Linking State and Federal Data

Use Case Report (SFPE-25-N7)

Jeremy Seeman, Devlin Hanson, Judah Axelrod, Rebecca Berger, Jordan Hardenburgh, Evy Park, and Ridhi Purohit

August 2026

America's DataHub Consortium (ADC), a public-private partnership, implements research opportunities that support the strategic objectives of the National Center for Science and Engineering Statistics (NCSES) within the U.S. National Science Foundation (NSF). These results document research funded through ADC and are being shared to inform interested parties of ongoing activities and to encourage further discussion. Any opinions, findings, conclusions, or recommendations expressed in this report do not necessarily reflect the views of NCSES or NSF. Please send questions to ncsesweb@nsf.gov. NCSES has reviewed this product for unauthorized disclosure of confidential information and approved its release (NCSES-DRN26-059).

This report is intended to serve as a guide to government evaluation offices across federal agencies to help them streamline federal-state data linkages for future evaluations. The goal of this project was to determine how a National Secure Data Service (NSDS) could facilitate data linkages between federal and state entities for evaluation purposes given the persistent challenges with federal-state linkages. The project included a use case to test federal-state linkage of state data from the Child Care and Development Fund (CCDF) on families receiving child care subsidies with federal data from the US Department of Housing and Urban Development (HUD) Public and Indian Housing Information Center on families receiving housing assistance. The use case tested an innovative linking method called privacy preserving record linkage (PPRL) with the NSDS serving as a trusted third party. This report provides a road map for conducting federal-state data linkages for government evaluations, summarizes the critical lessons learned from this project, and provides recommendations for future state-federal data partnerships.

Introduction

Administrative data are critical to conducting rigorous program evaluations. Federal and state agencies maintain extensive administrative data—information collected through routine operations, registrations, and transactions—that can be leveraged to measure program utilization, performance, and outcomes. Administrative records provide relatively accurate data on program utilization—especially when tied to benefit payments or eligibility determinations—because they are collected for operational purposes rather than self-reported by recipients of program benefits or staff. Administrative systems are often designed to collect information from all program recipients. Large-scale administrative data allow for more precise estimates and greater flexibility in research design. This is particularly valuable for quasi-experimental methods that require substantial statistical power to produce valid results, such as propensity score matching, regression discontinuity designs, and difference-in-differences analyses. In addition, using existing administrative records can reduce the need to collect new survey data, thereby lowering respondent burden and reducing data collection costs. However, rigorous and comprehensive program evaluations can require administrative data stored across different entities. Evaluators often need to navigate decentralized state and federal program data systems.

Some research questions can only be answered by linking data across systems. Linking administrative data from federal and state systems allows researchers and policymakers to evaluate federal programs that capture outcomes in state systems and vice versa. These linkages also help researchers identify coenrollment across programs, allowing federal and state governments to improve coordination, increase efficiency, and strengthen the social safety net.

Despite these benefits, federal-state data linkages remain relatively uncommon because linking these data requires navigating many logistical, legal, and technical challenges. Complicated legal protections and administrative procedures are often the most difficult to manage. Linking two disparate datasets requires some form of personally identifiable information (PII). States are required to abide by both federal- and state-specific data privacy laws in addition to state- and agency-specific policies. Also, state agencies may be apprehensive about sharing record-level information because it

involves institutional, legal, or reputational risk. Even if these challenges are addressed, federal and state agencies might not have the shared interests, incentives, or resources needed to mutually benefit from participating in linkage. Two parties may be entirely willing to share data but are blocked by human capacity constraints, incompatible technical infrastructures, or unsuitable data quality. The NSDS has the potential to address many of these persistent challenges that have delayed and prohibited federal-state linkages in the past.

National Secure Data Service

The NSDS Demonstration is a governmentwide effort to strengthen data integration and data access infrastructure to facilitate evaluation and statistical analysis that can inform decisionmaking at all levels of government and in all sectors. The future vision is that NSDS would provide shared services for data and evidence building and potentially facilitate federal-to-state data linkages. To meet this goal, the NSDS Demonstration is testing and piloting various services to help facilitate data linkage and access.

Project Overview

This project was funded by the National Science Foundation as part of the America's DataHub Consortium. The foundation contracted with the Urban Institute to investigate how an NSDS could facilitate data linkages between federal and state entities for evaluation purposes. HUD's evaluation office partnered with NSDS to serve as a test case for some of the tools that NSDS is developing. This project aimed to assess the role of NSDS in serving as (1) a third-party intermediary facilitating data linkage and (2) the capacity of the secure compute environment (SCE) NSDS is developing as a core service to handle data linkage using PPRL.

Our project had three primary components. The first involved collecting and analyzing qualitative data to better understand past projects that aimed to link HUD and state administrative data. The qualitative data included a review of available documents on past linkages between HUD and state data systems and a limited set of semistructured interviews with HUD, state, and evaluation personnel who were involved in these linkages to glean relevant insights for broader applications of federal-state linkages. Additional insights from that work are reviewed in the final lessons learned report.

The second component was a use case that planned to link state CCDF data from two states to HUD's tenant data using PPRL. This use case fit the HUD evaluation office's goal to examine the overlap between HUD-assisted households and households receiving CCDF child care subsidies. Because CCDF child care data are available only at the state level, this overlap could best be measured through a state-federal linked dataset.[1] Data from one state would be used to develop and test end-to-end infrastructure and processes for conducting federal-state data linking. Data from a second state would then be used to assess replicability of the process created for the first state. The third component planned to analyze the linked dataset to examine the overlap between HUD voucher recipient households and households receiving CCDF subsidies.

Because of project delays, including a lapse in appropriations, we were unable to complete the second and third components of the project within the 10-month project time frame. We successfully completed the first component and the initial steps of the second component, the use case. We recruited one state CCDF agency to participate, developed most of the technical and legal frameworks for data linking, and drafted an initial data-sharing agreement. Although we were unable to execute the data linkage, we drafted preprocessing and linking code and tested it on simulated data. This code can serve as a foundation for entities interested in applying or learning more about PPRL.

Ultimately, the project suffered from many of the same administrative burdens that we were tasked to study. Developing and finalizing the data-sharing agreements proved complex, and completing the process during the project period became infeasible. This highlights that more efficient processes, regardless of the technical or policy dimensions underlying any linkage project, are needed to notably reduce administrative burden.

This report provides a road map designed to guide government evaluation offices interested in linking state and federal data. As mentioned, a core component of the use case was to test a novel method of linking data, PPRL, using the NSDS SCE. We begin by explaining PPRL, including its benefits and limitations and the role NSDS SCE can play in facilitating PPRL. Next, we outline the road map by describing the key phases involved in a federal-state data linkage, which include recruiting states to participate and addressing key questions related to developing the technical and legal designs. At each step, we describe the general process, highlight which components are specific to PPRL and NSDS SCE, and present our use case as an example. Although we were unable to preprocess and link actual HUD and CCDF data, we present the practical implications of using PPRL to conduct data linkage from our simulation study. Finally, we conclude with lessons learned that could guide NSDS as it builds infrastructure to help federal program evaluation offices link state and federal data.

