Review whether the tool has features to support establishing acceptance and review regions (e.g., use of the generalized Pareto distribution). Additionally, review how manual record review is completed and integrated into final matched results. Disparate approaches result in data degradation, and tools with interactive reviews of “possible matches” and reconciliation support maintaining data integrity.
- **Security and Compliance:** Ensure that the tool meets your security and compliance requirements, especially when dealing with sensitive or regulated data. Look for features such as encryption, access controls, and compliance with relevant standards (e.g., GDPR, HIPAA). Cloud-based tools should be used cautiously to ensure they meet your required security standards.
- **Ease of Use and Maintenance:** Consider the tool's usability and whether it requires extensive technical expertise to set up and maintain. A user-friendly interface and comprehensive documentation can simplify the implementation process. Code-based solutions developed in SQL, Python, or R should be well documented and have multiple staff trained on its use.
- **Performance and Speed:** Evaluate the performance and speed of the tool, particularly if you have large volumes of data. The tool should be able to process data efficiently without significant latency.
- **Cost and Licensing:** Compare the pricing models of different tools, including upfront costs, licensing fees, and ongoing maintenance expenses. Consider both the initial investment and long-term operational and staff training costs.
- **Trial and Proof of Concept:** Whenever possible, try out the tool through a trial or proof of concept to assess its suitability for your specific use case. This will allow you to evaluate its performance and functionality in a real-world environment before making a commitment and, if possible, compare it with other tools.

Although it may seem obvious, it is important to have linkage software that makes it easy to get both the data that matches and the data that does not and provides a mechanism for reviewing and reconciling potential matches. Most often, identifying the best tool and linkage algorithms is an iterative process. It is important to spend time understanding the underlying population demographics, racial distributions, and how different tools and algorithms perform with different naming structures. For example, SOUNDEX, a phonetic algorithm, will result in different linkage rates by racial groups. Working with a demographer may help facilitate the creation of a test linkage data set based on the underlying population distribution to help test the performance of different tools and the implementation of various algorithms. This process can help quantify any potential linkage error.

Scan of selected linkage tools

Table 5 provides a brief high-level description of commonly used or widely available tools.

Table 5. Selected linkage software and brief description.

Tool	Description
LinkSolv	LinkSolv computes Bayesian probabilities to identify true record matches. It assesses candidate record pairs by comparing data values and calculating match probabilities. It also provides an interface to prepare, compare, and validate linkages.
Link King	Link King is an SAS/AF tool that uses a data importing and formatting wizard, artificial intelligence to ensure the use of appropriate linking protocols, a powerful interface for manual review, the ability to generate random samples of links for validation, and easy "point-and-click" editing. Link King uses both probabilistic and deterministic algorithms.
Link Plus	Link Plus can detect record duplicates and link registry files with external files. It computes probabilistic record linkage scores with value-specific (frequency-based), last name and first name, middle name, date, social security number, generic string, or ZIP code matching methods. Missing values of matching variables are treated as null or empty automatically. Link Plus facilitates blocking, comparing pairs with identical values on at least one variable, and has two phonetic coding systems.
Match*Pro	Conducts probabilistic record linkage based on the Fellegi-Sunter model, using tools to conduct data validation using pre-defined or custom validators. The user-friendly interface allows manual review with flexible configurations and the ability to specify blocking and matching methods, adjust the blocking sensitivity, define unknown values, set weights, and perform substitutions.
R: RecordLinkage	Provides functions for linking and de-duplicating datasets with supervised and unsupervised classification. Methods are based on a stochastic approach using an expectation-maximization (EM) algorithm. Employs machine learning methods and thresholds can be determined by tools based on extreme value theory.
R: FastLink	FastLink uses a Fellegi-Sunter probabilistic linkage method, using the expectation-maximization (EM) algorithm. This method allows for missing data and the inclusion of auxiliary information.
Python: recordlinkage	This toolkit provides most of the tools required for data linkage and deduplication

	of small or medium-sized files. It includes indexing methods and functions to compare records and classifiers. Recordlinkage uses pandas and numpy.
Python: Splink	Splink uses probabilistic record linkage to deduplicate and link datasets without unique identifiers.
Python: RLTK	A full, scalable record linkage pipeline, including multi-core algorithms for blocking, profiling data, computing a wide variety of features, and training and applying machine learning classifiers based on Python's sklearn library.
Python: Zingg	Zingg uses machine learning to resolve entities in five phases: findTrainingData, label, train, match, and link.
FRIL	A JAVA-based product, FRIL contains a rich set of user-tunable parameters, advanced features of schema/data reconciliation, search methods (e.g., sorted neighborhood, blocking, nested loop join), transparent support for multi-core systems, and additional features.
ChoiceMaker	ChoiceMaker employs a three-step process of blocking, scoring, and transitivity analysis using machine learning.
SQL: PRIL	Point-of-contact Interactive Record Linkage (PRIL) prospectively links data using a probabilistic record linkage algorithm based on the Fellegi-Sunter model.
Chimera	An open source machine learning model-based approach to family based record linkages. For more information about this resource, please reach out to John Prindle at jprindle@usc.edu.

Document Linkage Process

A hallmark of successful linkage projects is quality documentation. Documenting the linkage process, decisions made, and quality version control ensures that the methods and approaches used can be critiqued, communicated, and replicated. Documentation of the linkage process should include the following:

- **Brief project overview**: Describe the project purpose and goals and why the data linkage is necessary.
- **List and description of data sources**: List all data sources and describe the format, structure, size, identifiers included, years available, who owns/stewards the data, limitations on use, or population coverage. Include the paths of where the original data are stored locally (if applicable).
- **Overview of data preparation**: Describe the steps taken to prepare/harmonize the data for linkages and what cleaning, standardization, and transformations were completed. Include any changes to the original data sources or elements required for the linkages.
- **Linkage methods deployed**: Provide an explanation of the linkage algorithms or methods used to match records, what parameters or thresholds were used, and how they were determined. Include how potential errors, manual review, or uncertainty in the linkage process were resolved.
- **Matching quality assessment results**: Describe how the quality of the linkage results was determined (e.g., precision, recall, F-score, or other metrics). Describe any manual or automated

validation process used to assess the accuracy of the matches. If a training dataset was used, describe how it was constructed.

- **Results and outputs:** Provide a summary of the linked dataset, including the number of matched records and any additional variables generated through the linkage process. Tables 2 and 3 under the Linkage Assessment subsection should be included.
- **Software and tools used**: List the software and version, including any libraries or tools used for data linkage. Provide a description and path to any custom scripts or code developed for the project.
- **Challenges and limitations**: Describe challenges encountered during the linkage process and how they were addressed. Include any limitations or assumptions inherent in the linkage methodology and any limitations that should be considered for analyses and further use or sharing of the linked data.

Validation

Linking PRAMS data with administrative sources offers a distinct advantage – the ability to compare sample estimates to the entire birth cohort. PRAMS jurisdictions should validate the linked PRAMS weighted estimates by directly comparing them with at least one full birth cohort year, but preferably with multiple years.

In addition to comparing statewide weighted estimates with observed full birth cohort measures, PRAMS jurisdictions should stratify by various demographic factors and conduct similar comparisons. It is recommended that comparisons be conducted across 10 to 20 different strata.

As a guiding principle, PRAMS weighted estimates are considered valid population estimates when the PRAMS 95% confidence intervals of the strata compared contain the observed full birth cohort measure of the linked outcome 90% of the time. If the observed population is outside the weighted confidence interval, a review should be conducted to ensure consistency in both the linkage methods and outcome specifications in the outcome source.

Linking the entire birth cohort does increase resource requirements, so it should be purposefully considered when determining the number of birth years to link. Refer to Figure 6 for an overview of the general process for validating linked PRAMS weighted estimates with a full birth cohort. Additionally, Table 6 outlines the suggested information that should be recorded and compared.

When calculating observed measures with the full birth cohort to compare with, it is important to limit the cohort to the sampling frame, which is generally limited to mothers of in-state live births, with multiple birth events limited to only one if sampled.

Figure 6. Process for validating the PRAMS respondent sample estimates with the observed full birth cohort.

