Utilizing Privacy Preserving Record Linkage (PPRL) to Link Data from Two Federal Statistical Agencies

Final Project Summary Report

August 15, 2025

Jason Meyer; Chris Frederick; Rick Edwards; Rui Wang, PhD; Honoka Suzuki MA; Alex Bohl, PhD; and David Jones, PhD

Prepared for National Center for Science and Engineering Statistics (NCSES) within the U.S. National Science Foundation (NSF) by HealthVerity and Mathematica

ADC Base Agreement #: 2023-604

ADC Project Agreement #: ADC-PPRL1-23-N03

This page has been left blank for double-sided copying.

Table of Contents

1. Executive Summary

1.1. National Secure Data Service Demonstration

The National Secure Data Service Demonstration (NSDS-D) is a federal initiative authorized under Section 10375 of the CHIPS and Science Act of 2022. The initiative is operated by the National Center for Science and Engineering Statistics (NCSES), which is a federal statistical agency in the U.S. National Science Foundation (NSF), legislatively mandated in the National Science Foundation Act of 1950 (42 U.S.C. 1862 (a) (6)) to serve as a central federal clearinghouse for the collection, interpretation, analysis, and dissemination of objective data on science, engineering, technology, and research and development and to provide a source of information for policy formulation by other agencies of the federal government.

The goal of the NSDS-D is to inform the future establishment of a government-wide hub of services and resources for data linkage, data sharing, secure access infrastructure, and user training. This shared services model is intended to enhance evidence-based policymaking across the federal government, while preserving data privacy and confidentiality. Many of the activities under the NSDS-D are implemented through the America's Data Hub Consortium (ADC), a private-public partnership sponsored by NCSES to facilitate research collaboration across the government, industry, and academia in topics including but not limited to data linkage, access, security, and privacy.

1.2. Project Overview

To support the NSDS-D efforts, this project serves as a proof of concept to develop a data sharing agreement between two federal statistical agencies that have not previously developed data sharing relationships, deploy a privacy preserving record linkage (PPRL) tool to link data from two federal statistical agencies and utilize a secure environment to analyze the resulting linked data file. PPRL is a method that can be used to link de-identified data, using encrypted tokens[1] and a trusted third-party cloud provider. In this project, PPRL was performed in a HealthVerity secure environment that meets CIPSEA (Confidential Information Protection and Statistical Efficiency Act) and FedRAMP (Federal Risk and Authorization Management Program) standards. Once the data were linked using a PPRL tool and stripped of the encrypted tokens, they were then made available for analysis in a secure analytical environment.

NCSES supported this opportunity, using the America's Datahub Consortium (ADC) managed by Advanced Technology International (ATI), to award a prime contract to HealthVerity. HealthVerity is supported by its subcontractor Mathematica. The Centers for Disease Control and Prevention's (CDC) National Center for Health Statistics (NCHS) is a technical partner for this project.

[1] A token is a privacy-safe, encoded form of PII used to match records without revealing the underlying data. HealthVerity extends this process by matching tokens to their master patient index to assign a persistent HealthVerity Identity (HVID), enabling consistent, privacy-protected linkage across datasets. Throughout this document, the term "token" is used in the generic PPRL context, while HVID refers specifically to HealthVerity's proprietary process and resulting identifier.

2. Introduction

2.1. Request for Solution (RFS) Background

In 2023, NCSES issued a Request for Solutions (RFS) via the ADC titled "Utilizing Privacy Preserving Record Linkage (PPRL) to Link Data from Two Federal Statistical Agencies." The project aimed to:

- Develop a data-sharing agreement between two agencies with no prior collaboration
- Deploy a PPRL tool to link data from both agencies
- Analyze the linked data within a secure environment

The RFS noted that, if successful, the project will inform the development of a National Secure Data Service (NSDS), and will inform linkages across the federal government, using the development of agreements and deployment of PPRL as a model to improve the availability, quality, accessibility, and interoperability of data sharing within the NSDS.

America's Datahub Consortium

America's DataHub Consortium (ADC) is a public-private partnership established in 2021 by the National Center for Science and Engineering Statistics (NCSES), part of the National Science Foundation (NSF). Its mission is to serve as a national asset that brings together qualified individuals and secure data to enable collaborative research and evidence-based decision-making that benefits the American public. ADC focuses on enhancing data access, security, and analysis by supporting projects that involve data collection, linkage, privacy protection, and innovative statistical methods. It also plays a key role in informing the development of the National Secure Data Service (NSDS), a federally authorized initiative aimed at modernizing the U.S. data infrastructure

Managed by Advanced Technology International (ATI), ADC facilitates agile collaboration among federal agencies, academic institutions, nonprofits, and businesses, including non-traditional partners new to working with NSF. The consortium has supported over 30 projects since its inception, addressing critical issues such as STEM workforce development, privacy-preserving technologies, and the integration of diverse data sources. Through its flexible acquisition processes and emphasis on innovation, ADC aims to strengthen the federal data ecosystem and promote public trust in data-driven policymaking.

In order to conduct the linkage a competitive procurement was conducted by NCSES (using ATI to facilitate) for a compliant PPRL solution. The goals were to acquire a solution that is readily available for commercial use, and to procure a tool that used advanced techniques, including Artificial Intelligence (AI) & Machine Learning (ML)-enabled probabilistic matching, to optimize matching accuracy, optimize matching speed, reduce re-work, and enable cross-functional scaling across government. In addition, a PPRL solution that can integrate with a secure environment managed by NCSES and comply with security and privacy standards (including CIPSEA and FedRAMP) was required.

Several bidders were evaluated. HealthVerity's PPRL solution, a commercial tool used widely across life sciences, government, and healthcare for data linkage in research, surveillance, and analytics was selected. HealthVerity's solution applies advanced techniques, including AI/ML, to deliver secure, scalable linkages. HealthVerity included Mathematica as a partner in order to complete the analysis of the data within a secure environment.

2.2. Stakeholders and Roles/Responsibilities

The project required the linkage of data from two federal statistical agencies – the National Center for Science and Engineering Statistics' (NCSES) Survey of Earned Doctorates (SED) and the National Center for Health Statistics' (NCHS) National Health Interview Survey (NHIS). Therefore, the technical components of the project were co-led by individuals from NCSES, and the NCHS. The following table lists each core project team member (by organization):

Table 1: Core Project Team Members

Organization	Project Role(s)
National Center for Science and Engineering Statistics	• NCSES Project Lead; Agreements Officer

(NCSES)	Representative (AOR) ● Programmer ● Survey Manager for SED ● Mathematical Statistician
National Center for Health Statistics (NCHS)	● NCHS Project Lead ● Statistician ● Programmer
Advanced Technology International (ATI)	● Program Manager ● Program Administrator
HealthVerity	● Project Leadership ● Compliance and Contracts ● Solutions Engineer ● Project Manager
Mathematica	● Project Leadership ● Senior Data Scientist ● Data Scientist

The successful execution of this demonstration project required a multi-disciplinary team approach that extended beyond the core project team members. Depending on the Task Area and the specific tasks being worked on, an extensive list of individual stakeholders and teams were also actively engaged, including:

- Office of General Counsel
- Chief Information Security Officer (CISO)
- Chief Statistician
- Associate Director of Science
- Confidentiality Officer
- Information Technology Team
- Secure Data Access Facility Managers
- HealthVerity Data Integrations Team
- HealthVerity Legal Team
- HealthVerity Security Team
- HealthVerity Software Engineering Team

* NORC manages the NCSES' Secure Data Access Facility (SDAF)

2.3. Project Organization by Task Area

This project was organized into four distinct Task Areas, as shown below:

- Task Area 1: Data Sharing Agreement (DSA)
- Task Area 2: Privacy Preserving Record Linkage (PPRL)
- Task Area 3: Validation Statistics and Modeling
- Task Area 4: Project Management

As mentioned in Section 2.2, the completion of this demonstration project required a broad spectrum of multi-disciplinary expertise, engagement, and oversight - above and beyond the core project team. In the table below, each of the four Task Areas are listed, along with; (a) a summary of the major tasks included in the scope of each Task Area, (b) a list of the core project team members involved in each Task Area, and (c) a list of the other additional multi-disciplinary stakeholders required for each Task Area:

Table 2: Summary of Major Project Tasks

Task Area	Summary of Major Tasks	Core Project Team Members	Multi-Disciplinary Stakeholders
Task Area 1: Data Sharing Agreement (DSA)	● Development and Full Execution of the Software License Agreement (SLA) ● Development and Full Execution of the Data Sharing Agreement (DSA)	● NCSES ● NCHS ● HealthVerity	● Office of General Counsel ● Chief Information Security Officer (CISO) ● Chief Statistician ● Associate Director of Science ● Confidentiality Officer ● NSF OCIO Change Control Board (CCB) ● NSF Engineering Review Board (ERB) ● HealthVerity Legal Team
Task Area 2: Privacy Preserving Record Linkage (PPRL)	● HealthVerity Provisioning of FedRAMP PPRL Matching Environment Specific for NSF ● PPRL Data Layout Submission Form - Completion and Submission ● Set-Up, Configure, and Test Secure Data Transfer Connections ● PPRL DeID Engine Configuration, Deployment, and Installation ● PARTIAL Dataset Submission (10k records) through the PPRL DeID Engine ● FULL Dataset Submission through the PPRL DeID Engine ● Data Delivery of the HVIDs and Row IDs to NCSES Lead ● Ingestion of HVIDs, Row IDs, and Covariate Data into the NCSES SDAF ● Completion of Data Destruction Process	● NCSES ● NCHS ● HealthVerity	● Information Technology Team

Task Area	Summary of Major Tasks	Core Project Team Members	Multi-Disciplinary Stakeholders
Task Area 3: Validation Statistics, and Modeling (VSM)	● NCSES Grants Mathematica Access to the Combined NCSES and NCHS Datasets, within the NCSES SDAF Environment ● Mathematica Links and Cleans the Combined NCSES and NCHS Dataset ● Mathematica Completes Descriptive Analysis ● Mathematica Prepares and Delivers the Task Area 3 Final Report (Draft #1, Draft #2, Final Version)	● NCSES ● NCHS ● HealthVerity ● Mathematica	● NCSES Secure Data Access Facility managers
Task Area 4: Project Management (PM)	● HealthVerity Develops and Manages Project Plan, Project Timeline, and other Project Management Artifacts ● CIPSEA Compliance ● Bi-Weekly Project Status Meetings ● Monthly Project Status Reports	● NCSES ● NCHS ● HealthVerity ● Mathematica	● ATI, the managing firm for the ADC

3. Task Area 1 - Data Sharing Agreement (DSA)

3.1. Overview

A foundational component of this project was the creation and execution of a Data Sharing Agreement (DSA) between NCSES and NCHS. This agreement was essential to enabling the secure and privacy-preserving linkage of the NCSES SED and the NCHS NHIS[2]. As these two federal statistical agencies had not previously collaborated through direct data sharing, establishing a formal DSA required careful navigation of legal, policy, and technical frameworks, including compliance with CIPSEA and relevant standards under the FedRAMP.

The DSA process was designed not only to facilitate linkage for this specific proof-of-concept project but also to inform a broader framework for future interagency data collaborations. The development involved identifying and aligning stakeholder requirements, defining roles and responsibilities for data stewardship, and ensuring that all privacy, confidentiality, and security obligations were addressed through a shared understanding. Additionally, the agreement outlined how the selected PPRL tool would be used within a secure environment hosted by a HealthVerity trusted third party cloud provider, creating a replicable model for future linkages under the envisioned NSDS.

This section provides a summary of the steps and activities undertaken to draft, negotiate, and finalize the DSA, highlighting the institutional considerations, challenges encountered, and lessons learned. These insights contribute directly to the project's objective of advancing technical infrastructure and governance models that support rapid, secure, and ethical data sharing across federal agencies.

3.2. Summary of Steps Taken

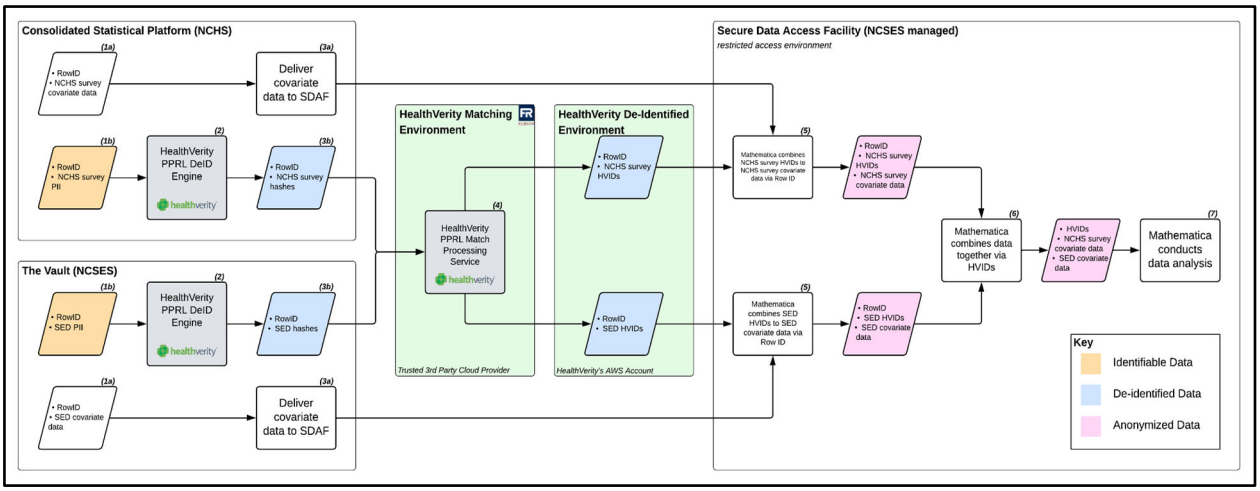
Task Area 1 encompassed the development and full execution of two critical agreements: (1) a Data Sharing Agreement (DSA) between NCSES and NCHS), and (2) a Software License Agreement (SLA) between HealthVerity and each of the two participating agencies. The development of the DSA between NCSES and NCHS began with the circulation of a draft template in November 2023.

Early in the project, during Task Area 1, HealthVerity's Solutions Engineer created detailed documentation of the PPRL solution. This included the "Utilizing PPRL to Link Data from Two Federal Statistical Agencies (PPRL1-23) Detailed Solution Overview" and an accompanying PPRL Data Flow Diagram for NCHS and NCSES Linkage (see Figure 3 below). These materials were instrumental in communicating the PPRL process, technical concepts, and data flow structure to the diverse, multidisciplinary stakeholders involved in finalizing both the DSA and SLA.

In order to link the NCHS NHIS data with the NCSES SED data, the PPRL process required the secure processing and transfer of each agency's data through a series of controlled steps and environments. Figure 1 illustrates the end-to-end data flow for the PPRL process.

[2] The 2023 RFS noted that the NCHS National Hospital Care Survey (NHCS) was to be one of the linked datasets, but prior to the initiation of the DSA process, the project co-leads determined that NHIS would be the data shared by NCHS for this project.

Figure 1. PPRL Data Flow Diagram for NCHS and NCSES Linkage



After a few months, it was determined that a designated agent agreement (DAA) would need to be attached to and accompany the DSA, incorporating specific language around data security and PPRL technology that required HealthVerity's input. HealthVerity provided tailored PPRL documentation and data flow diagrams, reviewed draft DAA language, and offered feedback to refine the agreement. As the project progressed, HealthVerity revised its technical documentation to clarify roles and responsibilities for each organization, encryption protocols, secure file transfer protocol (SFTP) implementation, and the process for data storage, retention, and destruction.

Key project related adjustments impacting the DSA included:

- During the DSA development process, NCSES and NCHS made a joint decision to handle covariate data from both surveys (SED and NHIS) outside of the HealthVerity de-identification process, in order to further protect data from the risk of disclosure. This decision prompted further updates to the supporting PPRL solution documentation, which included the "Utilizing PPRL to link data from Two Federal statistical agencies (PPRL1-23) Detailed Solution Overview" document and the accompanying PPRL Data Flow Diagram for NCHS and NCSES Linkage (see Figure 1). Covariate data refers to analytic, survey variables which are not personally identifiable information (PII) used in the PPRL de-identification process.

- This project required a PPRL technology with a FedRAMP Moderate security certification, due to the need to link two distinct federal government agency datasets. As a result, the NSF Chief Information Security Officer (CISO) required an agency-specific FedRAMP Authorization to Operate (ATO) to utilize the HealthVerity software on this project. This also required each agency's security teams and Information Security Officers to review the current HealthVerity Department of Health and Human Services (HHS) FedRAMP ATO and approve the ATO for use under this agreement.