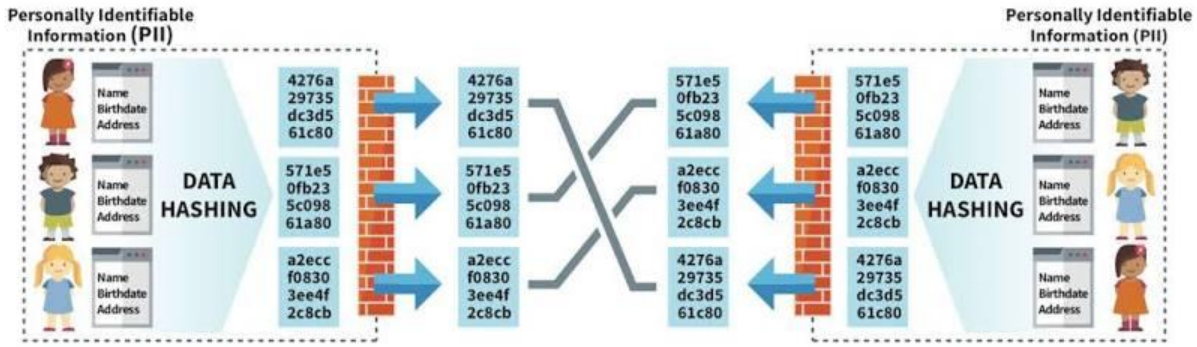
Privacy Preserving Record Linkage

Privacy protection has long been a significant barrier to linking data between different agencies. Federal and state agencies are hesitant to share PII (e.g., Social Security number, full names) because of concerns about violating privacy laws or policies and losing the trust of those they serve. An emerging technique that mitigates this privacy concern—PPRL—allows agencies to link data without sharing human-readable PII to outside parties. As illustrated in figure 1, the process starts with two or more datasets containing PII. The PII from each dataset is transformed into a non-human-readable format using a cryptographic method (hashing), which converts PII into long, string-based representations. Hashing is designed to be a one-way function; that is, given an input (e.g., a name or Social Security number), it is trivial to compute an output (e.g., a secure hash), but the inverse (starting with a secure output and using it to find the original input) is designed to be difficult. Hashing functions that produce more complex outputs typically have greater security protections, as more complex outputs are harder to reverse engineer. Security can be further enhanced through salting, which involves adding a unique random string of characters to PII before hashing. When the same salt and hashing function are applied across datasets, the transformed values remain consistent, enabling

accurate matching. After PII from each data source is hashed, datasets can be linked without sharing human-readable PII.

FIGURE 1

Visual Illustration of Privacy Preserving Record-Linkage Process



Source: Stacy J. Peterson, Robert D. Lieberthal, Keith J. Miller, and Niyati H. Vakil, *Privacy Preserving Record Linkage (PPRL) Strategy and Recommendations* (McLean, VA: MITRE, 2023), https://www.nia.nih.gov/sites/default/files/2023-08/pprl-linkage-strategies-preliminary-report.pdf.

PPRL may mitigate legal and privacy concerns involved in linking data, but it also has several limitations. PPRL limits the linkage methodologies that can be used by restricting how linked records may be assessed. Every linkage process involves modeling decisions about how related PII fields imply linkage. These decisions may be deterministic (e.g., two records from disparate sources match if they have the same Social Security number), or probabilistic (e.g., two records from disparate sources that have the same first name, last name, and date of birth have a 95 percent probability of representing the same person, but if the records have only the same fist initial, last name, and year of birth, they have only an estimated 60 percent probability of representing the same person). PPRL can only assess whether two fields from disparate sources match or do not match, not the degree to which they match. For example, the hashes for full names "John Smith" and "Jane Doe" do not match, but the hashes for "John Smith" and "Jon Smith" also do not match. Neither example would be considered a match with PPRL, but standard probabilistic models would have been able to assess the degree of similarity in the latter example and may have been able to match it.

Building probabilistic record-linkage models with PPRL is still possible. A probabilistic record-linkage model with PPRL is based on using different combinations of hashed PII. PPRL still assesses variable matches in a binary manner (i.e., two fields must match 100 percent or they do not match). While the use of exact matches within a probabilistic model may seem confusing, the interpretation would work like this in practice: if hashed first names match exactly, this increases the overall likelihood that these records are a match by 25 percent. In other words, the binary comparisons of each variable are added to estimate an overall match probability.

When data contain errors, inconsistencies, or variations, PPRL may miss a substantial number of true matches. The extent to which PPRL accurately matches across datasets determines its usefulness for evaluation. If the matches are inaccurate, any downstream analysis on the linked data will be inaccurate. This implies that PPRL is less appropriate for program administration, such as verifying eligibility or delivering services, and may be more suited for research and evaluation.

Although PII is transformed through hashing, some risk of reidentification remains, depending on how the resulting linked dataset is used or disseminated. PPRL addresses the "input privacy" problem, which prevents parties from sharing PII at the point of creating a linked dataset. But it does not solve the "output privacy" problem, which is the risk that output data products reveal information about individuals. This output privacy risk is especially high for PPRL, because each data-sharing partner has PII that could be used with the linked dataset to reidentify individuals contained in multiple datasets. Data-sharing partners could attempt to reverse engineer PII—for example, by using the salt key, or even without it, by combining enough indirectly identifiable variables to infer an individual's identity. Therefore, PPRL requires disclosure avoidance policies (i.e., technical and procedural measures used to prevent the accidental exposure or reidentification of sensitive personal data) for any statistics derived from the linked dataset. These critical procedures mitigate the risk that the careful privacy protections enforced during PPRL cannot be easily undone by data contributors or any third party using the linked dataset.

The Role of the National Secure Data Service in Privacy Preserving Record Linkage

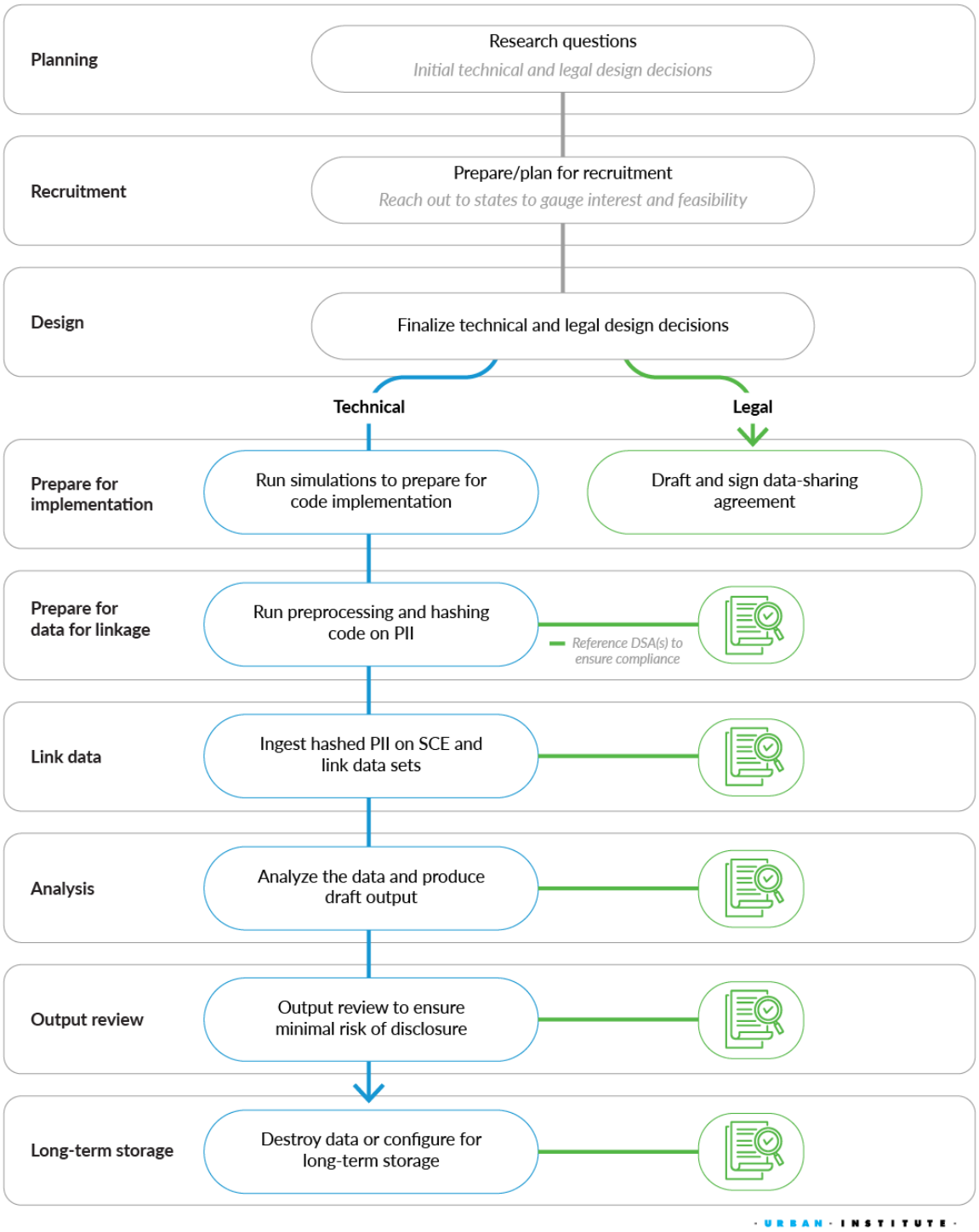
NSDS's goal is to serve as a one-stop shop for creating and analyzing linked data sources in a secure manner. In the context of data sharing using PPRL, NSDS can serve as a trusted third party and provide a secure compute environment for linking, analyzing, and storing data under the assumption that a trusted third party will not maliciously attempt to undo any part of the deidentifying process or attempt to reidentify the data. The NSDS filled the role of a trusted third party for our project, and in developing the roadmap we assumed that any analysis of the linked dataset would occur within the same computing infrastructure where the linkage occurred. This assumption aligns with NSDS's goal to serve as a comprehensive system for creating and analyzing linked data sources in a secure manner, with analyzing PPRL data as one potential service NSDS could offer. The NSDS SCE provides a restricted governmentwide data-sharing framework, which allows approved researchers to securely access, link, and analyze restricted federal and state data for evidence building.