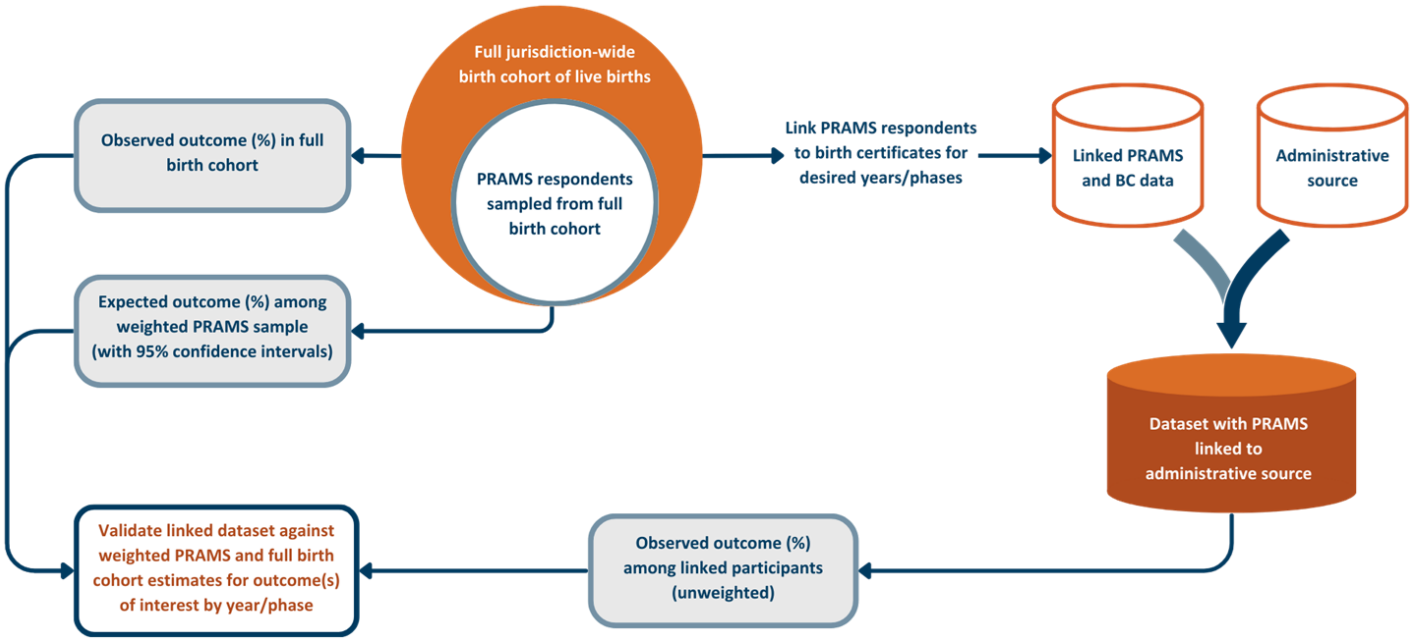


Table 6. Linked PRAMS weighted estimates validated against the full birth cohort.

	Unweighted PRAMS %	**Weighted PRAMS % (95% CI)**	**Full Birth Cohort %**
Outcome			
[Add rows for additional outcomes]			

Notes:

1. If you used multiple PRAMS years, please add rows to document results for each year and overall (all years combined).
2. If you assessed multiple strata levels (e.g., sex, race, education), please add rows to document results.

Example: Consider a scenario where a jurisdiction aims to examine the association between maternal access to healthcare, as measured in PRAMS data, and severe maternal mortality (SMM) identified in hospital discharge records (HDD). First, an algorithm can be employed to identify SMM cases within HDD records. These cases can then be linked with all recorded birth events in the state over a specified number of birth years. This dataset can then be subset to match the PRAMS sample from the same birth years.

Weighted estimates of SMM and their corresponding 95% confidence intervals can be calculated using the PRAMS-linked cohort. These estimates can then be compared with the observed SMM measures derived from the full birth cohort, considering all years combined, individual years, and various demographic strata. The demographic strata could include five levels of maternal education, five levels of maternal age, six levels of race, two levels of infant sex, and two levels of infant birth weight, amounting to a total of 20 strata.

Each PRAMS weighted estimate of SMM within these strata could then be compared against the corresponding SMM measure from the full birth cohort and documented. It is expected that approximately eighteen of the PRAMS 95% confidence intervals will encompass the population measure (i.e., the population measure is between the upper and lower bounds of the PRAMS estimated confidence interval). This simple process is extremely effective at validating PRAMS-weighted estimates produced from linked data and helps identify potential linkage errors, coding errors, or other issues that could lead to biased results.

PHASE IV: Research Dataset Creation and Analysis

Once the PRAMS respondents have been linked with the administrative source(s) and validated, analysis to support the original purpose of the project can occur. In this phase, the PRAMS respondent data elements are selected and coded and the administrative data elements are coded and organized to capture health outcomes of interest.

Create Analysis Plan

Each analysis should begin with an analysis plan that provides clear documentation and establishes how the research dataset should be created. Appendix F provides a basic template for documenting the PRAMS data linkage protocol/plan. The bulleted list below describes what is typically included in an analysis plan:

- Research lead contact information
- Clear project aims and research question(s)
- Data sources linked and years
- How the research dataset was constructed
- Exposure, covariate, and outcome variables to be used and how variables were constructed
- Inclusion/Exclusion criteria
- Statistical analysis

Creating Research Datasets

The following outline is based on the recommended data structure described in Phase I. If a different data structure is used, the actual mechanics of the research data construction will vary.

When creating a research dataset:

- Create an analysis plan and document the process for creating the research dataset. Appendix F provides a basic template for documenting the PRAMS data linkage protocol/plan, which should have been developed during Phase I and kept updated. A separate analysis plan should be created for long-term projects.
- Maximize confidentiality by ensuring that identifiers used for linkages are removed from analysis datasets.
- Create a data dictionary.
- Ensure that the full PRAMS respondent sample is used in the analysis.
    - Nonlinked records should be treated as missing data.

Using the recommended data structure, individual research data sets can be created that are free from identifiers and focused on answering the specific research question(s). The linked source elements should be processed to identify the outcome, covariate, or exposure of interest. The PRAMS data should be combined across selected years, with the survey questions aligned if including multiple Phases, and limited to the variables of interest. The processed PRAMS and source data should then be combined using the ID_Master and Source_ID tables (Figure 7). As indicated above, it is important to make sure

that the final research dataset has the same number of responses as the PRAMS respondent data used. This is critical for ensuring that the PRAMS post-stratification weights are applied correctly and produce correct results. The best practice is to verify the PRAMS sample with the linked research data set and verify that the estimates for a PRAMS response variable produce consistent estimates and standard errors.