- HealthVerity provided clarifications on salt[3] encryption and SFTP key management, submitted essential compliance forms, and confirmed SFTP's inclusion under its FedRAMP PPRL certification. About nine months into the project, the HealthVerity PPRL solution gained approval

[3] Salt or salting refers to the process of adding a random string (called a salt) to sensitive data—like names or identifiers—before hashing it, making it more difficult for an attacker to reverse-engineer or match records using dictionary or frequency attacks. This technique enhances security by ensuring that identical input values (e.g., "John Smith") result in different hashed outputs across datasets unless the same salt is used and shared under controlled conditions.

from the NSF Change Control Board and was presented to the NSF Engineering Review Board. With final agency-level reviews completed, the agreement was fully executed between NCSES and NCHS.

The DSA development and negotiation process took approximately ten months to complete, requiring a multi-disciplinary approach, inter-agency collaboration, and iterative support and documentation from HealthVerity.

Development of the Software License Agreement (SLA)

In addition to the DSA, each participating agency was required to negotiate and fully execute a SLA with HealthVerity, given that each agency would be deploying HealthVerity's PPRL de-identification (DeID) engine software within their respective local environments. The SLA included a summary of the work to be completed and established the terms and conditions for the deployment and use of the commercial HealthVerity PPRL software.

The SLA development and negotiation process took approximately four months to complete.

3.3. Summary of Results/Output

The DSA was intentionally drafted using a modular structure, with distinct sections addressing specific components of the project. This design was intended to make the final DSA adaptable as a template for future shared data service initiatives. Key sections of the finalized and fully executed DSA included the following:

Purpose – This section outlined the DSA's primary objectives:

- Establish a data sharing agreement between two federal statistical agencies.
- Deploy and utilize PPRL methods to link NCSES SED and NCHS NHIS data.
- Conduct analyses using the linked data to address research questions that could not be answered by either dataset alone.
- Document processes and lessons learned.

Roles and Responsibilities – Summarized the roles of NCHS, NCSES, HealthVerity, and Mathematica in the execution of the project.

Data Security and Safeguards – Covered topics such as:

- Access, storage, and disposition of data; including tokenization procedures, data destruction, encryption protocols, and NCSES SDAF access controls.
- Data transport methods, specifically the configuration and use of SFTP for secure file transfers.
- Disclosure and confidentiality standards, including requirements under CIPSEA for handling tokenized and covariate data.

The DSA development process exemplified the close collaboration and rigorous review needed across both agencies to ensure full compliance with legal and security requirements. HealthVerity played a pivotal role in aligning technical documentation with agency expectations and enabling secure, privacy-compliant data sharing throughout the project.

3.4. Lessons Learned

Given the "proof of concept" nature of this demonstration project, the importance of the documentation of Lessons Learned was stressed before and throughout the duration of the project. As the project team worked through the four Task Areas of the project, Lessons Learned were discussed and documented by

the core project team. On a quarterly basis, the Lessons Leamed were reviewed and approved by the AOR.

A complete list of all of the approved Lessons Learned from this project can also be accessed and reviewed at: https://www.americasdatahub.org/adc-lessons-learned-pprl1-23-n03/.

Below are some of the Lessons Learned that were captured during Task Area 1:

1. While a government agency cannot define the exact documents/agreements that will be required in future linkage projects, if they ask the following questions, it should help define the needs for each specific project in a timely and efficient fashion. Once the questions are addressed and the information is gathered, the government agency or designated entity can determine what needs to be created and/or revised, who needs to be involved, and in what sequence the activities must be completed in order to successfully get all data sharing parameters in place:
    - Who are the stakeholders involved with the project (software owners, data owners, analytics providers, data platform, etc.)?
    - From the list of defined stakeholders, does your organization have any established data use agreements or other materials that need to be used?
    - What entity(ies) should be involved from your agency in order to craft, review, approve, and sign any agreements established (Office of General Counsel, Contracts, Procurement, etc.)?
    - Are there any specific data elements, data uses, or other use cases that are restricted or require higher levels of approval?
2. Ensure that a similar level of involvement and engagement by government agency leadership can be provided throughout the course of projects because decisions cannot be made unilaterally by just one of the partners; decisions must be bi-lateral, and they need to be handled in near real time to keep the project moving forward efficiently.
3. For data linking engagements, it's critical to understand where the data resides, therefore the government agency should consider asking these questions at the outset of the project:
    - Where does the data reside?
    - Is the environment cloud-based or on-premises?
    - Who maintains the environment (contractor or government)?
    - Can the government agency provide a system diagram that shows where the data resides and how it will be extracted for linkage?
    - Who is responsible for approving the transfer of data outside of the environment?
4. For data linking engagements, it's critical to understand the details and attributes of the data that will be linked, because the data sharing agreement must ensure that all data elements required to link the data are adequately covered. Questions that should be considered:
    - What PII fields does the data contain and how is the data formatted (e.g., name, address, social security number, zip code, etc.)?

- Are there unique features of the PII that the government agency needs to be aware of (e.g., multiple rows for one person and multiple derivatives of their name [Joseph, Joe, Joey]; columns with low fill rates; restrictions on certain fields)?
- Are hashed/encrypted PII fields still considered PII by the government agency?
- Is there any sensitivity around the covariate/transactional data (i.e., does it need to stay on government agency servers? Could there be quasi-identifiers (or indirect identifiers) in the data?)
- What is the overall data layout that will be processed?
- What are the expected fill rates for each field that will be processed?
- What is the source of the data? Where is it being pulled from?

5. In order to utilize the PPRL tool for this project, NCSES had to obtain an agency-specific ATO requirement. This project required a PPRL technology with a FedRAMP security certification. The lesson learned from this experience is that additional time may be needed in the project timeline to meet the IT security requirements to utilize a PPRL technology with a FedRAMP security certificate.

6. Successfully developing, negotiating, and executing a data sharing agreement between two agencies requires multiple steps, and activities. However, this demonstration project found that the two most critical elements of this process are: (1) ensuring commitment from strong project leads from each agency; leads who can serve as committed advocates and champions of the initiative, and (2) forming a multidisciplinary team approach that actively engages key stakeholder groups throughout the process, including agency leadership, legal and regulatory staff, data security experts, IT technical support, and statistical and analytical teams.

   This Lessons Learned further supports the themes, keys to success, and other findings in the America's DataHub Consortium (ADC) Privacy Preserving Technology Phase 1 – Environmental Scan Report [4]

[4] See America's DataHub Consortium (ADC) Privacy Preserving Technology Phase 1 – Environmental Scan Report (January 2024). Available at https://www.americasdatahub.org/wp-content/uploads/2024/05/ADC-PPT_FinalReport.pdf

4. Task Area 2 - Privacy Preserving Record Linkage (PPRL)

4.1. Overview

Following the successful execution of the DSA between NCSES and NCHS, the program was positioned to begin the process of linking data between NHIS and SED. As stipulated in the DSA, NCHS was responsible for tokenizing all Personally Identifiable Information (PII) from using an approved encryption algorithm before any data could be shared with NCSES or its contractors. In parallel, NCSES was required to apply the same tokenization protocol to the PII within the SED dataset. By leveraging a standardized tokenization and matching algorithm with a consistent PII layout, both agencies formatted the applicable demographics data to generate compatible tokens which enabled privacy-preserving record linkage without revealing individual identities.

To ensure compliance with the agreement and the security standards mandated by a FedRAMP Moderate environment, HealthVerity established a dedicated and isolated infrastructure to support this linkage task. The following section outlines the technical and operational steps undertaken to configure this secure environment and execute the de-identification, tokenization, and linkage of records in accordance with the agreed-upon data protection framework.

4.2. Summary of Steps Taken

Provisioning of the PPRL environment

In parallel with the development of the data sharing agreement, the HealthVerity engineering team began building the FedRAMP PPRL Matching Environment. The creation of the development environment took two months to complete, followed by a validation and testing phase that ran for an additional month. Upon successful completion of this phase, the team proceeded to create the production environment and finalized its validation and testing.