Road Map for Using PPRL to Match State and Federal Data

We used what we learned from this project to develop a road map to guide federal evaluation offices interested in federal-state linkages, specifically those focused on applying PPRL within the NSDS SCE. We distinguish between information that is relevant to all federal-state linkage processes and what is relevant specifically to PPRL. The road map outlines each key step in the process, including developing the research concept, identifying data requirements, recruiting state partners, making technical and legal design decisions, preparing for implementation, preprocessing and hashing data, conducting data ingestion and linkage within NSDS, and performing analysis. Figure 2 presents a high-level overview of the process and key steps required to complete a federal-state data linkage. We describe each step in the process in detail below. For each step, we also include a box explaining how it was implemented in our use case. For steps that we were unable to complete in the project period, we provide details on how we planned to complete those steps.

FIGURE 2

Workflow for a State-Federal Data-Linking Project Using Privacy Preserving Record Linkage and the National Secure Data Service Secure Compute Environment



Source: Authors' analysis.

Note: DSA = data-sharing agreement; PII = personally identifiable information; SCE = secure compute environment.

1. Planning

Every research project requires a planning phase to develop the scope and objectives. Clearly defined research questions are the cornerstone of research and provide a guide for identifying the specific data needed from state and federal agencies. This planning aligns the project's research goals with two data governance goals: data minimization (sharing only the data necessary to answer a research question) and purpose limitation (ensuring that any shared data are prohibitively difficult to use for nonresearch purposes). The planning phase includes making initial technical and legal design decisions based on information from existing project partners. If the federal data partner is already involved at project onset, they should answer as many of these questions from their perspective as possible to help expedite the legal and technical design phases and to help determine the minimum requirements for states.

BOX 1

Planning Phase: Use Case

The primary research questions for this linkage were developed in consultation with HUD based on their desire to identify the overlap of households receiving HUD assistance with those also receiving CCDF child care subsidies and to examine the characteristics of those who did or did not overlap.

2. Recruitment

Federal-state data linkages initiated by federal evaluation offices will need to recruit states to participate as data partners, but recruiting states to share data with the federal government can be difficult. State agencies may be reluctant to participate in federal-state data linkages for several reasons, including limited staff capacity, legal considerations, and, in the case of PPRL, uncertainty about the security of methods that may be unfamiliar to their legal and technical staff. Below, we outline recommended steps for recruiting states to participate in this type of project.

2.1. Determining Which States to Recruit and Developing a Recruitment Plan

Federal-state data linkage projects will first need to identify how many and which states to recruit. The answer will depend on project goals, including geographic and sample size requirements, whether the data need to be representative to answer the project's research questions, and the subpopulations the project is interested in studying. In addition, a recruitment plan with a well-defined outreach process and timeline can ensure that recruitment is conducted in a timely and uniform manner.

2.2. Finding the Right Contact

Identifying the right initial contact is the first step. Ideally, this person should benefit directly from the project and be willing to serve as a champion—someone willing to shepherd the project through all the legal and technical processes within the agency. Potential initial contacts may include the following:

- Program staff who would benefit from answers to evaluation questions
- Evaluation staff interested in the analysis
- Evaluation or technical staff interested in learning to use PPRL
- External partners (e.g., local researchers) who have strong relationships with the agency

Finding the right contact can be challenging. Many state agencies do not publish up-to-date organizational charts or staff contact information online. Even when this information is available, it may be outdated or otherwise not reflective of a person's roles and responsibilities.

Starting with someone in the state agency who has an existing connection with the project team or federal agency—even if they will not serve as the project's champion—can be an effective strategy. These individuals can often help identify the appropriate person and provide an introduction. Contacting states with no prior relationship or common contact, particularly via cold email, is less effective because messages may be filtered as spam or ignored by recipients unfamiliar with the sender. Whenever possible, outreach should be conducted through existing relationships.

2.3. Performing Initial Outreach

The initial outreach should be via an email that provides the key information a state needs to decide whether to engage further. At a minimum, the outreach should clearly communicate the following:

- The goal of the project
- The benefits of participation for the state
- The extent of participation required from the state
- A concise plain-language explanation of the process of safeguarding potentially sensitive data, including PII

If the initial outreach does not yield a response, friendly follow-up emails are a next step. Timing and spacing of follow-up should balance respect for staff boundaries with maximizing the chances of successfully making contact.

2.4. Scheduling the Initial Call

The initial outreach email should include a request for a follow-up meeting or call to discuss the project in more detail. The goal is fourfold:

1. **Ensure project comprehension.** It is critical that the initial call successfully establish foundational information with state staff attending the call. They may be the ones communicating the project information to other staff across the state agency (e.g., legal, leadership, IT/tech). Providing a one- or two-page summary that describes the project can help ensure that details are communicated accurately with other staff at the agency. For PPRL and NSDS, the summary should include plain-language explanations of the technique, the organization, and how they protect PII, ideally including visual aids, such as figure 1.

2. **Highlight benefits and address concerns.** Because states have limited capacity to participate in projects, it is important to highlight the benefits (e.g., answering programmatic or evaluation questions, building familiarity with PPRL for potential future use). It is also critical to be up front about what will be asked of staff and how, if at all, the burden can be mitigated. Participating in any kind of data sharing and linkage project can impose a high burden as it requires substantial staff time across multiple roles. To reap the full security benefits of PPRL, state and federal IT or evaluation staff must prepare and hash data themselves before uploading to a secure location such as NSDS. Trusted partners can write the code and documentation, but staff must carry out these instructions.
3. **Determine feasibility.** Not all state data will be suitable for every federal-state data linkage project or be suitable for PPRL. Suitability depends on whether the state agency has the identifiers needed to match the federal data and the data elements needed to answer the research questions. In addition to understanding data availability, it is important to understand data quality and coverage. If the desired administrative data are a priori not fit for linkage or research purposes, the linkage will not help address the research questions. Additionally, in the context of PPRL, it is important to understand whether the agency has the technical capacity and ability to preprocess or hash the PII. Agencies regularly face restrictions on hardware or software that can be used on their data and the availability of technical staff who could run the preprocessing and hashing code.
4. **Learn about data-sharing processes and steps.** Finally, it is important to understand the reviews and approvals that must be completed before the state agency can participate in the project, and how long that approval process is likely to take. State agency approval processes for data sharing differ significantly, so facilitators should understand up front which staff at the agency need to provide approval (e.g., leadership, IT, tech, legal, IRB/research review committee), and in what order. Obtaining a realistic estimate of how long each approval process takes is critical, particularly when key decisionmakers have limited availability.

Staff who attend this preliminary meeting can have varying degrees of knowledge and responsibilities related to data linkage and may not be able to answer questions or make decisions regarding the feasibility of the project. During the meeting, it is important to identify the best points of contact to move the conversation forward. The initial meeting is just one of a series, but it is critical for opening a line of communication and building a partnership.

2.5. Establishing Regular Communication

Once a state has expressed interest in moving forward, consistent and proactive communication is essential for advancing a project through the design phase, approvals, and data-sharing agreements. Early discussions should focus on identifying the steps necessary to access state data as well as expected time frames for each step. In addition, early discussions should attempt to identify any potential challenges the project may face at each step. Regular meetings and check-ins can ensure that processes are not delayed and that any issues are handled quickly so the project does not lose momentum. During these meetings, topics for discussion should include steps completed, next steps, changes to the timeline, and any new challenges and potential solutions.

BOX 2

Recruitment: Use Case

Our recruitment process followed the steps outlined above. First, we developed a recruitment plan and identified states for outreach. Given our short timeline, our plan and identification focused on recruiting two states as quickly as possible. Because our use case planned to recruit only two states, we focused on a convenience sample.