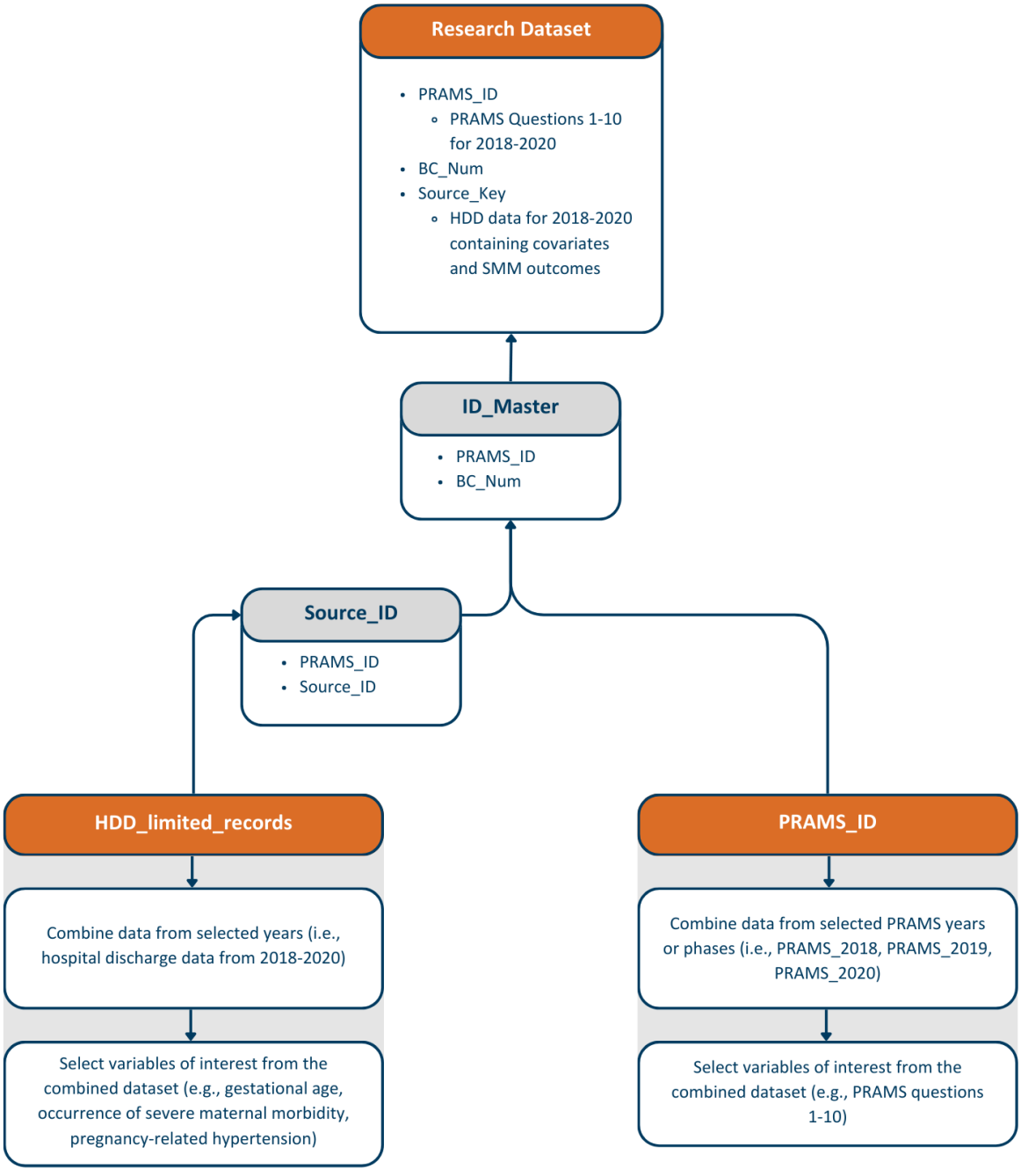
Creating a Data Dictionary

It may not be feasible to create a specific data dictionary for the linked data, although a general dictionary that outlines the elements and type of data contained in each source should be created. For all created analytical research files derived from the linked data, a specific dictionary should be created. An analytical research file data dictionary outlines the structure and content. It serves as a reference guide for data analysts, scientists, and other stakeholders to understand the data elements and ensure consistency and accuracy in analysis.

Recommended elements to include in a linked PRAMS research file data dictionary:

- Data Element Name: The name of each data element (e.g., column names in a dataset).
- Description: A brief description of what each data element represents or measures.
- Data Type: The type of data (e.g., text, numeric, date, Boolean) for each element.
- Format: The format or structure of the data (e.g., YYYY-MM-DD for dates, currency format for monetary values).
- Source: PRAMS, birth record, linked source.
- Units: For numerical data, specify the units of measurement (e.g., meters, kilograms, dollars).
- Validity: Any rules or constraints that apply to the data (e.g., valid range of values, allowed formats).
- Dependencies: Relationships or dependencies with other data elements or datasets.
- Quality Issues: Known data quality issues, such as missing values, duplicates, or outliers.
- Last Updated: Date when the data dictionary was last updated to reflect changes.
- Metadata: Any other relevant metadata, such as creation date, author, or version.

Figure 7. Creating a research dataset from linked PRAMS - Hospital discharge data using the recommended data structure. Illustrated using an example research topic on severe maternal morbidity outcomes among PRAMS respondents, 2018-2020.



Conduct Analysis

To address a key research question using PRAMS-linked data, researchers typically begin with univariate analysis, examining individual variables such as maternal age, race/ethnicity, and smoking status to understand their distributions and characteristics. This provides an initial overview of the data and

identifies potential associations between individual variables and the outcome of interest, such as low birth weight or preterm birth. Next, researchers conduct bivariate analysis to explore relationships between pairs of variables, such as the association between maternal smoking and birth outcomes stratified by maternal age or race/ethnicity. Bivariate analysis allows for the identification of crude associations and potential confounders that may need to be adjusted for in a multivariable analysis. Finally, researchers perform multivariable analysis, using regression modeling techniques such as logistic regression, Poisson, log-binomial, or linear regression, or survival analysis methods to assess the independent effects of multiple predictors on the outcome while controlling for confounding variables. This enables the identification of factors that are independently associated with the outcome of interest, providing insights into the complex relationships between maternal characteristics, pregnancy experiences, and health outcomes for mothers and infants. It is critical to remember to include the full PRAMS cohort and use appropriate software functions to incorporate the PRAMS survey design, which includes the annual weight, strata, and FPC and/or total count.

Generating Reports

Once the linked dataset is established and a research dataset created, a range of analytical procedures can be applied to extract meaningful insights and metrics. This involves employing various statistical analyses, data mining techniques, and modeling approaches to explore relationships, detect patterns, and generate actionable findings. While scientific papers serve as one avenue for disseminating results, their impact within government and policy spheres may be limited. Conversely, alternative communication channels such as memos, presentations, or factsheets can wield substantial influence within government settings. These formats allow for a concise and accessible presentation of findings tailored to the specific needs and interests of policymakers, public health officials, or researchers. The synthesized results are then compiled into comprehensive reports or information products, integrating visualizations such as charts, graphs, and tables to effectively convey complex data in a digestible format. These visual aids play a crucial role in facilitating informed decision-making and program evaluation, empowering stakeholders to utilize evidence-based insights for policy formulation and programmatic interventions.

For peer-reviewed publications, researchers should follow the GUILD (GUidance for Information about Linking Data Sets) to ensure that the key project details are described to facilitate scientific scrutiny and replication. Researchers should provide enough detail (often in appendices) to replicate the linkage project and, at a minimum, review the matching methods and linkage rates.

For other information products, balancing how much of the methods and procedures should be included to help contextualize the results will depend on the format of the information product and the intended audience. An online resource that describes the linkage methods and highlights limitations can be referenced in information products like infographics that have limited space to adequately describe methods. It is important to help your audience understand the source of information and types of data included, as disseminating the information will prompt additional collaborations and new investigations.

As mentioned before, a key benefit of PRAMS is the representativeness of a jurisdiction's population. Sharing analyses from linked data with community partners can be empowering as it represents the

collective voice of the jurisdiction. While scientific peer review is important, sharing results with community leaders, legislators, non-profits, and other partners is paramount and will lead to direct impacts on the population from which the data were derived.

Example: Let's consider how linked data can be used to influence policy. The Alaska Longitudinal Child Abuse and Neglect Linkage project (ALCANLink) explored pre-birth household challenges and the accumulation of Adverse Childhood Experiences (ACEs) by age three years (Study 1) and how changes in these household challenges between the pre-birth and early childhood periods impact child welfare involvement (Study 2). Both studies were published in peer review literature but had minimal impact on state policy. These data were shared with the director of the Division of Behavioral Health through a presentation and brief report summarizing key points, who then included the results to apply for and obtain an 1115 Medicaid waiver that enables clinical providers to bill for activities that can mitigate ACEs. The specific billable activities identified z-codes that directly mapped onto the associated factors identified in the studies.

PHASE V: Sustainability

Embedded within a PRAMS linkage project should be a plan for the potential continuation of linkage activities. Phase V discusses the several elements that can be implemented to ensure a comprehensive sustainability plan.

Strengthen Capacity and Relationships

In a state or other jurisdictional public health department, securing adequate funding and resources is essential for the sustainability of a PRAMS data linkage project. However, staff turnover and limited expertise can pose significant challenges. High turnover rates can disrupt project continuity and knowledge transfer, resulting in gaps in data management and analysis. To address this, prioritizing staff training and professional development is crucial to cultivate expertise and retain skilled personnel.