It is critical to note that the matching environment was required in order to meet multiple critical parameters set out in the DSA:

1) Enabling the destruction of the project data after a set amount of time
2) Restricting the duplication or capturing of any meta data from tokenized data
3) Meeting FedRAMP Moderate Authorization to Operate (ATO) security requirements

Table 3: Data Sharing Agreement Parameters

Parameter	Data Sharing Agreement Language
Data Destruction	*After 30 days from data delivery to the Secure Data Access Facility, HealthVerity will delete this AWS project specific account, which will destroy all project specific data. A certification of destruction form will be submitted to NCHS and NCSES.*

Parameter	Data Sharing Agreement Language
Restriction on Creating Copies or meta data from Tokenized data	*Except as needed for operational purposes, copies of NCHS and NCSES tokenized data (e.g., paper documents, electronic files, video records, or records of other kinds) are not to be made. Any duplicated copies made of NCHS tokenized data must be documented in Attachment B and must be destroyed as soon as operational requirements permit, using appropriate means of destruction such as shredding, burning, and macerating, or for electronic media intended for reuse, deleting files and then overwriting so that NCHS tokenized data cannot be recovered.*
FedRAMP ATO	*For this project, HealthVerity will leverage its FedRAMP compliant PPRL technology within Amazon Web Services (AWS). As part of the deliverable, HealthVerity will provide documentation and learnings from the process to help inform data sharing and future interagency linking initiatives…NCSES and their contractor, HealthVerity, agrees to provide moderate level security arrangements (as defined by NIST SP 800-60 and FIPS-199) for access to, storage of, and disposition of all files, extracts, printed listings, paper forms, or outputs to prevent unauthorized use of NCHS tokenized data.*

Due to the first two parameters described above, the project team determined that no data or metadata from the tokenization and linkage activities on this project should be used via AI, ML, or traditional means, to enhance HealthVerity's identity resolution capabilities or commercial master patient list, which is used for identity resolution and patient mastering across dozens of other clients. To ensure that all data was destroyed and no copies or metadata from the project data were available after the project, a read-only copy of HealthVerity's de-identified master patient list (as of January 4, 2024) was used to conduct the tokenization and linking. This was housed in a discrete AWS cloud environment intended only for use by this project.

This ensured that data within this project remained isolated, allowing it to be destroyed in accordance with the DSA. Additionally, HealthVerity destroyed this AWS account and all associated materials, including the project-specific master patient list, within 30 days of data delivery into the NCSES SDAF, allowing the project to remain compliant with the DSA.

When HealthVerity was awarded this project in August 2023, it had already obtained an ATO from HHS (October 17, 2022) to use the HealthVerity PPRL software for tokenizing and linking CDC immunization information system (IIS) data under contract #75D30122C13317. At the start of the NCSES project, NCSES and NCHS leadership determined that the HHS-sponsored ATO for CDC linkage did not meet the requirements for the NCSES PPRL project. Therefore, after completing the development of the PPRL infrastructure for this initiative, the HealthVerity team collaborated with NCSES and NCHS to obtain an ATO specific to the NCSES project. The high-level steps can be found below.

1. In April 2024, NCSES began working with their security representatives to obtain ATO sign-off for the project. At the outset, HealthVerity provided NCSES with the required system documentation for review and approval. Over the following 6–8 weeks, NCSES engaged in detailed discussions with their security office to ensure that the DSA accurately reflected the necessary security parameters—a key factor in achieving the ATO. (Note that NCHS deferred to NCSES during these steps and determined that once NCSES obtained the ATO, NCHS would be authorized to move forward with the project).

2. Following this collaboration, the NSF OCIO Change Control Board (CCB) requested a formal briefing from NCSES to evaluate and approve the solution. The team was subsequently invited to present at the NSF Engineering Review Board (ERB), where the NCSES Project Lead shared the PPRL Data Flow Diagram for NCHS and NCSES Linkage (see Figure 1). At that point, NSF had the information it needed to initiate the agency-specific ATO request.
3. Two months later, the agency approved the DSA, enabling the issuance of the ATO. NCHS then reviewed the DSA and ATO package, and agreed to move forward with the project, making the matching environment fully available for tokenizing and linking the project data.

Preparation of NHIS and SED data to send to PPRL environment

After provisioning the PPRL environment and ATO approval, HealthVerity met with NCHS and NCSES to configure their respective PPRL DeID Engines (JAR files, or Java ARchive executables). The engine is installed behind the data owner's firewall and transforms sensitive individual-level records into de-identified formats, enabling secure delivery into HealthVerity's centralized matching service for the assignment of unique and persistent HealthVerity IDs (HVIDs).

The HealthVerity Integrations Team conducted individual technical meetings with NCHS and NCSES in order to align on data source layouts and ensure accurate mapping of input fields to hashed identifiers. This included agreement on field types, header locations, data formats (e.g., YYYY-MM-DD v. MM-DD-YYYY; Male v. M), and join keys to apply the de-identification logic, and support downstream integration with covariate data after HVID assignment.

Name handling and alternative survey records

In preparing data for optimal matching in the PPRL process, HealthVerity typically recommends that middle names and initials, prefixes (e.g., Mr., Dr.), and suffixes (e.g., Jr., Sr.) are excluded from first and last name fields.

While HealthVerity's AI-powered matching engine can account for common nicknames (e.g., "Elizabeth," "Liz" and "Lizzy" all pertaining to the same person), the way some of the source data were structured supported a slightly different approach to handle name variation for their survey records. The source data structure maintained a primary and secondary relationship between a respondent's suspected legal name (primary) and alternative names, nicknames, or ambiguity around which name is the first name or middle name (secondary), grouping them under one survey identifier.

Figure 2: Illustration of primary/secondary name relationship issue

Survey ID	First Name	Last Name	DOB	Zip	Covariate Data
00001	Jonathan	Thomas	1/1/60	12345	xxxx
Potential Issues with these records	John	Thomas	1/2/60	12345	xxxx
	Thomas	John	1/3/60	12345	xxxx
	Johnny	Thomas	1/4/60	12345	xxxx
	Jonathan	Thomas	1/5/60	12345	xxxx
	Thomas	Jonathan	1/6/60	12345	xxxx

Note: Fabricated data used above

This structure conflicted with HealthVerity's matching logic, which treats each row as a unique individual. If a nickname (secondary) row were selected as the match, rejoining to the primary survey ID would fail due to the missing survey ID reference.

To resolve this, the team proposed and implemented a two-part solution for both NHIS and SED data:

Add a unique row ID for each name variation (see Figure 2) and include survey ID as a covariate field.

Figure 3: Illustration of proposed solution to primary/secondary name relationship issue

Survey ID	First Name	Last Name	DOB	Zip	Covariate Data
00001	Jonathan	Thomas	1/1/60	12345	xxxx
00002	John	Thomas	1/2/60	12345	xxxx
00003	Thomas	John	1/3/60	12345	xxxx
00004	Johnny	Thomas	1/4/60	12345	xxxx
00005	Jonathan	Thomas	1/5/60	12345	xxxx
00006	Thomas	Jonathan	1/6/60	12345	xxxx

Note: Fabricated data used above

This solution of adding row IDs, which uniquely identifies each row of data, allows the PPRL engine to de-identify each row separately and run each row against the master patient list during match processing. The survey ID can then be used to identify rows that belong to the same survey record and deduplicate them post-matching. These rows, called alternative survey records, carry identical data in all fields (both PII and covariates) except for first and last name and have the same survey ID but different row IDs. Alternative survey records serve to maximize the probability of matches to HVIDs by supplying the PPRL environment with multiple options for name variations. Note that after the PPRL process, alternative survey records were deduplicated to ensure there is only one row per survey record in the analysis of the linked NHIS-SED file.