We identified appropriate contacts for states through the publicly available website listing all CCDF lead agency contacts. When possible, we used an introduction. We reached out to 20 states over two months. Of the 20 states contacted, 13 declined to participate, 6 never responded or stopped responding to outreach, and 1 agreed to move forward. Five states initially expressed some interest, but four of these later declined to participate. State agency staff who showed interest wanted to learn more about PPRL, because they could use the knowledge later for other data-linking projects and because they supported data sharing and evaluation more generally. The primary reasons for declining were capacity constraints and the project not aligning with the agency's strategic priorities.

3. Design

Before implementing the data linkage or data-sharing agreements, key decisions regarding technical and legal architecture must be made. Technical design consists of the logistical and functional details of how the linking will be implemented. Legal design consists of terms that both parties must agree on to allow data sharing and linking. Legal and technical choices do not operate in isolation; decisions in one area frequently inform the other. However, the choices often engage different types of staff. For example, IT and evaluation staff are often consulted for technical design, while legal staff are consulted for legal design. For this reason, we recommend working on both tracks concurrently, as applicable, to save time. This allows facilitators to serve in a translational role, whereby the ways legal and technical requirements inform one another are made clear to all relevant parties.

The following sections outline key technical and legal design choices to consider when preparing for a data linkage between state and federal entities. These design choices should guide conversations with the state and federal agencies that will be involved in the data-linking project.

3.1. Technical Design Choices

Federal-state data linkage projects can vary widely in their design based on the goals and limitations of both the federal and state partners. Partners need to work together to make technical design decisions. Table 1 outlines the questions that need to be answered to develop the technical design. Details are provided below the table. While some questions are universal to all federal-state data linkages, others are specific to PPRL implementation with a trusted third party (these are noted in the table). For each question, we provide the answers from our project's use case.

TABLE 1

Technical Design Choices for Using Privacy Preserving Record Linkage in State-to-Federal Data Linkages

Questions	Use case
Who?	
What entities/parties have ownership of the data and will be providing the data to be hashed (e.g., state or federal agencies)?	HUD and state CCDF agency
PPRL specific: Who is the trusted third-party entity providing the SCE for matching the hashed data and protecting output privacy?	NSDS
PPRL specific: Who will conduct or facilitate the hashing? Who will generate the code to hash the PII and provide technical assistance to the parties involved?	Urban Institute, MRAIA (technical)
Who will conduct the linking?	Urban Institute
Who will be allowed to analyze the matched dataset?	Urban Institute
What?	
What PII will be used for matching?	Names, dates of birth, SSN
PPRL specific: What PII will be hashed?	Names, dates of birth, SSN
What data will be used for analysis?	Demographics, HUD program participation and voucher receipt data, CCDF child care subsidy data
Where?	
Where will the PII be preprocessed (i.e., cleaned and standardized)?	State CCDF agency's PII was to be preprocessed within their systems. HUD's PII was to be securely transferred to Urban, where it would be preprocessed.
PPRL specific: Where will the PII be hashed?	State CCDF agency's PII was to be hashed within their systems. HUD's PII was to be securely transferred to Urban, where it would be hashed.
Where will the data be linked?	NSDS SCE
Where will the linked data be stored?	NSDS SCE
Where will the linked data be analyzed?	NSDS SCE
When?	
When will each step occur? Will any steps occur concurrently?	Simulation was to occur before preprocessing. Preprocessing and hashing for CCDF data was to occur while the data-sharing agreements were reviewed. Preprocessing and hashing for HUD data was to occur after the data license was in place.

Questions	Use case
How?	
How will the data be cleaned and standardized to prepare for linking?	Standard cleaning and standardization, including metaphones for names
PPRL specific: How will the data be hashed and salted (if using a salt)?	Hashing/salting functions within Python
How will the data be transferred?	SFTP
How will the data be linked (e.g., probabilistic record linkage)?	Deterministic match with composite hashes
How will the resulting linked dataset be analyzed? What output products will be produced from the linked dataset?	Undetermined
How will the project prevent information about individuals from being revealed by output data products?	Undetermined
Once these methods are applied, who will have access to the output data products?	Undetermined
How will the linked dataset be stored or destroyed once the analyses are complete and data products are disseminated?	Destruction of all data at the end of the project

Source: Authors' analysis.
Note: CCDF = Child Care Development Fund; HUD = US Department of Housing and Urban Development; MRAIA = Medical Research Analytics and Informatics Alliance; NCSES = National Center for Science and Engineering Statistics; NSDS = National Secure Data Service; PII = personally identifiable information; PPRL = privacy preserving record linkage; SCE = secure compute environment; SFTP = secure file transfer protocol; SSN = Social Security number.

Who? In a federal-state data linkage, the primary parties include one or more federal agencies and one or more state agencies. The federal and state agencies are the data owners and are responsible for providing the data to be linked. For the remainder of this road map, we assume one federal and one state agency are involved. If more than one state or federal agency are involved, projects will likely need to review these decisions separately for each state or federal agency participating in the project, because security requirements, technical capacity, and data structures often vary across agencies. In addition, the project must determine the level of consistency needed for each decision across agencies and coordinate accordingly.

A trusted third party may also be involved to facilitate data linkage. In the context of PPRL, a trusted third party facilitates linkage in a secure manner. The trusted third party must have access to only the minimal information necessary to facilitate PPRL; it must not have access to the information needed to reidentify individuals (such as the salt key). Furthermore, it must enforce strict output privacy protections when hosting and processing hashed PII.

Additional entities may also be involved, such as an evaluator who analyzes the linked dataset or, in the case of PPRL, a PPRL facilitator who either conducts preprocessing and hashing for one of the

data-sharing partners or develops preprocessing and hashing code and provides technical assistance to participating agencies.

What? Before implementation, federal and state agencies must agree on the specific data elements to be shared. It is critical for accurate linkage that PII variables are identical across datasets and formatted consistently. For example, differences in how dates are recorded (e.g., "1/1/2020" versus "January 1, 2020") will result in different hashed values and prevent accurate matching. Agencies must also determine which nonidentifiable variables will be sufficient to answer critical evaluation questions, along with the variables' accuracy and coverage.

Where? It is also necessary to define where each computational step will occur and who will have access to the data at each stage. The first two steps—preprocessing and, in the context of PPRL, hashing PII—can take place within federal or state agency systems, within a trusted third-party environment, or within a separate secure environment operated by another approved partner.

Preprocessing PII is easier to complete within the same environment as the data linking. When using PPRL, conducting preprocessing steps within the originating agency's system is generally the most secure approach because it ensures that no human-readable PII leaves the agency. A PPRL facilitator can provide code and technical assistance to the originating agency to clean and encrypt PII. If agencies are unable to run the code internally, a trusted partner, separate from the trusted third party used for linkage, operating under appropriate legal agreements, may conduct the work. After hashing, the transformed PII and associated nonidentifiable data are transferred to the trusted third party for linkage.

Once the hashed PII has been used to link the data, the hashes derived from the PII should be destroyed, assuming there are no plans to introduce new data that need to be linked to one or more sources. The resulting linked dataset can then be stored and analyzed within the secure environment or, if the linked dataset meets disclosure and privacy standards, shared with authorized users.

When? Many steps in the process must occur in a defined sequence. In the context of PPRL, PII must be preprocessed before hashing, and hashing must occur before linkage. However, federal and state agencies may conduct preprocessing and hashing either sequentially or concurrently. Concurrent processing may reduce the overall timeline, but sequential processing allows lessons learned from one dataset to inform the preparation of the other, potentially improving data quality and match rates.

Throughout the use case process, the Urban Institute simulated preprocessing and hashing processes to prepare for actual data. The project planned to preprocess and hash the state CCDF data while the data-sharing agreements were being finalized because this step did not require the agency to share any data. The HUD PII needed to be preprocessed after the data-sharing agreements were in place because it required transferring PII.

How? Agencies must establish clear protocols for preprocessing, hashing, data transfer, linkage, and output controls. Federal and state agencies both face limitations that must be considered when determining protocols. They may face technical infrastructure limitations, such as restrictions on the

availability and applicability of data processing software and hardware, which must be considered when determining the protocols. They may also face staff limitations, such as insufficient staff capacity. In addition, security limitations often restrict how data are transferred or what level of encryption is necessary for security. These limitations will guide how state-federal linkages can be performed.