Developing standardized operating procedures and comprehensive documentation is equally vital. This approach ensures continuity and facilitates the seamless onboarding of new team members. Collaborating with academic institutions or external experts can provide additional resources and specialized knowledge, strengthening the project's resilience against staffing challenges.

Moreover, fostering a supportive organizational culture that values knowledge sharing and mentorship can promote staff retention and enhance the project's long-term sustainability. Jurisdictions must prioritize funding to support at least one dedicated staff member for the PRAMS linkage project.

Sustainability also hinges on establishing reliable infrastructure for data collection, storage, and analysis, leveraging emerging technologies for efficiency while adhering to ethical guidelines. These are often influenced by broader infrastructure development occurring in individual departments. However, each PRAMS data linkage project should clearly document processes, where data is stored, security measures and remediation plans made, a select few tools used, and a register of prior analysis, publications, and code used.

To facilitate ongoing linkages with PRAMS, jurisdictions should maintain engagement with partners within and external to their agency to support data sharing, routine linkages, and future investment in linkage activities. Longitudinal linkage of PRAMS with various administrative sources involves connecting PRAMS data with other datasets over multiple time points, creating a longitudinal dataset that tracks individuals' experiences and outcomes over time. This longitudinal approach offers several advantages over one-time linkages. It allows agencies and researchers to analyze trends and changes in maternal and child health outcomes over time, identify patterns, and assess the effectiveness of interventions or policies longitudinally. Additionally, longitudinal linkage can serve as an ongoing tool for evaluating the impact of public health programs and policies, providing valuable insights into long-term health trajectories and informing programmatic decision-making. Moreover, maintaining a mechanism to easily add additional PRAMS years as data is released ensures the continuous updating of the longitudinal dataset, enhancing its utility for ongoing research and evaluation efforts. Demonstrating the utility of establishing and maintaining a longitudinal PRAMS linkage environment can generate buy-in from leadership and secure funding sources for sustained investment in such projects.

Jurisdictions should prioritize presenting and sharing analyses results with partners. Ongoing partner engagement and communication are essential to address emerging needs and adapt to evolving research priorities, thereby enhancing the project's longevity and impact in improving maternal and child health outcomes.

Establish Data Use Policy

Once PRAMS data has been linked, jurisdictions should identify if their agencies have existing data use policies for linked data and establish a protocol if one does not exist. It is important to ensure that any existing data use policy will still be applicable for a linked dataset, as some agencies may find their existing policies only cover the original non-linked datasets. These protocols should provide guidelines for ensuring the secure transfer, storage, and handling of linked PRAMS data, particularly when identifiers are included. External researchers should also be made aware of any necessary data-sharing agreements, IRB, and associated costs for accessing the data. Information on how jurisdictions in the PRAMS Data Linkage Learning Community outlined their processes for external use of linked PRAMS data can be found as part of the learning community final report.

Important Note: When de-identified PRAMS data is linked with another source, it introduces the potential for re-identification or disclosure of sensitive information. Despite the de-identification process, the combination of variables from multiple sources can inadvertently lead to the identification of individuals, especially if the dataset contains rare or unique combinations of characteristics. To mitigate this risk, standard review procedures should be implemented, including careful scrutiny of cell sizes and suppression of small cell counts in aggregated data to prevent the disclosure of sensitive information. Additionally, statistical techniques such as data perturbation or noise addition may be needed to add randomness to the data and reduce the risk of re-identification while preserving the integrity of the analysis. Regular assessments of disclosure risk should be conducted throughout the linkage process, and appropriate safeguards should be implemented to protect the privacy and confidentiality of individuals included in the linked dataset. Collaboration with experts in data privacy and security is essential to ensure compliance with ethical guidelines and legal requirements regarding data disclosure and privacy protection.

Documentation

One of the most critical aspects of creating a sustainable PRAMS linkage project is documenting the process. Returning to Phase I, the protocol/plan document originally created should be updated, with shell tables created and completed. Adequate documentation ensures the project can be repeated, replicated, critiqued, and communicated effectively. Appendix F provides recommended documentation.

Any code developed should be well documented, with the paths to code and data used contained within the protocol document. If non-code-based systems are used (e.g., Microsoft Excel), an explanation of how these tools were used, what was performed (including screenshots), and any other information that will facilitate replication should be provided.

At a minimum, the documentation should include:

- All research notes/documentation for completed and ongoing projects.
- Specification of challenges and resolutions.
- Location of all raw, processed, and relevant databases or research datasets.
- Updated standard operating procedures (i.e., the linkage methods and routines established).
- Software used, cost, renewals, description of special waivers needed by IT, and experts who can be consulted.
- Any critical institutional knowledge related to policies, ethics, regulations, deadlines, or other helpful information to mitigate loss of productivity or replication of prior efforts.

Conclusion

The integration of PRAMS data with administrative sources holds substantial promise for enhancing maternal and child health research, policy development, and program evaluation. This comprehensive Framework, informed by experiences from multiple PRAMS jurisdictions, underscores the importance of systematic processes across five distinct phases: preparation, data identification and harmonization, linkage approach design, dataset creation and analysis, and sustainability planning. By addressing common challenges such as data governance, ethical considerations, and resource allocation, this Framework equips public health agencies and researchers with the tools necessary to maximize the impact of PRAMS data linkage projects in addressing critical public health challenges and improving health outcomes. Through strategic implementation of this Framework, jurisdictions can harness the potential of integrated data to inform evidence-based decision-making and drive meaningful advancements in maternal and child health practice and policy.

Appendices

Appendix A. Description of state linkage approaches that participated in the ASTHO learning community.

The table below illustrates how jurisdictions linked various data sources, such as Medicaid claims, hospital discharge data, home visiting programs, and child maltreatment reports. With this wide variety, no site would have the same approach and therefore emphasized the need for a linkage approach that was well-tailored to the needs of the jurisdiction and the resources available to them. For additional details of each jurisdiction's project, including their project purposes, see ASTHO's final report on the PRAMS Data Linkage Learning Community.

State Linkage Approaches					
Jurisdiction	**Data Sources Linked**	**Data Linkage Tool**	**Identifiers Used**	**Linkage Methods**	**Proportion of PRAMS Records Linked**
Alaska	Medicaid Claims Database incorporated with prior linkages with child welfare, vital birth/death records, permanent fund dividend data, and education records.	R (Packages used: RecordLinkage, RODBC, dplyr, tidyr, lubridate, tidyverse, reshape2, stringi)	First name* Last name* Middle name* Date of birth Sex Year of birth (exact match needed) *Name data obtained from prior linkage of PRAMS to birth certificate data.	Deterministic linkage with probabilistic linkage enhancement and deduplication	59.9%
New Mexico	New Mexico State Home Visiting Database	SAS 9.4 and Match*Pro	Mother first name Mother last/maiden name Mother date of birth Child first name Child last name Child date of birth SSN when available	Deterministic linkage with probabilistic linkage and manual review	3%
Texas*	Birth Certificates Texas Health Care Information Collection Research Data File				-

Washington**	The Comprehensive Hospital Abstract Reporting System (CHARS) Washington Health and Life Events System (WHALES)	R version 4.1.0 using RStudio 1.4.1717 (Packages used: tidyverse, DBI, odbc, janitor, stringdist, data.table, e1071, stats, mltools, doParallel, snow, fuzzyjoin, stringr, pbapply, geosphere, scales, IDPmisc)		Pairwise matching followed by machine learning and deterministic matching	-
Georgia	Vital Records (Births, Fetal deaths, Deaths, Marriage) Laboratory reports Surveillance case reports	Senzing	Birth Certificate number (BCN) Record Number (e.g., BCN) Current name (first, middle, last, suffix) Maiden name (first, middle, last) Base name (first and last) Date of birth Date of death SSN Resident address Gender Employer Email address Phone number	Deterministic and probabilistic matching with machine learning and artificial intelligence	100% to birth certificates
Massachusetts	Pregnancy to Early Life Longitudinal (PELL) datasets, including: live birth certificates, fetal death records, delivery/birth hospital administration records, and non-delivery/birth hospital care (inpatient admissions, observational stays, emergency department visits), and other maternal and child health data systems linked to PELL	LinkPro v3.0 (SAS-based)	Birth Certificate number (BCN) Mother first name Mother last name Mother date of birth Date of delivery Zip code	Deterministic and probabilistic matching	100%
Montana	Vital Records (Birth Certificate and Death Certificate)/VSIMS Medicaid files (Montana's Program for Automating and Transforming Healthcare (MPATH))	Link Plus	Birth Certificate number (BCN) Infant first name Infant last name Infant date of birth Infant sex Mother maiden/ last name	Deterministic and probabilistic linkage	12-25% (depending on the data year and linked source)