(Note that these records were destroyed in accordance with the project DSA, however retention of these records for a longer duration during the project could have supported further data quality reviews that were critical to the project during Task 3. See Task Area 3, Lesson Learned #3 for more details)

Delivery of data to PPRL environment

Once the data were prepared, each agency ran their PII data through the HealthVerity PPRL DeID Engines on their respective servers and delivered the hashed values to HealthVerity for PPRL match processing. The following steps were involved in this process:

- **Data staging.** NCHS and NCSES extract and stage their data within flat files (e.g., psv, csv) for processing. For each agency, the first flat file contains all non-PII fields (i.e., covariate data), along with a row ID and survey ID, to be delivered directly to the NCSES SDAF. The second flat file contains the PII, along with a row ID, to be processed through the PPRL DeID Engine and delivered directly to the FedRAMP Moderate AWS HealthVerity environment.
- **PPRL DeID Engine ingestion.** NCHS and NCSES process the PII flat file through the HealthVerity PPRL DeID Engine utilizing a command line interface (CLI). The PPRL DeID Engine replaces the PII values with one-way cryptographic hashes. To do this, the DeID Engine appends a small value to each piece of PII, called a salt, before applying a cryptographic hash function taking the input and producing a fixed-size 256-bit (32-byte) output called a hash (SHA256). This salt was randomly generated by a trusted third-party cloud provider and then built into the DeID engine at compile time. This allows the salt to be consistent every time the application runs, as well as being consistent across NCHS and NCSES. HealthVerity does not have access to that salt.
- **Data delivery to PPRL Matching Environment.** The hash file is the result of replacing the PII values with one-way cryptographic hashes. The PPRL DeID Engine automatically encrypts this file with two layers of encryption. The first encryption method, symmetric encryption (AES256), ensures the records within the file are encrypted at rest and in transit. The second encryption method, asymmetric encryption (RSA 4096), encrypts the overall file utilizing a public key. The

private key, which is held only by HealthVerity's matching environment, is able to decrypt this file, however the underlying cryptographic hashes remain intact. Each DeID Engine uses its own encryption key to ensure that the hashes cannot be intercepted or deciphered by anyone other than the intended destination. Although the original plan was for NCHS and NCSES to both deliver this file to HealthVerity via SFTP for PPRL match processing, a later change was required for NCHS. Due to the Centers for Disease Control and Prevention (CDC) no longer permitting SFTP traffic into or out of their environment, AWS CLI was adopted and implemented successfully for NCHS.

4.3. Summary of Results/Output

PPRL match processing and HVID assignment

The PPRL Matching Environment is an automated sequestered server managed by a HealthVerity trusted third party cloud provider, therefore enabling greater security and separability. The HealthVerity PPRL Match Processing Service is a centralized probabilistic matching engine that matches each set of hashes against a shared master patient list, using any and all evidence in the hash to accumulate the evidence of the best match, providing high levels of precision and recall based on previous research[5]. The processing service creates universally unique and persistent HVIDs for each file delivered to HealthVerity by NCHS and NCSES. These resolved HVIDs are then delivered to HealthVerity's De-Identified Environment through HealthVerity's internal SFTP, and HealthVerity's integration engineers ensure data quality on the resolved HVIDs. Before delivering the data to the NCSES SDAF via SFTP, HealthVerity salts and hashes the HVIDs uniquely for this project. This ensures the HVIDs can only be used for this use case and cannot be linked to other HVIDs outside of this project. HealthVerity does not have access to the NCSES SDAF, The resulting HVIDs are delivered back to NCSES SFTP pickup folder, where they were manually moved to the NCSES SDAF.

Null and valid HVIDs

During PPRL match processing, each row processed is either matched to a valid HVID or assigned a null HVID. A null HVID indicates that the row did not meet the requirements for the PPRL process to match it to a valid HVID. This occurs when 1) the row has data quality issues in the necessary PII fields (i.e., invalid record), or 2) the row has multiple candidates of HVIDs or does not meet HealthVerity's confidence threshold for matching (i.e., non-matches).

Partial and full dataset submission

In order to complete end-to-end testing of the entire PPRL solution and process, including matching and HVID assignment, HealthVerity recommends that each agency submit a partial dataset (10,000 rows) of their PII data through the PPRL DeID Engine and submit the resulting hash files to HealthVerity via their data transfer connection.

This allows HealthVerity to complete the matching and quality assurance process (QA) of the partial datasets, to determine if there are any underlying data quality issues that negatively impact match rates. Two primary QA controls include verifications on record validity and match rate. Valid records are those which contain the necessary PII fields for matching (first name, last name, date of birth, and zip code). A partial dataset containing more than 5% invalid records triggers review. HealthVerity's standard matching acceptance rate is 95% of total valid records, with a 5% deviation allowance (90-100%) depending on the use case. Due to the extensive nature of HealthVerity's data ecosystem, typical match rates are 97-99% of valid records.

[5] https://surveillance.cancer.gov/reports/TO-P1-PPRLS-Landscape-Analysis.pdf; https://surveillance.cancer.gov/reports/TO-P2-PPRLS-Evaluation-Report.pdf

After both agencies successfully completed end-to-end testing of the entire PPRL solution and process, including matching and HVID assignment of a partial dataset file, both agencies proceeded with full dataset submission.

Results of PPRL match processing and HVID assignment

The results of the PPRL HVID match processing of the full datasets from each survey are provided in Table 1. Out of all records processed, records are either deemed valid or invalid based on data quality of the necessary PII fields. For rows deemed valid, each row is either matched to an existing HVID, matched to a new HVID, or non-matched due to not meeting HealthVerity's confidence threshold for assigning a HVID. Existing HVIDs refer to HVIDs that existed on HealthVerity's master patient list prior to this project, whereas new HVIDs are those newly generated for this project because the individual did not previously exist on the master patient list. Records deemed invalid and records deemed valid, but were non-matched, are assigned to null HVIDs. The high match rates (above 90% for each survey and combined) shown in Table 1 indicate that the majority of processed rows in this project were matched to valid HVIDs.

Table 4. Matching Results to Master Patient List

	NHIS	SED	Combined
Total records processed	295,772	531,786	827,558
Total invalid	433	51	484
Total valid	295,339	531,735	827,074
Existing HVID matches	258,630	469,019	727,649
New HVID matches	22,157	18,934	41,091
Total non-matches	14,552	43,782	58,334
Match rate	94.93%	91.76%	92.89%

Note: Null HVIDs are assigned to the "total invalid" and "total non-matches" records. The match rate is calculated as the number of "existing HVID matches" and "new HVID matches" divided by the total number of records processed.

Artificial Intelligence Use in the Project

On April 3, 2025, the administration released the following AI-related memos:

- Driving Efficient Acquisition of Artificial Intelligence in Government
  https://www.whitehouse.gov/wp-content/uploads/2025/02/M-25-22-Driving-Efficient-Acquisition-of-Artificial-Intelligence-in-Government.pdf
- Accelerating Federal Use of AI through Innovation, Governance, and Public Trust
  https://www.whitehouse.gov/wp-content/uploads/2025/02/M-25-21-Accelerating-Federal-Use-of-AI-through-Innovation-Governance-and-Public-Trust.pdf
- Fact Sheet: Eliminating Barriers for Federal Artificial Intelligence Use and Procurement
  https://www.whitehouse.gov/wp-content/uploads/2025/02/AI-Memo-Fact-Sheet.pdf

In compliance with these memos, the project currently deploys flexible machine learning techniques to predict the probability of matches and handle complex relationships between features. HV uses Artificial Intelligence (AI) / Machine Learning (ML) methods as part of the PPRL in the following areas: (*proprietary information is omitted*):

Error correction: Plaintext data is vetted for potential issues before encoding at the data provider's site. This is performed at both the individual row level as well as at the aggregate level to identify inconsistencies that could adversely affect matching later.

Anomaly detection: Also operating on the plaintext data, anomaly detection is employed to detect unusual patterns in the data that could indicate discrepancies in the data. For instance, if there is an unusual distribution of the date of birth, this could indicate that a different date field was incorrectly mapped in.

Probabilistic matching: The matching model framework developed by HealthVerity is itself probabilistic AI, built as a multi-target tracking (MTT) problem.

Migration models: Patient locations change over time, as represented in three-digit zip codes. These movements are not random, and depend on multiple factors, including the source location and the patient's age

and gender. The solution uses a Zero-shot learning model to understand these movements to compute the probability of a match.

Name distributions: The matching engine is unable to use external data to identify the frequency and conditional distributions of names. Instead, it learns these interactions using the unlabeled data itself.

Unobserved population: HealthVerity's solution uses AI modeling, based on internally observed statistics combined with external estimates of population distributions and change rates, to predict what the unobserved individuals may look like to better identify when a patient is truly a new individual versus a change in an existing patient or even just a noisy record.