Preprocessing: PII must be preprocessed for all federal-state data linkages. This involves correcting errors, removing invalid values (e.g., future birth dates), and standardizing formats across datasets. It may also include restructuring variables—for example, separating date fields into day, month, and year—to improve linkage flexibility. If unhashed PII are being used to link data, preprocessing can occur either before or after transferring data. However, for PPRL, preprocessing should occur prior to data transfer because the hashed PII are not human readable, making it hard to detect inconsistencies after the PII has been transferred. Data analysts should agree in advance on preprocessing standards and the tools or software to be used to ensure consistency.

Hashing: For PPRL, parties must determine how hashing will be conducted, including whether salting will be used and which hashing algorithms and platforms will be employed. Any inconsistencies in how PII are hashed or discrepancies in which salt is used will make it impossible to link the data.

Transfer: Parties must establish secure methods for transferring hashed PII and associated data, such as secure file transfer protocols. While hashed PII is more protected than unhashed PII, it is still important to transfer the data securely.

Linkage: For PPRL, the linkage methodology impacts how the PII is preprocessed and hashed; therefore, parties must agree on the linkage methodology to be used within the trusted third-party environment. The two common techniques for matching disparate datasets are deterministic and probabilistic. Deterministic (or rules-based) record linkage looks for *exact* matches across one or more fields used for matching. Probabilistic record linkage is a statistical approach that considers a wider range of evidence, including partially matching fields, typos, and missingness in estimating match likelihood.[2] Deterministic record linkage has a higher risk of missed matches (i.e., true matches that do not match when linking the data), and probabilistic record linkage has a higher risk of mismatches (i.e., matching two observations that are not true matches). The preferred technique depends on project goals and the quality of the PII that will be matched on.

PPRL limits the types of probabilistic record-linkage techniques that can be used. After PII are hashed, many of the usual tools of probabilistic record linkage are removed for increased data privacy. Probabilistic record linkage relies on measuring the similarity of two observations (i.e., how closely two data fields resemble each other). For example, in standard probabilistic record linkage, if two last names differed by just one letter, we could flag these as having a high fuzzy similarity. However, the concept of similarity does not hold for hashed data; to guarantee security, hashes do not encode *any* meaning from the underlying PII fields, which means that two hashes that differ by just one character do not signify anything about the similarity of their underlying data. One method to account for this limitation uses composite hashes, which generate several variables (or combinations of variables). The

power of composite hashes is that they retain the underlying data privacy guarantees while creating potentially dozens of new variables for which an exact match might signify two records correspond to the same person.

Output controls: Finally, PPRL requires output controls for any statistics derived from the linked dataset to prevent reidentification of PII. Federal and state agencies, along with the trusted third party, must agree on the specific output controls to be implemented.

BOX 3.1

Technical Design: Use Case

Who?

The intended data-sharing parties were HUD and two state CCDF agencies. NCSES was to serve as the trusted third party, responsible for providing the SCE and enforcing output privacy rules. The Urban Institute played several roles. First, Urban was to serve as a trusted third party for HUD and would be responsible for preprocessing the HUD PII. Second, Urban was to serve as the PPRL facilitator by preparing code and providing technical assistance for the state CCDF agency to run on their systems. Third, Urban was to link the datasets within the NSDS SCE. Urban also collaborated with the Medical Research Analytics and Informatics Alliance (MRAIA) to aid in developing the code for pre-processing, hashing, and linking, and with the Public Interest Privacy Center (PIPC) to help facilitate data-sharing agreements. Finally, Urban was to serve as the evaluator analyzing the linked dataset to answer the relevant research questions.

What?

The project planned to collect Social Security number, date of birth, name, and address as potential PII for linking because these were available in both the HUD and CCDF data. The project also planned to collect demographics and housing assistance receipt from HUD data and demographics and child care subsidy receipt data from the CCDF agency for analysis.

Where?

For this project, we used different locations and entities to preprocess and hash HUD and CCDF data. Urban has existing data licenses with HUD and has received identified HUD data in the past. In addition, HUD has technical staff capacity constraints. Therefore, the project planned for HUD to send PII and housing assistance data to Urban, which would preprocess and hash the PII within its secure systems.

In contrast, the state CCDF agency did not want to share unhashed PII and had the staff capacity and technical infrastructure to perform preprocessing and hashing. Therefore, the project planned for CCDF PII preprocessing and hashing to take place within their system. Urban wrote code and planned to provide technical assistance to the CCDF agency staff to facilitate this process. Once the PII were hashed, the data were to be transferred securely to the NCSES NSDS SCE to be linked, analyzed, and stored.

How?

The project planned to use standard data cleaning practices to preprocess the PII (e.g., removing white space and suffixes) and to create a metaphone variable for names, which would create one code for names that sound the same but have different spellings (e.g., John and Jon). In addition, the project planned to create composite hashes to allow additional flexibility in matching. Data would then be linked using probabilistic record linkage. The project ended before we determined output controls for the linked data.

While the technical design choices are being determined, the project should be making legal design choices in tandem because some technical design choices will impact legal design choices and vice versa.

3.2. Legal Design Choices

Table 2 outlines the questions that need to be answered to make key legal design choices that are critical in developing a data-sharing agreement. Details are provided following the table. Some choices are universal to linking federal-state data, and some are specific to PPRL. For each question, we provide the design choices made for the use case. Because the project ended before all legal steps were complete, some choices had not been made. These are designated "undetermined" in the use case column.

TABLE 2

Legal Design Questions for PPRL

	Questions	Use case[a]
Parties	Who are the parties participating in the project? What parties are they entering into agreements with? What are the parties' roles and responsibilities?	HUD, state CCDF agency, NCSES, Urban: The Urban Institute was to receive a data license from HUD. State CCDF agency was to enter into a data-sharing agreement with NCSES, with individuals accessing the data on NSDS signing confidentiality pledges
Linkage purpose	For what ultimate purpose are the data being linked? Could the linked data be used for other purposes, either encouraged or discouraged?	The purpose of the linkage was to answer research questions about the overlap of HUD voucher recipient households and CCDF child care subsidy recipients
Statutory policies	What existing nonstatutory policies will shape the data-sharing agreement?	Undetermined
Existing organizational policies	What existing organizational policies will shape the data-sharing agreement? Are these policies required by statute?	Undetermined
Incentives and risks	What do different parties stand to gain or lose depending on how the linkage is executed?	Incentives: achieve policy goals of HUD and state CCDF agency, receive technical assistance, risk reduction of PPRL, discussion of template DUAs Risks: reidentification of PII
Data life cycle security	How will the data be protected while it is collected, stored, and preprocessed? How will the data be protected when it is transferred? How will the data be used? How will the data be destroyed?	Data were to be stored on secure data enclaves, transferred via secure file transfer protocol, analyzed on NSDS with output privacy protections, destroyed at the end of the project
Transparency	What information about the input data, code, methods, or derived results will be publicly available? Available to some project members?	Undetermined

	Questions	Use case[a]
Enforcement and accountability	Which parties will be able to audit or enforce the agreed-upon terms in each agreement? Who will have visibility into how each part of the process is executed?	Undetermined

Source: Authors' analysis.
Notes: CCDF = Child Care Development Fund; DUA = data use agreement; HUD = US Department of Housing and Urban Development; NCSES = National Center for Science and Engineering Statistics; NSDS = National Secure Data Service; PII = personally identifiable information; PPRL = privacy preserving record linkage.
[a] Some rows in the use case column are marked as undetermined because the project did not complete the use case data-sharing agreements.

Parties: Data linkages will often include multiple parties. At a minimum, data contributors, entities that own and hold the raw data, will be involved. In addition, other entities may be involved as facilitators, evaluators, or third parties. As discussed in the technical design considerations above, multiple partners are involved in a federal-state data-linking effort. The federal and state agencies, as data contributors, play a central role in determining whether a data-sharing agreement is required, what it should include, and which parties must be covered. In most cases, federal and state agencies will require any entity receiving either hashed or unhashed PII to enter into a formal data-sharing agreement.

Federal and state agencies will typically establish agreements with the trusted third party responsible for facilitating the linkage. If an external organization is engaged to preprocess and hash PII on behalf of a federal or state agency, that organization must also be covered by a data-sharing agreement. In contrast, individuals or entities, such as evaluators, that access the linked dataset solely within a secure environment for analysis purposes may be subject to a confidentiality pledge rather than a full data-sharing agreement.