	Newborn Screening (metabolic/bloodspot and hearing/critical congenital heart defect)/CHRIS Child Maltreatment Reports from Child Protective Services (CPS) Evidence-Based Home Visiting Services/MTmechv Children's Special Health Services (CSHS)				
Nebraska	Vital Birth Records Nebraska Hospital Discharge Data	Match*Pro	Mother first name Mother last name Mother date of birth Age Zip Code	Probabilistic linkage	87.6-92.7% (depending on the data year)
Rhode Island	Vital Birth Records Hospital Discharge Data	Match*Pro and SAS	Patient first name Patient last name Mother date of birth Medical record number Discharge date Zip code	Deterministic linkage and probabilistic linkage	99.3%
South Dakota	Vital Records- Birth file South Dakota Medicaid Claims Data	SQL	Mother social security number Mother first name Mother last name Mother date of birth	Deterministic linkage	49.1%
Tennessee	Vital Birth Records Hospital Discharge Data System	SAS 9.4	Mother social security number Two-name components and DOB or one-name component, DOB, and address	Deterministic linkage	100% to birth statistical data 38.1-97.5% to hospital discharge data depending on years linked and prenatal, birth, and postpartum hospital encounters
Virginia	Vital Birth Records Hospital discharge data from the Virginia Health Information (VHI) system	SQL	Mother social security number	Deterministic linkage	67.9%

* Due to staffing shortages, some jurisdictions were unable to complete or report on their data linkages during the timeframe of the learning community.
** Due to delays in IRB processing, some jurisdictions could not complete or report on their data linkages during the timeframe of the learning community.

Appendix B. PRAMS Data Linkage Readiness Assessment

Instructions: The following questions are designed to help PRAMS jurisdiction teams assess their readiness to conduct a linkage project with the PRAMS respondent data. Based on the experiences of multiple prior statewide PRAMS linkage projects, the key components of successful projects are highlighted in this questionnaire. The tool is generalized and should be used as a guideline.

To use this tool, teams should consider each question and informative bullets, and choose the most appropriate answer that reflects their current situation. Be as realistic as possible. At the end of the questionnaire, total the sum of your scores to determine your readiness level.

For each question, select a score between 0 and 5, with 0 being no development in this area and 5 being complete development and all aspects in place.

	Question	Score (0 – 5)
1	Does your program have a clear purpose for undertaking a data linkage project? ● Are the key objectives well-defined? ● Will the linked data provide something that isn't already available? ● What are the expected impacts of linking these data?	
2	Do you have a well-defined project plan? ● Do you have clear timelines? ● Do you have specified milestones? ● Will the project involve ongoing or single (static) linkages? ● Do you have anticipated products or other deliverables?	
3	Does your program have the necessary technical infrastructure and expertise for data linkage? ● Will you require staff training on linkage methods and/or the software used to conduct linkages? ● Do you have a data integration solution and has it been calibrated to your population? ● Is the data linkage solution affordable? ● Does the data linkage solution require unique or additional IT support? ● Will the linkage solution scale if desired?	

4	Have you identified relevant data sources and assessed their quality and compatibility? ● Do the data have identifiers, and can you access them? ● Is the system managed in-house or by a vendor? ● Do the data have a cost? ● Are agreed data standards and dictionaries in place or is there a process to develop them? ● Have there been any system transitions? Are the data comparable and combinable across system iterations? ● Do the years of available data coincide with the years of available PRAMS data? ● Do the data cover a statewide population or only a subset? ● Do you have data cleaning and harmonization standards?	
5	Can you access and link the data sources? ● Does the PRAMS Informed Consent (Appendix I in the PRAMS protocol documents) for the years to be linked include language about combining their responses with information the health department has? ● Do policies and procedures exist for requesting access to and use of the data?	
6	Does your agency have an existing data governance framework? ● Does the existing data governance framework cover privacy, security, and confidentiality of data? ● Are there legal and regulatory frameworks in place to govern data sharing and data protection? ● Is the process documented and clearly defined? ● Does the existing data governance framework include data linkages? ● Is an IRB required for linkages? ● Does your agency cover data governance for all data sources that will be listed?	
7	Do you have key partners engaged? ● Have you identified and engaged all required partners (e.g., IT, data stewards, legal, and analysts)? ● Do you have a project champion within each relevant partner agency? ● Have you established roles and responsibilities of key personnel/partners?	
8	Have you conducted a comprehensive risk assessment for data linkage, including potential privacy and security risks? ● Do you have a documented plan for mitigating risk associated with data sharing, linkage, storage, and use?	

9	Have you conducted a pilot or feasibility study to test the data linkage process? ● Do you have metrics in place to evaluate data quality? ● Do you have a plan for establishing acceptance thresholds? ● Do you have metrics in place to test the validity of the linkages?	
10	Have you established the rules related to analyzing the linked data? ● Do they include provisions for sharing with internal and/or external researchers? ● Do they include provisions for notifications and review of developed products?	
	Total	

Scoring:

Overall readiness to link data: Total the sum of your scores from each question.

- 41-50: Excellent Readiness – You have demonstrated a high level of readiness for a data linkage project with PRAMS data.
- 31-40: Moderate Readiness – You have made substantial progress but should address several areas to ensure an efficient and successful PRAMS linkage project.
- 21-30: Low Readiness – You have several critical areas to address and develop that will improve the likelihood of a successful PRAMS linkage project.
- <20: Very Low Readiness – Your organization is not yet ready for a PRAMS linkage project. Use the ASTHO Framework for Linking PRAMS with Administrative Data to get started and/or seek expert guidance to address the identified areas of improvement.

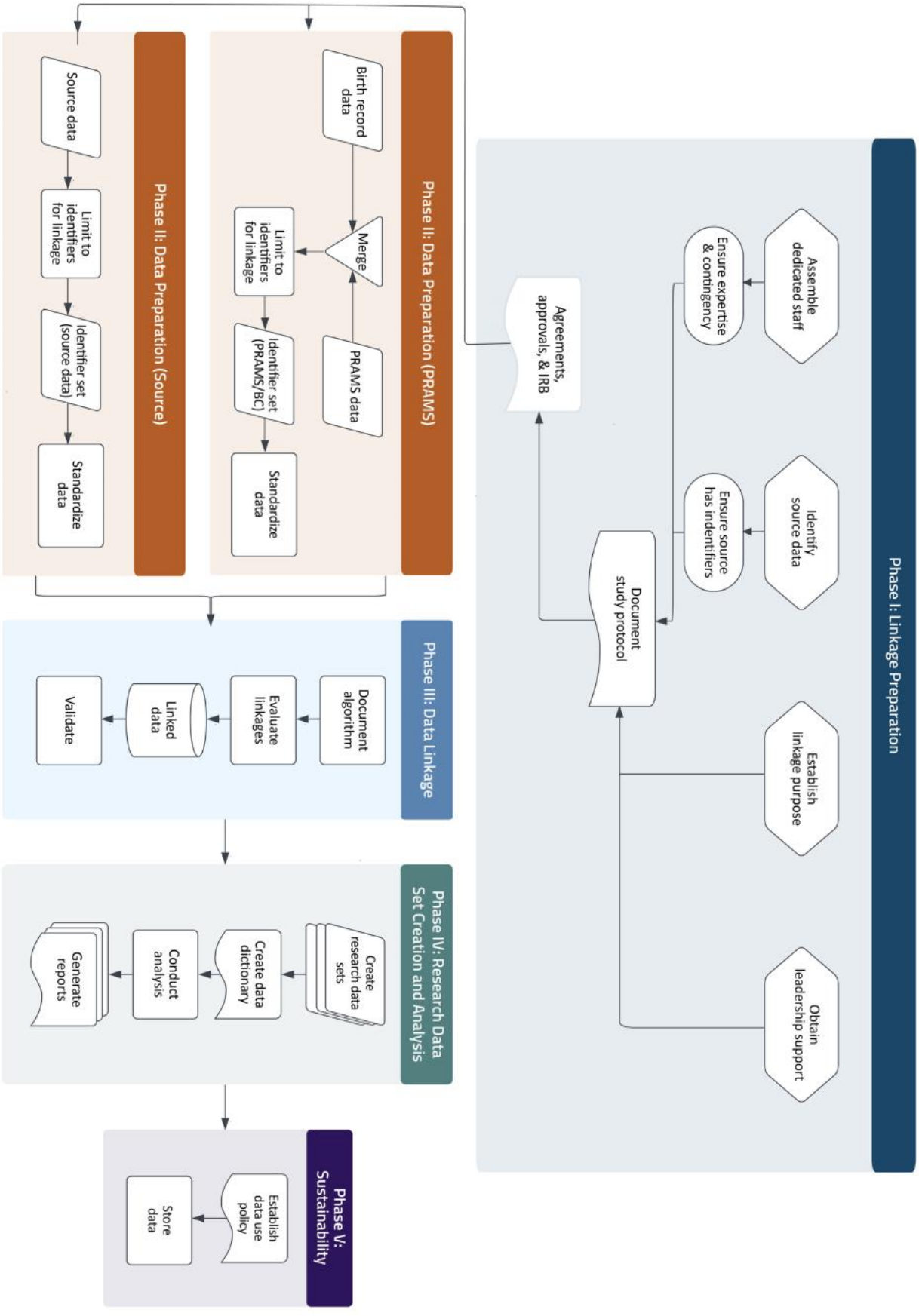
Readiness to initiate data linkages (evaluate questions 3-6):
Did you score:

- question three > or = 4? Y/N
- question four = 5? Y/N
- question five = 5? Y/N
- question six > or = 4? Y/N

If you answered "No" to any of the questions above, data linkage is either not possible (e.g., don't have access to the data) or will be extremely problematic. Addressing any identified issues are paramount before initiating a linkage project.

NOTE: This assessment only serves as general guidance as each state or other jurisdiction may have additional factors to consider. This assessment should not replace a comprehensive evaluation.

Appendix C. PRAMS Data Linkage Process Map



Appendix D. PRAMS Data Linkage Process List

Prior to linking multiple administrative data sources to the PRAMS respondents, it is highly recommended that states conduct a pilot or simplified initial project integrating a single PRAMS phase with a single patient-centered administrative source. The process list below is designed to support states in developing their initial linkages and approach but can be scaled as projects evolve. As programs integrate new sources and/or additional phases, having a simple data infrastructure and documented data governance is critical to the long-term sustainability and use of these novel data.

This tool is designed to be a phased guide by providing a synthesized overview of the various processes that may need to be accomplished within each phase of a linkage project. It is important to remember that not all processes or components will be required within each phase for every linkage project and some processes may need to be conducted multiple times.

PHASE	**Process**
PHASE I: Linkage Preparation	● Establish linkage purpose* ○ Clearly define what you are trying to accomplish by linking the data (what can be answered with the linked data that otherwise couldn't?) ○ Describe the public health practice the linked data will support ○ Identify challenges or barriers for obtaining the data (remember to consider political barriers) ○ Create a process for managing the data ● Conduct feasibility analysis (see linkage readiness assessment tool) ○ Identify gaps in learning, experience, or tools that need to be addressed ○ Determine if this will be a one-time linkage or an ongoing effort ● Form relationships and identify mutual benefits from the data linkage ● Obtain leadership support (buy-in) ● Assess capacity ○ Ensure there are dedicated staff with expertise ○ Ensure there is sufficient funding and resources ○ Identify and obtain appropriate linkage software ○ Establish a contingency plan for when critical staff leave ● Create meta-data repository (governance structure for how things work) ○ Determine data ownership and governance ○ Document data security ● Complete all required agreement(s)

	• Obtain IRB approval (if required)

PHASE II: Data Preparation	• Identify administrative data source(s)
	  ○ Ensure source(s) have common identifiers that can be accessed and used for linkage
	  ○ Identify potential limitations or unique challenges with linking to the data source(s)
	  ○ Assess data quality (including missingness) through exploratory data analysis (e.g., quality of linkage variables, distribution of outcomes to be investigated, and frequency distributions)
	  ○ Evaluate missingness and heterogeneity in elements to be used for linkages
	  ○ Conduct data deduplication (if applicable)
	  ○ Conduct data harmonization between sources including standardization and record cleaning
	• Identify PRAMS data year(s) to use
	  ○ Align variables across PRAMS years/phases
	  ○ Merge with birth records to obtain identifiers (e.g., names, birth date, and sex)
	  ○ Conduct data harmonization including standardization and record cleaning
	  ○ Evaluate missingness and heterogeneity in elements to be used for linkages

PHASE III: Data Linkage	• Establish linkage methods and approach
	  ○ Document algorithm used (e.g., rule-based deterministic, score-based probabilistic, or ML methods), iterative approaches, and blocking
	  ○ Establish rules for manual review and conduct agreement assessments
	• Create basic flow diagram prior to linkage and set general expectations
	• Link full birth cohort (all PRAMS years or subset) with administrative source
	  ○ Merge linked file back to full birth cohort file (include match scores if used)
	  ○ Describe linkage rates overall and by subpopulations (e.g., race, region, sex, income, and age)
	  ○ Evaluate for any bias in linkages

	○ Subset to PRAMS cohort and compare weighted estimated linkage rates to full birth cohort (pay particular attention to the PRAMS sampling strata variables) ● Link PRAMS with administrative source (if the full birth cohort wasn't linked for all years) ○ Merge linked file back to PRAMS file (include match scores if used) ○ Describe linkage rates, unweighted and weighted, and by demographic characteristics ○ Evaluate for any bias in linkages (pay particular attention to the PRAMS sampling strata variables)
PHASE IV: Research Dataset Creation	● Create research dataset(s) ○ Create an analysis plan ○ Organize linked data through query or other data wrangling techniques to identify health outcomes for linked participants ○ Ensure data are de-duplicated if applicable (e.g., the data had a one-to-many relationship between PRAMS respondent and source[s] linked) ○ Limit PRAMS elements to variables of interest and ensure that identifiers are removed ○ If relational, merge source outcomes identified with PRAMS elements using the linked common ID ○ Verify sample size (final analysis dataset should match the total PRAMS respondent file for the years linked)
PHASE V: Analysis	● Conduct appropriate analyses, perform data consistency checks, and additional benchmarking if necessary ○ Investigate potential bias in results due to differential linkage rates, differences in emigration that could impact detection of outcomes in source(s) linked, and record in summary tables any difference in linkage rates of match probabilities (particularly if comparing by racial groups) ● Summarize results and generate report(s) ○ If peer-review manuscript developed, follow the GUILD guidance for reporting linked studies
PHASE VI: Sustainability	● Establish and refine data use policy to support use by external researchers ● Ensure data are securely stored, process is documented, and a data map with data books for each data set are available ● Ensure data and process can be replicated and documented

*Note: When developing a linkage purpose, keep in mind that PRAMS is a sample. Depending on your individual states' sampling fractions and sampling strata, the types of outcomes (for the mother, child, or father) that can be investigated may be limited. Rare outcomes may have a low probability of being included in the linked data set. CDC guidance on necessary sample size for reporting PRAMS responses should be followed for linked outcomes as well. Additionally, for outcomes that are associated with a sampling stratum (e.g., a sampling stratum of low birth weight and outcome of infant mortality) that may also be influenced by differential reporting (e.g., mothers of LBW babies that may be more likely to respond and recall events in their history) can result in biased results. Therefore, it is important to include in the defined purpose a specification on the hypothesized potential influence of these on estimates.

Appendix E. Template Data Use Agreement for Public Health Data Linkages

This Data Use Agreement ("Agreement") is entered into between [Name of First Agency], located at [Address of First Agency], hereinafter referred to as the "Data Provider," and [Name of Second Agency], located at [Address of Second Agency], hereinafter referred to as the "Data Recipient."
Effective Date: [Date]

Purpose and Scope:
Describe the purpose and necessity of the data linkages. This section should generally outline the legal authority to collect and use identifiable data and specifies that the agreement governs/outlines the use, sharing, and linkages of the data. It should also specify the datasets to be linked and provide a general description of what they include as well as what elements will be used within the data.