Delivery of data to the NCSES Secure Data Access Facility (SDAF)

Once the PPRL match processing and HVID assignment was completed, two types of files were then delivered to the NCSES SDAF environment so that Mathematica could construct the linked NHIS-SED file, which was then used for the analyses described in section 5 (as further detailed in section 4.1, the NCSES SDAF is a secure compute environment managed by the National Opinion Research Center (NORC), which is the analytic environment approved via the DSA where all analysis on the project were to take place). Below, we describe the two types of files that need to be delivered to the NCSES SDAF for these analyses.

PPRL response files

For each survey, the PPRL response file is the file returned from the PPRL environment and contains the assigned HVIDs and row IDs for each processed record. HealthVerity delivered the NHIS and SED PPRL response files to NCSES, via SFTP. NCSES then ingested the NHIS and SED response files into Mathematica's project workspace, within the NCSES SDAF.

Covariate files

For each survey, the covariate file contains the non-PII, analytic variables ("covariates"), along with the row IDs and survey IDs, for the survey records included in the PPRL process. As mentioned in section 4.1, the covariate files are directly transferred from each agency to the NCSES SDAF without entering the HealthVerity FedRAMP PPRL environment. For this project, NCHS delivered the NHIS covariate file to NCSES via a secure data transfer method. NCSES then transferred both the NHIS and SED covariate files into Mathematica's project workspace, within the NCSES SDAF.

Data destruction and shutdown of PPRL environment

The DSA included a requirement for HealthVerity to shut down the PPRL environment provisioned for this project, as well as destroy all data related to this project, within 30 days of data delivery to the NCSES SDAF. HealthVerity completed the following; (a) data destruction and provided NCSES and NCHS with a Certificate of Data Destruction and (b) shutdown of the PPRL Matching Environment.

4.4. Lessons Learned

As stated previously, this project was purposely defined as a demonstration project, with the goal of using it as a template and to inform future NSDS projects and initiatives. Therefore, the documentation of pertinent Lessons Learned was especially vital to the overall success of this project. As the project team worked through the four Task Areas of the project, Lessons Learned were discussed and documented by the core project team. On a quarterly basis, the Lessons Learned were reviewed and approved by the AOR.

A complete list of all of the approved Lessons Learned from this project can also be accessed and reviewed at: https://www.americasdatahub.org/adc-lessons-learned-pprl1-23-n03/.

During the completion of Task Area 2, the following Lessons Learned were documented and approved:

1. It is important to have discussions about the data flow. The team developed a plan to limit the covariate data exposure to a trusted third-party cloud provider by adding a step in the process to separate the hashed personally identifiable information data from the covariate data until a HealthVerity ID (HVID) has been assigned.
2. Data quality, particularly levels of missingness, should be considered prior to the development of hashed tokens
3. It is important to have discussions about name fields and name variants prior to the creation of hashed tokens. This type of name standardization can aid in the linking process and should be agreed upon by both parties prior to deployment of encryption tools.
4. When working with a commercial privacy preserving record linkage (PPRL) vendor (such as HealthVerity) for a PPRL solution, it is important to test data transfer methods prior to full scale implementation to ensure they are meeting agency requirements and permissions and to troubleshoot for future implementation.
5. The NSDS is envisioned as a government-wide set of shared services for data and evidence building. It will be used by a diverse group of individuals with differing levels of technical, analytical, and informatics experience. Therefore, technical assistance materials (e.g., PPRL related documentation, FAQs, training manuals, etc.) should be created in a manner that will be usable by a diverse set of users. In this specific project, the supporting PPRL Documentation provided by HealthVerity is currently written for technical IT staff. This can be problematic for someone who doesn't have a technical IT background. The team on this project has worked together jointly with NCSES to create user friendly documentation for different skill levels and expertise in IT infrastructure.

5. Task Area 3 - Validation Statistics and Modeling (VSM)

5.1. Overview

After the files were successfully de-identified, tokenized, linked to HealthVerity's master patient list (or assigned a new HVID), and delivered to the NCSES SDAF, the HealthVerity team relied on Mathematica to combine the NHIS and SED data using HVID and perform the required analyses of the linked NHIS-SED dataset. As noted in the RFS, the project's three overarching aims included (1) demonstrating the ability to establish a DSA between two federal agencies, (2) applying PPRL to two datasets to link the data, and (3) analyzing the resulting linked data file in a secure compute environment. Led by the subcontractor Mathematica, Task Area 3 - Validation Statistics and Modeling (VSM) was designed to fulfill the third of these requirements.

As noted in previous sections, the two datasets used in this demonstration project are the National Health Interview Survey (NHIS) and the Survey of Earned Doctorates (SED). Both are described in detail below:

National Health Interview Survey (NHIS)
Conducted by NCHS since 1957, NHIS is a cross-sectional household survey that collects information on a range of health-related topics, including illness and chronic conditions, injuries and chronic pain, health-related behavior, functional limitations, healthcare access and use, health insurance coverage, and preventive health services. The survey is representative of the U.S. civilian, non-institutionalized population. For this project, NHIS data from 2012 to 2022 were used, limited to sample adult respondents only. Key analytic variables for this project from NHIS included survey year, education level, sex, race/ethnicity, and functional limitations.

Survey of Earned Doctorates (SED)
Conducted by NCSES since 1958, SED is an annual census of all new research doctorate recipients from accredited U.S. institutions. It collects information about graduates' educational history, funding sources, and post-graduation plans. For this project, SED data from 2012 to 2022 were used. Key analytic variables for this project from SED included year of doctorate, education history, sex, race/ethnicity, and functional limitations.

Overview of the linked NHIS-SED file
The linked file was anchored by NHIS, such that it included all eligible survey records (see section 5.3.1 for linkage and data eligibility definitions) from NHIS. Each record was either linked or not linked to a SED record. For NHIS records that were linked, data for both NHIS and SED variables were included; for records that were not linked, only NHIS variables were available. A detailed description of the process used to construct the linked NHIS-SED file is provided in section 6.2.

This section provides a detailed overview of the steps taken to prepare and analyze the data in accordance with the project requirements, including key considerations presented by Mathematica for interpreting the PPRL linkage results and potential use of the linked NHIS-SED file if it were to be released.

5.2. Summary of Steps Taken

HVID assignment among alternative survey records

After the files were moved to the NCSES SDAF, but before constructing the linked NHIS-SED file, Mathematica conducted data quality checks on each survey to assess the extent to which rows that belong to the same survey record (i.e., alternative survey records as previously described in PPRL

section) were assigned to the same HVID during the PPRL match processing. There were five possible scenarios of HVID assignment among alternative survey records:

1) All rows were assigned to a null HVID only
2) All rows were assigned to a single valid HVID only
3) Rows were assigned to a null HVID and a single valid HVID
4) Rows were assigned to a null HVID and multiple valid HVIDs
5) Rows were assigned to multiple valid HVIDs only

Alternative survey records should ideally be assigned to a single HVID given that in truth, these rows belong to the same survey record, meaning they belong to the same survey participant. However, it is possible for alternative survey records to be assigned to multiple different HVIDs due to the name variation (as previously described in PPRL section), which may have resulted in multiple matches of HVIDs that meet the confidence threshold for HVID assignment among those rows. This may especially be the case for participants with common names in densely populated areas, such that it is more likely to have multiple individuals with the same name in the same zip code.

In this project, Mathematica found that a large proportion of alternative survey records were assigned to multiple HVIDs. The true causes of multiple HVIDs among alternative survey records are unknown, given that in this project, the hash files were destroyed prior to conducting the data quality checks. Therefore, we recommend conducting this check before the destruction of data in the PPRL process, to enable further investigations into why certain alternative survey records were assigned to different HVIDs.

Construction of the linked NHIS-SED file

Starting source files and identifier columns

The construction of the linked NHIS-SED file began with four separate source files (see Figure 3): the NHIS PPRL response file, NHIS covariate file, SED PPRL response file, and SED covariate file. Data in all example illustrations in this report are fake and are for demonstrative purposes only.

Figure 4. Example illustration of the four source files used to construct the linked NHIS-SED file

1. NHIS PPRL response file

HVID	NHIS row ID
123abc	1
123abc	2
000jkl	3
456def	4
NULL	5
789ghi	6
NULL	7

2. NHIS covariate file

NHIS row ID	NHIS survey ID	NHIS covariates
1	1	…
2	1	…
3	1	…
4	2	…
5	3	…
6	3	…
7	4	…

3. SED PPRL response file

HVID	SED row ID
123abc	10
000jkl	11
NULL	12
NULL	13
111zzz	14
456def	15
456def	16

4. SED covariate file

SED row ID	SED survey ID	SED covariates
10	10	…
11	10	…
12	11	…
13	12	…
14	12	…
15	13	…
16	14	…

Note: All data in the figure are fabricated and for illustrative purposes only. For each survey, there is an exact one-to-one match of row IDs between its PPRL response file and covariate file.