Linkage purpose: For federal-state linkages to be successful, parties must agree on the data linkage purpose. Many federal agencies have compelled states (by various mechanisms) to produce administrative data that could be used for administrative instead of statistical purposes (for example, investigating fraud in the Supplemental Nutrition Assistance Program). States often do not participate in federal-state data linkages because of this concern. Specifying the explicit analytic questions that will be answered with access to linked data protects against data being used outside the scope of the project. The linkage purpose should also determine the terms and conditions under which data and statistical outputs may be used.

Conversely, specificity can also limit the ability of evaluators to maximize the benefits of the linkage. Conducting data linkage requires a significant amount of effort and time at every stage, and federal-state data linkages can often be used to answer a multitude of questions. Linkage purposes should be sufficiently broad or flexible so that the full benefits of the linkage can be realized.

Statutory policies: Federal and state agency administrative datasets are governed by a sectoral patchwork of laws, contracts, and existing agreements that determine the legal permissibility of

entering into new data-sharing agreements for linkage purposes. Federal agencies are often subject to broad-based privacy laws, such as the Privacy Act and the Confidential Information Protection and Statistical Efficiency Act, as well as data domain–specific laws such as the Health Insurance Portability and Accountability Act (commonly known as HIPAA). State agencies may be subject to additional state-specific data privacy laws or preexisting agreements between state and federal agencies (for example, mandatory reporting requirements for state-administered, federally funded programs like CCDF).

Each of these policies places different restrictions on different components of the end-to-end data-sharing process needed for federal-state data sharing. In some cases, federal-state data sharing may be a priori impermissible solely because of these policies (e.g., some agencies are not allowed to share sensitive subpopulation data, such as data on children, incarcerated individuals, or other potentially vulnerable groups). While PPRL could alleviate some of these restrictions, each policy operates with different working definitions for personally identifiable information, and few policies explicitly treat hashed PII differently in the policy's text. Therefore, agency counsel who will ultimately represent and sign data-sharing agreements must be involved as early as possible in establishing these agreements, particularly for interpreting how PPRL may affect the interpretation of data sharing policies.

Existing organizational policies: Federal and state agencies have numerous nonstatutory policies that may operationally impede a legal agreement. For example, state agencies may refuse to share PII with federal agencies unless the research explicitly benefits state residents. Similarly, federal agencies may only issue certain kinds of data use agreements to nonfederal organizations that, a priori, restrict how administrative data can be accessed and used. Such policies are often set by agencies to serve their constituents, even if they are not enforced by law.

Incentives and risks: Federal and state agencies operate under different incentive structures; each is beholden to different constituencies and has different institutional goals. Federal and state agencies do not necessarily share the same upsides or downsides related to participating in PPRL.

The primary incentive for data sharing in the context of evaluation is to answer questions relevant to improving program implementation and testing program effectiveness. For PPRL, state and federal agencies may benefit from learning a new methodology or learning about new software that could be used in future data-linking projects. For federal-state data linkage projects initiated by federal offices, federal agencies may benefit more from learning about state agency data, because they have already identified a data linkage as important to answering federal evaluation questions.

The primary risk of data sharing is the reidentification of PII. Federal agencies may be more insulated from legal and operational privacy risks related to reidentification than are state agencies. For example, state privacy laws may enable greater private rights of action that create higher risks for state agencies, but these higher risks may not apply to federal agencies that often operate with sovereign immunity. Moreover, the relative technical, personnel, and infrastructural costs needed to participate in PPRL and ensure compliance with privacy policies may differ between federal and state

agencies. Federal and state agency counsels are best equipped to understand and plan legal agreements that appropriately mitigate the relative risks.

Data security life cycle: Each stage of a state-federal data-linking project entails risk. Accordingly, data-sharing agreements between federal and state partners must clearly define security requirements for the entire life cycle of the data-linking project to ensure compliance with applicable federal and state privacy laws and policies. These agreements should specify how data will be prepared, transferred, accessed, stored, and ultimately destroyed, ensuring that all parties adhere to consistent and legally sufficient privacy and security standards. In addition, the agreements must outline output control mechanisms that regulate or protect any statistics derived from the linked dataset to prevent reidentification. In general, any disclosure avoidance practices that apply to an individual constituent dataset should be applied, at a minimum, to the linked dataset. For example, HUD data publication guidelines require suppression of any count tabulation cells of size less than or equal to 10.

Transparency: Many federal-state data linkage projects produce lessons or tools that could be useful for the field in general. Sharing legal and technical processes, such as data-sharing agreements, preprocessing code, and hashing code, may help others implement similar projects. Sharing information about the data-generating process, coding, and derived results increases the usefulness of these products for informing future research and evaluations. However, increasing this transparency also increases potential privacy and security vulnerabilities. Parties must agree in advance what should be disclosed for the results to be verifiable and to inform policymaking, what adds unnecessary risk, and how much detail on PPRL mechanics is necessary to instill confidence in the results without adding disclosure risk.

Enforcement and accountability: Data-sharing agreements alone are not always sufficient to prevent data mishandling or misuse. Legal agreements might create incentives to prevent data use that is outside the linkage purpose, but additional technical safeguards must be part of this process to align incentives, which, at a minimum, should include disclosure avoidance techniques, data archiving and deletion policies, purpose limitation mechanisms, and other techniques that ensure agencies do not misuse linked datasets in ways that threaten privacy concerns.

Federal and state agencies may need different assurances that each part of the data-sharing process is unfolding according to plan. By design, PPRL makes it more difficult to audit the use of agency data, because encryption can obscure efforts to understand exactly which data are used in downstream processing and analysis and how the data are used. This is where trusted third parties serve both computational and administrative roles to ensure that PPRL is implemented and used in accordance with legal and operational terms defined in any agreements.

BOX 3.2

Legal Design: Use Case

Parties

As discussed in the above design section, our use case involved multiple parties, including HUD, the state CCDF agency, NCSES, the Urban Institute, MRAIA, and PIPC. The roles and responsibilities outlined above, specifically the extent of data access, determined the legal agreements needed. The HUD PII was to be transferred to Urban to be preprocessed and hashed; therefore, HUD had planned to grant the Urban Institute a data license. The state CCDF agency had planned to send hashed PII and other data elements directly to NCSES; therefore, the state CCDF agency needed to enter into a data-sharing agreement with NCSES. Because individuals outside NCSES would be granted access to the data, the data-sharing agreement included a list of authorized users as well as requirements for these users. MRAIA and PIPC were not permitted access to the data and therefore were not parties to the data-sharing agreements. These agreements were never finalized, so their structure and the parties involved might have changed had the data-sharing agreements moved further along in the review process.

Linkage Purpose

As this project was a use case, the linkage purpose was narrow—to answer a limited set of research questions about the overlap of HUD voucher recipient households and CCDF child care subsidy recipients. In addition, the linkage purpose limited the output from the analysis to aggregate statistics, and the data were to be destroyed after analysis was complete.

Incentives and Risks

While our use case was initiated by a federal agency, both state and federal agencies stood to benefit and both faced some risk. The benefit for both was the opportunity to learn more about the overlap of housing assistance and child care subsidies. Additionally, the state would benefit from learning about PPRL, which they could apply to future projects. The primary risk of participating in this project was the reidentification of PII. The data security life cycle protocol was put in place to help mitigate this risk.

Data Security Life Cycle

The CCDF PII was to be preprocessed and hashed within the state systems. The code and software used were reviewed by CCDF agency staff to ensure security. The HUD PII and housing assistance data were to be sent to Urban via secure file transfer protocol. Urban staff who signed confidentiality agreements were to preprocess and hash the PII on a secure Urban server. Both datasets would then be transferred to NSDS SCE via secure file transfer protocol, where they would then be linked, stored, and analyzed on the NSDS SCE until the end of the project, then destroyed. NCSES would then be responsible for output privacy via the disclosure review process.

Notes: Statutory policies, existing organizational policies, transparency, and enforcement and accountability were still being determined before the end of the project and are therefore not included in our description of the use case.

Once all decisions are made regarding technical and legal design, the project can follow that design in implementation. The technical design may change over time as the project learns more about each agency's data. As the technical design is implemented, it is important to continually refer to the data-sharing agreements to ensure compliance.