Data Access and Use:
Typically, this section describes the following:

- *Data Provider agrees to provide access to the specified datasets to Data Recipient for the purpose of conducting data linkage and analysis.*
- *Data Recipient agrees to use the data solely for the purpose stated and to comply with all applicable laws and regulations.*
- *Data Recipient shall not use the data for any unauthorized purpose.*
- *Either specifically or broadly individuals that will have access to the data (define who is an "authorized user").*

Data Security and Confidentiality:
Outline the required security and measures to be taken to ensure confidentiality, limitations on access, and that the minimum necessary elements are used. This section can also include descriptions of how the data will be secured in transit and at rest, what tools will be used to share the data, and the process for sharing the data. Typically, both parties take responsible measures, use appropriate/approved technology, and abide by organizational safeguards. Nearly all agreements include the explicit language that "Data Recipient shall restrict access to the data to authorized personnel only," or something similar in this section.

Privacy and Legal Compliance:
Outline how both parties shall comply with all relevant data protection and privacy laws and regulations – specify laws/regulations. In addition, describe what the Data Recipient will do in the event of a data breach or unauthorized disclosure. Usually, language is included on how the Data Recipient will address the breach and the duration of time to notify the Data Provider and what actions, as required by law, will be taken.

Data Linkage and De-identification:
Describe the linkage and de-identification methods and what tools will be used (especially if the tool uses external servers or web-based tools). This section should detail (usually in generalities) the elements (identifiers) used for linkages and how the research data will be de-identified to meet organizational policies. When required, language related to re-identification of de-identified research datasets would be included in this section.

Data Sharing and Reporting:
Describe how (the process) and when the Data Recipient may share de-identified and aggregated results of data analysis with other parties for research or public health purposes. Any shared results or datasets will generally not contain personally identifiable information, but provisions for such sharing should be specified here (e.g., under an approved IRB and approved data use agreement signed by all data stewards).

Ownership and Intellectual Property:
Describe original raw data and information products derived from the linked data. In state-based agreements, this usually also includes provision for the Data Provider to review information products using the linked data, duration of time for the review, and a method for reconciliation if disagreement occurs. This section should also include provisions for how the Data Recipient can share de-identified data with other internal/external researchers. Most often, the Data Provider retains ownership of the original data, and the Data Recipient acknowledges this ownership. The Data Recipient retains intellectual property rights related to analysis and research conducted using the data.

Duration and Termination:
Outline the length of the agreement, method for termination, and requirements of use of the data after termination. In general, when possible, open-ended opposed to term limits are preferred, and provisions for continued use of the linked data with destruction of raw data or identifiers used to conduct linkages destroyed.

Governing Law and Jurisdiction:
This section may be required if not already clearly described and may be included regardless to be explicit about the legal authority of both the Data Provider to share the data, and for the Data Recipient to receive the data. This is often a list or short description with links to applicable laws, regulations, and policies.

Amendments and Modifications:
This section can detail how amendments will be addressed, for example. Any amendments or modifications to this Agreement shall be in writing and signed by authorized representatives of both parties.

Signature:
[Signature of Authorized Representative of Data Provider] [Signature of Authorized Representative of Data Recipient]
[Name of Authorized Representative of Data Provider] [Name of Authorized Representative of Data Recipient]
[Title of Authorized Representative of Data Provider] [Title of Authorized Representative of Data Recipient]
[Date] [Date]

Appendix F. Basic template for documenting the PRAMS data linkage protocol/plan.

Short Informative Title
Short sub-title

Author Date

Background
[3-4 paragraphs that provide details on the motivation for the linkage project. This should provide enough background to understand the project's purpose.]

Purpose
[One sentence statement on the overarching goal of the project.]

Objectives
[Clearly stated research questions. This should be as specific as possible.]

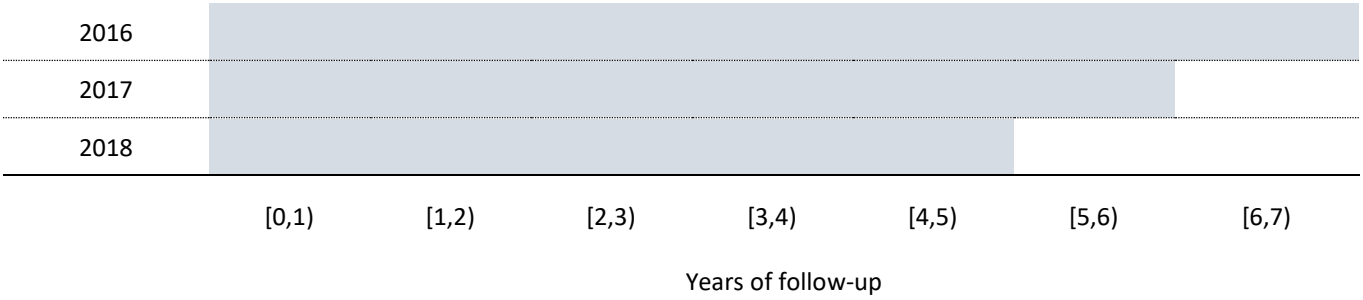
Research Questions
[Optional but recommended for projects that are anticipated to be ongoing. Questions should be grouped by anticipated short-, mid-, and long-term analyses.]

Data infrastructure
[Brief description (1-2 paragraphs) outlining any prior linkage work and existing infrastructure that will assist or limit the linkage activities.]

Population

[Describe the population to be linked, including the specified PRAMS years and years of the administrative data to be used, and if any/all full birth cohorts will be linked. If applicable, describe the length of observation from birth (or before) each birth cohort will be followed (e.g., Figure 1). Figures make it easy to see that the entire birth cohort may be detected in the administrative source for just under 5 years of follow-up, with the 2016 cohort having just under 7 years.]

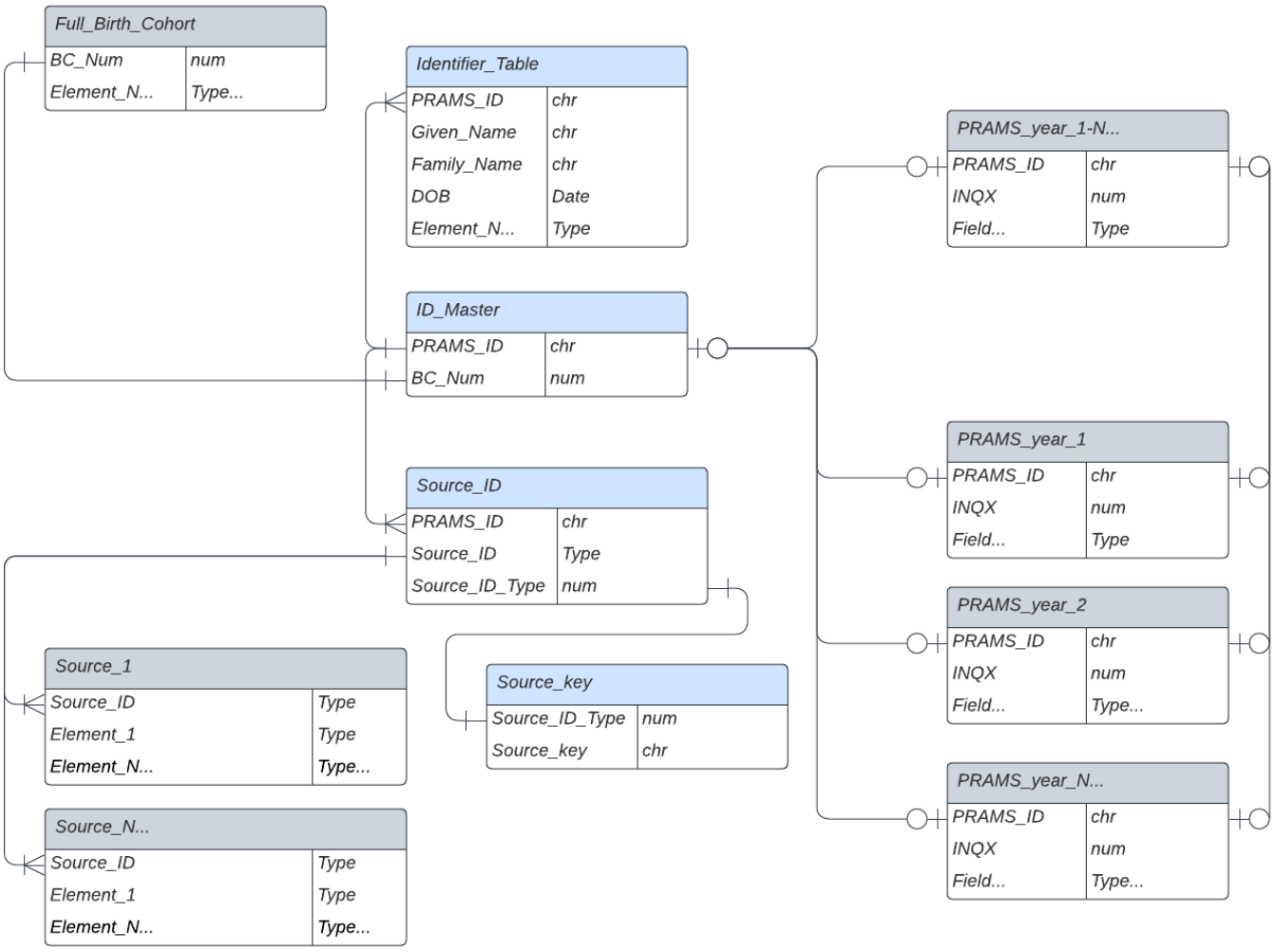
Figure 1. Years (age-scale) of follow-up among the 2016-2018 PRAMS cohorts linked with administrative records through 2023.