As described previously, the covariate file contained the non-PII, analytic variables (“covariates”) for each survey. Each row of the covariate file was identified by a unique row ID, and a unique survey record was identified by a survey ID. In the covariate file, alternative survey records, as discussed in section 5.3.2, were rows with the same survey ID but different row IDs. Rows with the same survey ID carried identical

covariate data since they belong to the same survey record. These rows were deduplicated during the construction of the linked NHIS-SED file.

As also discussed previously, the PPRL response file was the file returned from the PPRL process which contained the assigned HVID for each row of survey records included in the PPRL process for each survey. Each row of the PPRL response file was identified by a unique row ID, which corresponded exactly to the row IDs of the survey's covariate file. If a HVID appeared in both NHIS and SED, those survey records associated with that common HVID "linked" in the linked NHIS-SED file (see "row type A" in Figure 5).

Table 5 summarizes the roles of the key identifier columns (row ID, survey ID, and HVID) in the construction of the linked file.

Table 5. Description of key identifier (ID) columns used to construct the NHIS-SED linked file

Identifier column	Source file	Purpose	Possible to be duplicated in source file (multiple rows of source file have the same ID value)	Role in construction of linked NHIS-SED file	Possible to have missing ID value
NHIS survey ID	NHIS covariate file	Identify a unique NHIS record	Yes – possible when multiple rows exist for the same survey record due to creation of alternate survey records (see section 5.3.2)	Identify rows that belong to the same survey record and deduplicate	No
SED survey ID	SED covariate file	Identify a unique SED record	Yes – possible when multiple rows exist for the same survey record due to creation of alternate survey records (see section 5.3.2)	Identify rows that belong to the same survey record and deduplicate	No
NHIS row ID	NHIS PPRL response file and covariate file	Identify a unique row	No	Merge NHIS PPRL response file with NHIS covariate file	No
SED row ID	SED PPRL response file and covariate file	Identify a unique row	No	Merge SED PPRL response file with SED covariate file	No
HVID	NHIS PPRL response file; SED PPRL response file	Identify a unique individual	Yes – possible when an individual completed a given survey multiple times, or rows that belong to the same survey record were assigned to the same HVID during PPRL process	Merge NHIS and SED	Yes – a null HVID indicates that the PPRL process could not assign a row to an individual (see section 5.5.1)

End product: linked NHIS-SED file

Figure 5 is an example illustration of the linked NHIS-SED file once it has been created.

Figure 5. Example illustration of linked NHIS-SED file

HVID	NHIS row ID	NHIS survey ID	NHIS covariates	SED row ID	SED survey ID	SED covariates	Row type
123abc	1	1	…	10	10	…	A: Linked – valid HVID
456def	4	2	…	15	13	…	A: Linked – valid HVID
789ghi	6	3	…				B: Non-linked – valid HVID
NULL	7	4	…				C: Non-linked – NULL HVID

Note: All data in the figure are fabricated and for illustrative purposes only.

The linked file contained three types of rows (A, B, and C):

- **Type A: linked rows with a valid HVID**, where a NHIS record was linked to a SED record because they were associated with the same individual (same HVID assigned by PPRL process).
- **Type B: non-linked rows with a valid HVID**, where a NHIS record was not linked to any SED record because no SED records had the HVID associated with that NHIS record.
- **Type C: non-linked rows with a null HVID,** where a NHIS record was not linked to any SED record because no HVID was assigned to that NHIS record during the PPRL process). NHIS records with null HVIDs were included in the linked file as row type C, given that their covariate data were valid and could be used in certain analyses.

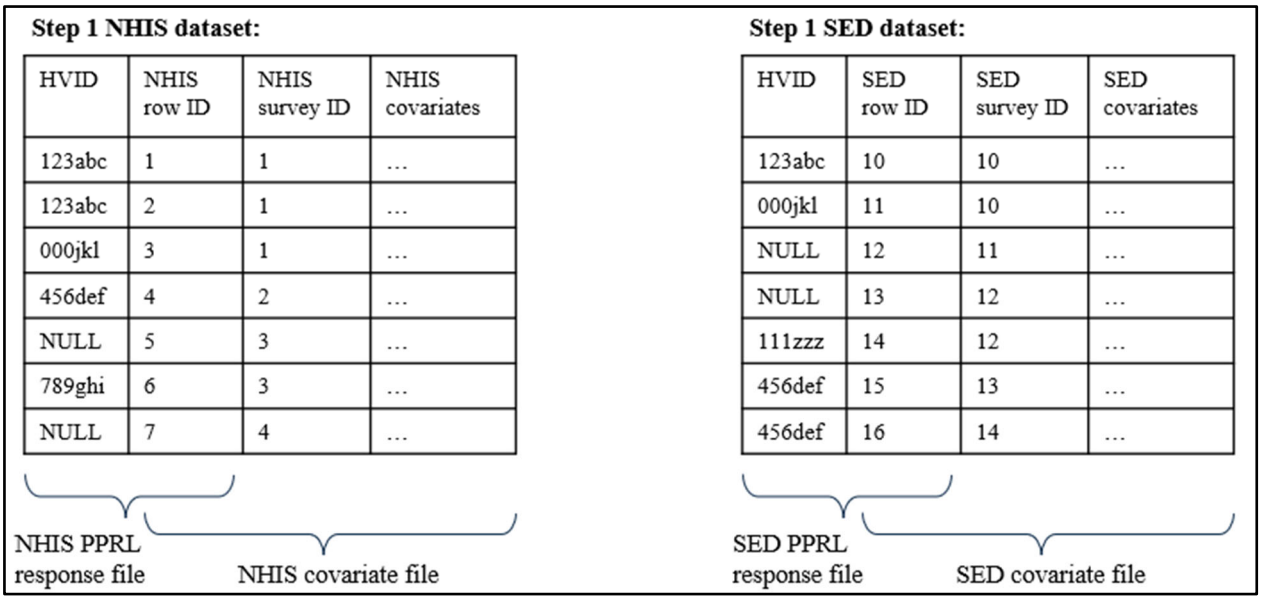
Each row of the linked NHIS-SED file was a unique NHIS record identified by a unique NHIS survey ID. Because multiple rows could exist for the same survey record within the starting *source* files (i.e., alternative survey records discussed in section 5.3.2), these rows were deduplicated during the construction of the linked file to ensure there was ultimately only one row per survey record in the linked file. Importantly, this deduplication only occurred *after* linking HVIDs across NHIS and SED, to ensure maximum linkage when alternative survey records were assigned to multiple different HVIDs.

Step-by-step process to create the linked NHIS-SED file

Step 1: Merge the PPRL response file with the covariate file for each survey (Figure 6).

In each survey, Mathematica merged the PPRL response file with the covariate file using row ID. When merging, there was an exact one-to-one match of row IDs across the PPRL response file and the covariate file per survey.

Figure 6. Illustration of step 1: merge the PPRL response file with the covariate file for each survey

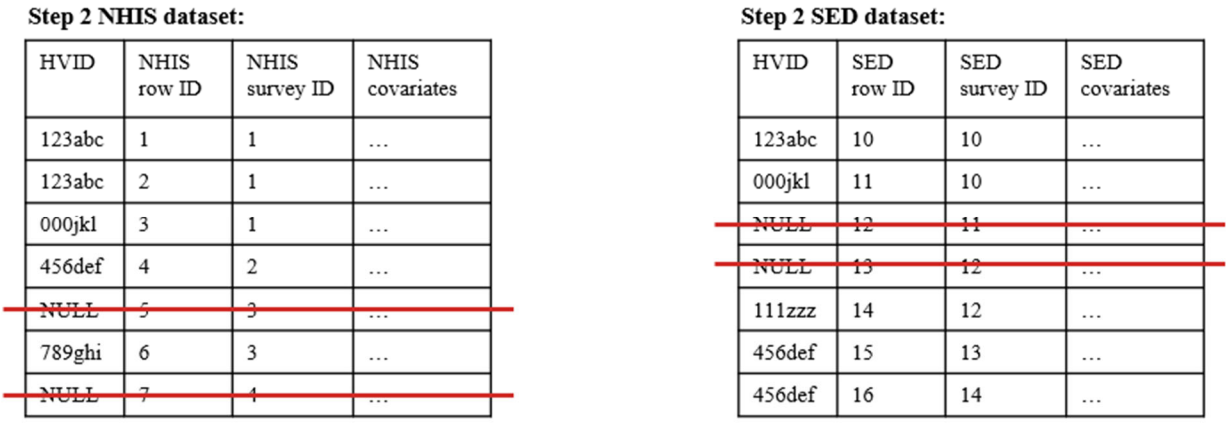


Note: All data in the figure are fabricated and for illustrative purposes only.