4. Preparing for Implementation

Before any data is transferred outside the data-sharing agencies, the full project data-sharing agreements must be drafted, approved, and signed. Even if all legal design decisions have been made, this process can take a long time. To maintain forward momentum, the project should conduct technical steps, such as testing code on simulated data or even running preprocessing and hashing code within the data-sharing partner agencies' systems.

4.1. Drafting and Signing Data-Sharing Agreements

Once legal and technical design decisions are finalized, agencies can start drafting data-sharing agreements. Sometimes agencies prefer to start with template data-sharing agreements from within the agency because they already include legal language addressing the agency's laws or policies. Often, templated language that meets the needs of one agency requires modification for another, so it may be more efficient to start from a more project-specific template.

Developing a data-sharing agreement is generally the most time-consuming part of the data-linking process because it often requires review from multiple levels of staff (legal, IT, leadership) and requires multiple rounds of review as agency staff react to edits. Meetings that include legal counsel from both agencies can help expedite this process.

4.2. Running Simulations to Prepare for Code Implementation

Preprocessing, hashing, and linking code can be developed and tested before accessing the actual data. The code should be developed using an agreed-upon software that can be run by the state or federal agency systems that will be processing data. The code can be developed using codebooks and metadata to determine how the data are structured and formatted. Once developed, the code can be tested on simulated data (i.e., an artificial dataset with the same properties, format, and distributions as real data).

In addition to developing the code, the SCE can also be prepared by ensuring that software necessary for linking and analyzing data is approved and available within the environment. Testing with simulated data could also be completed on the SCE to ensure that the code runs properly and efficiently in the environment.

BOX 4

Preparing for Implementation: Use Case

4.1. Drafting and Signing Data-Sharing Agreements

We developed a template data-sharing agreement in consultation with the PIPC that could be used as a starting point for an agreement between state agencies and NCSES. The state CCDF agency that was participating opted to use the existing state template, which was modified for the use case but did not complete review by NCSES and was not finalized before the project ended.

4.2. Running Simulations to Prepare for Code Implementation

While the data-sharing agreement was being developed, we used publicly available information on CCDF data, codebooks, and metadata from the CCDF agency, as well as toy datasets, to develop and test code for preprocessing, hashing, and linking.

5. Preparing Data for Linkage

As discussed in detail in the technical design decisions section, preprocessing PII involves cleaning and standardizing. For PPRL, it also involves creating new composite variables. If there are concerns from agencies about sharing granular PII (e.g., exact address) even after hashing, this step may also include the generation of coarser PII (e.g., zip code) before hashing. Hashing PII involves creating a salt key to attach to raw PII and then passing all PII through a secure hash algorithm to prevent reidentification. This code can be run within agencies before data-sharing agreements are in place to build familiarity with and test the efficacy of these steps. Code should incorporate frequent checks to ensure that assumptions made when creating the code hold in real life.

BOX 5

Preparing Data for Linkage: Use Case

We were unable to run the code we developed before the project ending. However, we did run simulations, which provided some insight.

6. Linking Data

Once data-sharing agreements have been finalized and signed, full implementation can begin. The preprocessed and hashed PII and the other data elements can be transferred to the trusted third party's computing environment. It is important that the data-sharing agreement be reviewed to ensure all parties are in compliance and following data security protocols for transferring and linking the data.

After the data are linked by either deterministic or probabilistic linkage methods, the linkage should be assessed for match quality. This can be difficult to assess for PPRL because the hashed PII are not human readable. Information from prior research could help assess whether a match rate is reasonable (e.g., if a prior survey or data match has shown that half of those participating in the federal agency program are also participating in the state agency program, this provides a reasonable expectation for match rate). Once data are linked, hashed PII should be destroyed.

BOX 6

Linking Data: Use Case

We were unable to link the data from our use case before the end of our project. However, we were able to run simulated data linkages using artificial data. We tested the reliability of PPRL methods on a publicly available dataset, FEBRL 4, which is widely used to evaluate record-linkage methods. The data contain 5,000 true matches with realistic noise to test how well record-linkage algorithms can identify the true matches. We compared seven PPRL models using hashed data with a baseline probabilistic model using unhashed records that is commonly used in standard data linkages. We found that hashing PII sacrificed little in match quality. Our strongest PPRL models identified 99.44 percent of the 5,000 true matches, which was nearly comparable with our best model that used the unhashed PII (99.96 percent). Our results were also robust to excluding any single identifiers—even after dropping names, Social Security identifiers, or postal codes, our process identified over 97 percent of true matches. Only one model performed poorly overall, with a recall under 70 percent; this model likely filtered out too many valid pairs during the blocking process. These results suggest that PPRL has strong potential in producing match rates comparable with standard methods. However, because the sample dataset used in the simulation contained only matches (and not mismatches), these percentages overstate the quality of match results we would expect in administrative data, in which the data would likely contain far more false positives (i.e., records that may seem similar but do not correspond to the same person) and would be messier than the FEBRL 4 data. However, hashing data only modestly reduced match quality compared with our benchmark, which suggests that these methods—even with the information loss required to protect individual privacy—hold promise and can detect real signal in fields used for matching, even after they are encrypted.

7. Analysis

Analyzing linked federal and state administrative data presents several challenges. Because administrative data are collected for operational rather than research purposes, substantial data cleaning is often required. Combining datasets from different sources also requires careful alignment of variables, definitions, and data structures before analysis can begin. In addition, as with all analysis including administrative data, accurate interpretation of findings depends on a thorough understanding of the quality, meaning, and limitations of the data.

For PPRL, data should be analyzed within the same system in which they were linked. Data leaving the system would need to be sufficiently masked to protect privacy. Even without hashed or unhashed PII, agencies contributing data could reverse engineer PII by matching it back with their original data on a sufficient number and granularity of variables. Therefore, any data leaving the SCE would need to be coarsened to ensure privacy is protected. This is not ideal for analysis, which may require more detailed data. Data should be analyzed in a way that aligns with the linkage purpose outlined in the data-sharing agreement.

Analyzing data within a third party's SCE presents several limitations. First, adding or removing user access may take time. Only individuals approved under the data-sharing agreement can access the system, and any additional staff may need to sign supplemental agreements and obtain approval

before access is granted. Second, required software may not be available in the environment, and installing new software may require review and approval by the third party. Third, as discussed below, all statistical outputs must be reviewed to ensure compliance with disclosure rules before they can be shared outside the environment. This review process can delay communication of results needed for data cleaning and analytical decisionmaking.

BOX 7

Analysis: Use Case

While we were able to develop an analysis plan, we were unable to execute it. For this use case, we worked toward linking HUD and CCDF data. Our primary research questions for our analysis of the linked data were about the extent to which households receiving HUD assistance overlapped with those receiving CCDF child care subsidies. In addition, we wanted to understand the characteristics of households that did or did not overlap. Among households that overlapped, we wanted to understand the type and quality of child care that families used and how long families used their subsidy to pay for child care. We planned to conduct descriptive analysis such as cross-tabulations, *t*-tests to test for statistically significant differences, and, possibly, descriptive regression analysis.

8. Output Review

Before any data or statistical output leave the SCE, they must be reviewed to ensure they do not pose a risk of reidentification and that they align with the linkage purpose specified in the data-sharing agreement. This applies to both data and statistical output, such as descriptive statistics.

BOX 8

Output Review: Use Case

Our use case did not reach the point of developing a detailed output review process; however, we decided that no data would leave the SCE and only statistical output would be allowed. While NSDS itself does not have an explicit disclosure review process, it does have technical staff who implement disclosure reviews with processes specified by data-donating agencies like NCSES. We had planned to develop a process following HUD's existing disclosure review guidance for statistical releases.

9. Long-Term Storage

Depending on the linkage purpose outlined in the data-sharing agreement and the data security life cycle, the linked data could either remain stored on the SCE for future use or be destroyed. Preserving the linked data has many advantages, especially considering that linked data can often be used to

answer additional research questions across multiple projects and that developing the data-sharing agreements and linking the data is time intensive. Partners would need to agree and codify approval processes in the data-sharing agreements for giving additional researchers access and using the data to answer additional questions. However, preserving the data poses additional privacy and legal risks. Destroying the linked data effectively eliminates the risk of reidentification. Although small, the risk of reidentification or misuse is inherent with long-term storage. In addition, as agency priorities shift, agencies may change their minds about long-term storage of their data or the linkage purpose. Data-sharing agreements should allow agencies to opt out of long-term storage at any time. In addition, by restricting access to linked datasets and enforcing output reviews, NSDS can help protect against some of these risks.