Planned data structure

[The recommended relational structure that balances normalization and utility and provides flexibility for future and ongoing linkages with individual adaptations should be described here. If some other structure is used that should be clearly articulated with a visual depiction provided. This ensures a clear target for how the linked data will come and stay together over time. See Figure 2 that depicts the generalized recommended data structure. Depending on source restrictions, the data flow from system extraction to linkage and then to storage within the structure proposed will vary.]

Figure 2. Planned data structure for final linkage data for the 2016 - 2018 PRAMS birth cohorts.



Agreements and approvals

[A list of all existing agreements and ones that are needed, the process for completing them, key contacts if available, expiration dates, and other relevant information.]

Linkage plan

[A short description of how the actual linkages will be accomplished. Supporting the description, a simple outline should be included and look something like the following:
The following process will be used.

1. Ensure the data source has identifiers that will facilitate high-quality data linkages
2. From the source data request only the identifiers with the system ID (limited identifier set)*
3. Merge PRAMS with birth records to obtain identifiers (names, dob, sex, location, race, birth weight…)

4. Link birth records with the source limited identifier set
5. Update outlined data structure
6. Request specified data elements for only those that are linked using the source ID only*

*Note: The source may choose to send all the information at once. Ideally, the identifiers and data elements will still be separated from the linked data to ensure that data from multiple systems are only connected through the PRAMS-ID.]

Tools

[Short description of the linkage tool used, any source code needed, cost, methods for renewal, or any other relevant information.]

Methods

[Short description of the anticipated linkage methods. This should include if you're using deterministic, rule-based, probabilistic, machine learning, or some combination. Generally most successful linkages are iterative, so this section may need to be fluid, but you should start with what is anticipated and then document the resulting method used. Include whether manual reviews will be completed or not, how thresholds will be set, if data cleaning will be conducted (with links to source code), and how errors will be handled or assessed.]

Validation

[A brief description of how you will validate the linkages and PRAMS estimates. Tables 1 and 2 below should be completed to support this work. The general tools that will be used for linkage validation and assessment of false positives, false negatives, and overall accuracy should be briefly described. Any prior linkage work could greatly inform this section. Table 1 documents linkage results and Table 2 compares the PRAMS results against the full birth cohort. Once the data are linked, a research data set can be constructed. This research data will include all the PRAMS/BC records and the outcomes of interest derived from the linked source. Ensure that no duplicate records are added or lost during this process.]

Table 1. The percentage of PRAMS records that were successfully linked with each administrative data source record, overall (all years of PRAMS linked), and by each PRAMS year.

Source Data	**# of PRAMS Records**	**# of Source Records**	**# of PRAMS Records Linked with Source Records***	**% of PRAMS Records Linked to Source Data**
Insert name of administrative data source linked with PRAMS]				
Overall (2016 - 2018)				

2016				
2017				
2018				

Notes:
*If relevant, specify the total PRAMS records linked and include the breakdown of the number linked deterministically and those detected through probabilistic determination or manual review (e.g., If there are 1000 PRAMS respondents and 900 of them were linked using full name, DOB, and sex using a basic edit distance approach, one would report the number linked (900) and the number with exact matches (n1) and those detected through manual review (n2), where n1 + n2 = 900.)

Table 2. PRAMS-linked outcome validation against birth cohort.

	Unweighted PRAMS %	**Weighted PRAMS % (95% CI)**	**Full Birth Cohort %**
Source 1*			
outcome 1			
outcome 2			
outcome n.			
Source 2*			
outcome 1			
outcome n.			
Source n.*			
outcome n.			

*Report the proportion that linked to the source (for the same years)

Analysis and Reporting

[A brief description (1-2 paragraphs) on what analyses and statistical tests will be conducted, and how these data/results will be shared, including what information products will be produced.]

Sustainability

[A brief 1-2 paragraphs describing some key considerations for enhancing or ensuring sustainability. This may include network paths to files and code, and any approvals, key contacts, and notes related to funding. This will likely result in an evolving section and should be adaptable. Teams should also maintain key information related to each source linked (Table 3) and update annually for ongoing linkage projects.]

Table 3. Key Information for Linked PRAMS Data

Source data	**Agency that owns these data and contact info**	**Permission required to link data[1]**	**Type of access to these data[2]**	**Years of data available[3]**	**Is IRB approval required for this source? [Y/N]**	**Are other internal approvals required? [Y/N]**
Vital Birth Records						
[Insert names of administrative data source(s) linked. Add rows for each additional source]						

Notes:
1. Describe the type of permissions required to obtain and link the data with PRAMS (e.g., MOU, MOA, DUA, internal document, etc.).
2. Is direct access to the data source available to extract and batch download individual records? Is a pre-defined record level dataset with a subset of data elements contained in the entire system, or some other level of data access?
3. Include all years of data available for the linkage project.

Appendix G. Additional data linkage resources.

Frameworks and Toolkits

- Linking Data for Health Services Research: A Framework and Instructional Guide
- CSTE Injury Data Linkage Toolkit
- CSTE Tribal Epidemiology Toolkit
- DaSy Center Data Linking Toolkit: Steps to Data Linking
- NAHDO Data Enhancement and Linkage
- Race and Ethnicity Data Improvement Toolkit
- UK Office for National Statistics - Developing standard tools for data linkage: February 2021
- UK Longitudinal data linkages
- Queensland data linkage framework
- Census Bureau data linkage policy
- The Linkage of the National Center for Health Statistics (NCHS) Survey Data to U.S. Department of Housing and Urban Development (HUD) Administrative Data: Linkage Methodology and Analytic Considerations
- NCHS data linkage program

Linkage Methodology and Tools

- GitHub: Data Matching Software
- Pareto Distribution
- Chimera open-source machine learning tool
- OpenRefine: open-source data cleaning and transformation tool
- Australian Institute of Health and Welfare data linkage
- Privacy preserving record linkage

Linkage Quality Assessment

- Quality and Complexity Measures for Data Linkage and Deduplication
- Data Linkage: A powerful research tool with potential problems
- Comparing record linkage software programs and algorithms using real-world data
- A guide to evaluating linkage quality for the analysis of linked data
- Challenges in administrative data linkage for research
- Assessing data linkage quality in cohort studies

Trainings/Organizations

- The International Population Data Linkage Network
- Record linkage training (4-part series) by Peter Christen: Part 1, Part 2, Part 3, Part 4
- CSTE Powerpoint on data linkage
- Linkage example using R RecordLinkage package