BOX 9

Long-Term Storage: Use Case

We were unable to reach the point of data destruction or long-term storage, but we had already decided that the data should be destroyed at the end of the project for privacy assurance.

Lessons Learned and Recommendations

This report outlined the workflow for federal-state linkage projects initiated by federal evaluation offices. First, federal evaluation offices must determine what research questions could be answered by federal-state linkages and make initial legal and technical design decisions based on federal agency limitations. Next, states must be recruited to participate in the data-linking project. Once a state has agreed, it is important for participating agencies and any partners to reach agreement on the technical and legal designs. Next, agencies must draft, review, finalize, and sign data-sharing agreements. This process takes substantial amounts of time and effort. To keep forward momentum, projects should prepare and test code for data linking. For PPRL, agencies can preprocess and hash PII while data-sharing agreements are being finalized. Preprocessed and hashed data can then be transferred to the trusted third party for data linking, analysis, and storage after data agreements are signed. Once the data are analyzed, any output should be reviewed to ensure that it does not pose a risk of reidentification and follows disclosure rules. Finally, data can then either be kept with the third-party system for long-term storage or destroyed as agreed upon by the data owners.

We tested this workflow through a use case in which we attempted to link federal HUD data and state CCDF data. We learned several key lessons from this use case. States are often reluctant to participate because of capacity constraints, misalignment with strategic priorities, or concerns about privacy. Successfully recruiting state partners requires providing incentives (e.g., answering state-specific research questions) and minimizing burden (e.g., providing technical assistance). Federal-state data linkages require a long timeline and significant staff time investment from all partners involved.

Much of this time is spent developing, reviewing, and finalizing data-sharing agreements. The timeline can be reduced by mitigating privacy concerns associated with exchanging PII by implementing strong security protocols, communicating security protocols in plain language, and bringing staff together to resolve roadblocks. PPRL can assuage some, but not all, of these privacy concerns about exchanging PII. However, it also introduces additional burdens (e.g., the need to explain the novel, technical processes to new users, preprocess and hash PII within agency systems, and account for matching limitations specific to PPRL). While the roadmap lays out the key steps for federal-state data linkages and offers suggestions for expediting these processes, prospective parties need to be prepared for long timelines and a significant investment from staff.

NSDS has the potential to help make these processes more efficient and less burdensome. In this use case, NCSES served as the trusted third party that hosted the NSDS SCE, where data could be linked, analyzed, and stored. As the SCE host, NSDS was responsible for the security of this environment (e.g., restricting user access) and for setting and enforcing output controls. While we were unable to ingest federal or state agency data into the NSDS SCE, our project suggests that NSDS could serve in this third-party role for federal-state linkages. This may be especially useful when data linkages are combined with privacy-enhancing technologies, such as PPRL. NSDS could also improve federal-state data linkage efficiency in a number of ways:

- Developing trust with state agencies to mitigate concerns about sharing data by providing details on security protocols and enforcement and conducting informative outreach to state agencies.
- Continuing to test and refine additional security technologies to further mitigate privacy concerns.
- Providing templates and technical assistance for common elements of federal-state data-linking projects, such as template language for data-sharing agreements, plain-language explanations of the security provided by the NSDS SCE and privacy-enhancing technologies, template code for preprocessing and linking data, and template code for hashing PII.
- Providing multiple software options for data linking and analysis on the NSDS SCE may also be helpful, as the ideal software for each step and for each project may vary.

By supporting a few additional technical and legal processes to better operationalize federal-state linkages and improve their efficiency, NSDS can move closer to its goal of being the front door of access to different tiers of confidential and sensitive data across the federal and state statistical systems.

Notes

[1] While this could technically be accomplished by linking state data with local Public Housing Authority data, the project wanted to cover a broader geographic area (state or multistate). Linking state CCDF and Public Housing

Authority data at the state level would require a data-sharing agreement with each of the state's Public Housing Authorities. Given the high number of Public Housing Authorities per state (some states have 100 or more), this would be prohibitively burdensome. While case-level CCDF child care data are available at the federal level (e.g., ACF 801 Monthly Child Care Data Report), it is de-identified and therefore cannot be linked with other data sets.

[2] Deterministic record linkage looks for exact matches on one or more variables between two records. This is ideal when the variables are unique, consistent, and reliable across the datasets or when the risk imposed by a mismatch is high (e.g., when determining eligibility for a program). Probabilistic record linkage uses a statistical approach that measures the probability that two records represent a true match. It uses weights to calculate the match scores and thresholds to determine a match, nonmatch, or possible match.

About the Authors

Jeremy Seeman is a senior research associate in the Tax and Income Supports Division at the Urban Institute. His research focuses on technical and policy solutions for responsible data sharing, with an emphasis on deploying privacy-enhancing technologies to expand access to sensitive data. Before joining Urban, Seeman was a Michigan Data Science Fellow at the University of Michigan and a faculty affiliate of the university's Center for Ethics, Society, and Computing. Seeman received his BS in physics and MS in statistics from the University of Chicago and his PhD in statistics from Penn State, where he was a US Census Bureau Dissertation Fellow.

Devlin Hanson is a principal research associate in the Family and Financial Well-Being Division at the Urban Institute. She is a labor economist, specializing in rigorous impact evaluations, whose research focuses on housing, and child welfare. She received her PhD from Boston College.

Judah Axelrod is the associate director of data science research methods at the Urban Institute. He works on Urban's Technology and Data Science team and in collaboration with the Center for Equity and Community Impact to drive policy research using data science tools and methods. Axelrod leads Urban's work on the equitable use of AI and automated systems and has expertise leading complex, interdisciplinary research projects that leverage tools such as web scraping, data linkage, imputation, statistical modeling, geospatial analysis, and natural language processing. Axelrod holds a BA in economics from Rutgers University and an MSc in data science from the London School of Economics.

Rebecca Berger is a senior research associate in the Family and Financial Well-Being Division at the Urban Institute. In her substantive research, she focuses on promoting access to services, particularly early care and education, for young children and their caregivers. Berger holds an MS and a PhD in family and human development from Arizona State University.

Jordan Hardenburgh is a research assistant in the Family and Financial Well-Being Division at the Urban Institute. She applies qualitative and quantitative methods to research on child welfare and early childhood education. Hardenburgh holds a bachelor's degree in criminal justice and international affairs from Northeastern University.

Evy Park is a software engineer on the Technology and Data Science team at the Urban Institute. She builds applications and tools for policy research. Park holds a BA in Economics from Georgetown University and a MS in computational analysis and public policy from the University of Chicago.

Ridhi Purohit is a data scientist on the Technology and Data Science team at the Urban Institute. at the Urban Institute. She supports policy research through tools that integrate data, machine learning, and responsible AI practices. Her recent work includes benchmarking GenAI models and developing public-facing web reports for policymakers and practitioners. Purohit holds an MS in computational analysis and public policy from the University of Chicago.

Acknowledgments

This brief was developed by the Urban Institute, under Government contract with America's DataHub Consortium (ADC), a public-private partnership that supports the strategic objectives of the National Center for Science and Engineering Statistics within the U.S. National Science Foundation. Strategic leadership for the development of the NSDS is provided by the Interagency Council on Statistical Policy's (ICSP) Subcommittee on the NSDS (S-NSDS). We are grateful to them and to all our funders, who make it possible for Urban to advance its mission. We thank MRAIA for their partnership and expertise in developing and refining our data preprocessing, encryption, and linkage methodology. We thank PIPC for their partnership and expertise in developing data-sharing agreements. The views expressed are those of the authors and should not be attributed to the Urban Institute, its trustees, or its funders. Funders do not determine research findings or the insights and recommendations of Urban experts. Further information on the Urban Institute's funding principles is available at urban.org/fundingprinciples.

ABOUT THE URBAN INSTITUTE

The Urban Institute is a nonprofit research organization founded on one simple idea: To improve lives and strengthen communities, we need practices and policies that work. For more than 50 years, that has been our charge. By equipping changemakers with evidence and solutions, together we can create a future where every person and community has the opportunity and power to thrive.