Step 2: Remove rows with a null HVID in each survey (Figure 6).

In each survey, Mathematica removed rows with a null HVID. Because null HVIDs by definition do not link across surveys, we temporarily removed them from each dataset in preparation for the NHIS-SED merge in step 3. Note that rows with null HVIDs from NHIS were added back in step 5, given that the linked NHIS-SED file ultimately contained both linked and non-linked rows.

Figure 7. Illustration of step 2: remove rows with null HVID in each survey



Note: All data in the figure are fabricated and for illustrative purposes only.

Step 3: Merge (inner join) the NHIS and SED datasets to create a linked dataset (Figure 8).

Mathematica merged the step 2 NHIS and SED datasets using HVID. The merge was an *inner join*, meaning only the linked rows (row type A), or rows where the HVID appeared in both datasets, were included in the resulting linked dataset.

Figure 8. Illustration of step 3: merge (inner join) the NHIS and SED datasets to create a linked dataset

Step 3 linked dataset:

HVID	NHIS row ID	NHIS survey ID	NHIS covariates	SED row ID	SED survey ID	SED covariates	Row type
123abc	1	1	…	10	10	…	A: Linked – valid HVID
123abc	2	1	…	10	10	…	A: Linked – valid HVID
000jkl	3	1	…	10	10	…	A: Linked – valid HVID
456def	4	2	…	15	13	…	A: Linked – valid HVID
456def	4	2	…	16	14	…	A: Linked – valid HVID

Note: All data in the figure are fabricated and for illustrative purposes only.

Step 4: Deduplicate NHIS survey ID in the linked dataset (Figures 9-11).

The goal of step 4 was to deduplicate rows of the step 3 linked dataset that were associated with the same NHIS survey ID (i.e., retain one row per NHIS survey ID). There were two possible scenarios of NHIS survey ID duplication that could occur in the step 3 linked dataset:

1) A single NHIS survey ID is linked to a single SED survey ID – see step 4a to address.
2) A single NHIS survey ID is linked to multiple SED survey IDs – see step 4b to address. Note that this scenario did not occur in this project.

Figure 9. Illustration of step 4: deduplicate NHIS survey ID in the linked dataset

Step 3 linked dataset:

HVID	NHIS row ID	NHIS survey ID	NHIS covariates	SED row ID	SED survey ID	SED covariates	Row type
123abc	1	**1**	…	10	10	…	A: Linked – valid HVID
123abc	2	**1**	…	10	10	…	A: Linked – valid HVID
000jkl	3	**1**	…	10	10	…	A: Linked – valid HVID
456def	4	**2**	…	15	13		A: Linked – valid HVID
456def	4	**2**	…	16	14		A: Linked – valid HVID

Scenario 1: A single NHIS survey ID is linked to a single SED survey ID **(see Step 4a)**

Scenario 2: A single NHIS survey ID is linked to multiple SED survey IDs **(see Step 4b)**

Step 4a + 4b: Deduplicate

Step 4 deduplicated linked dataset:

HVID	NHIS row ID	NHIS survey ID	NHIS covariates	SED row ID	SED survey ID	SED covariates	Row type
123abc	1	**1**	…	10	10	…	A: Linked – valid HVID
456def	4	**2**	…	15	13	…	A: Linked – valid HVID

Note: All data in the figure are fabricated and for illustrative purposes only.

Step 4a (Figure 10): To address scenario 1, Mathematica retained one row per unique combination of NHIS survey ID and SED survey ID. It did not matter which row, as all rows contained identical covariate information. Therefore, Mathematica simply kept the first row.

Figure 10. Illustration of step 4a: deduplicate NHIS survey ID linked to a single SED survey ID

HVID	NHIS row ID	NHIS survey ID	NHIS covariates	SED row ID	SED survey ID	SED covariates	Row type
123abc	1	1	…	10	10	…	A: Linked – valid HVID
123abc	2	1	…	10	10	…	A: Linked – valid HVID
000jkl	3	1	…	10	10	…	A: Linked – valid HVID

Step 4a:
Deduplicate scenario 1

HVID	NHIS row ID	NHIS survey ID	NHIS covariates	SED row ID	SED survey ID	SED covariates	Row type
123abc	1	1	…	10	10	…	A: Linked – valid HVID

Note: All data in the figure are fabricated and for illustrative purposes only.

Step 4b (Figure 11): To address scenario 2, we retained one SED survey ID per NHIS survey ID. Decisions as to which SED survey ID to retain should be made on a case-by-case basis to identify the most suitable set of SED covariates. For example, we may consider retaining the SED survey ID that was closest in year to its linked NHIS record, to maximize chances of concordance in covariate data between NHIS and SED. Note that scenario 2 did not occur in this project's data, so this is a proposed approach to address the scenario if the scenario were to occur in a different project.

Figure 11. Illustration of step 4b: deduplicate NHIS survey ID linked to multiple SED survey IDs

HVID	NHIS row ID	NHIS survey ID	NHIS covariates	SED row ID	SED survey ID	SED covariates	Row type
456def	4	2	…**Year = 2015**	15	13	…**Year = 2015**	A: Linked – valid HVID
456def	4	2	…**Year = 2015**	16	14	…**Year = 2023**	A: Linked – valid HVID

Step 4b:
Deduplicate scenario 2

HVID	NHIS row ID	NHIS survey ID	NHIS covariates	SED row ID	SED survey ID	SED covariates	Row type
456def	4	2	…**Year = 2015**	15	13	…**Year = 2015**	A: Linked – valid HVID

Note: All data in the figure are fabricated and for illustrative purposes only.

Step 5: Append the deduplicated non-linked NHIS records (Figures 12-13).

The goal of step 5 was to append (i.e., row-wise bind) the non-linked NHIS records (row types B and C) to the step 4 dataset (row type A) to create the linked NHIS-SED file.

To obtain the non-linked NHIS records to append, we started with a copy of the step 1 NHIS dataset. Then, Mathematica limited this dataset to non-linked NHIS records by removing rows where the NHIS survey ID appeared in the step 4 dataset (reminder: the step 4 dataset only includes linked NHIS records). Then, Mathematica deduplicated this resulting dataset by retaining one row per NHIS survey ID, choosing the row with a valid (as opposed to null) HVID, if any. Otherwise, Mathematica simply kept the first row as it did not matter which row was retained, as all rows carried identical covariate information.

Figure 12. Illustration of step 5: deduplicate non-linked NHIS records

Step 1 NHIS dataset:

HVID	NHIS row ID	NHIS survey ID	NHIS covariates	Row type
123abc	1	1	…	A: Linked – valid HVID
123abc	2	1	…	A: Linked – valid HVID
000jkl	3	1	…	A: Linked – valid HVID
456def	4	2	…	A: Linked – valid HVID
NULL	5	3	…	C: Non-linked – null HVID
789ghi	6	3	…	B: Non-linked – valid HVID
NULL	7	4	…	C: Non-linked – null HVID

Remove NHIS survey IDs that appear in Step 4 dataset

Non-linked NHIS records:

HVID	NHIS row ID	NHIS survey ID	NHIS covariates	Row type
NULL	5	3	…	C: Non-linked – null HVID
789ghi	6	3	…	B: Non-linked – valid HVID
NULL	7	4	…	C: Non-linked – null HVID

Retain one row per NHIS survey ID, prioritizing valid HVIDs if any

Deduplicated non-linked NHIS records
(each row is a unique NHIS survey ID):

HVID	NHIS row ID	NHIS survey ID	NHIS covariates	Row type
789ghi	6	3	…	B: Non-linked – valid HVID
NULL	7	4	…	C: Non-linked – null HVID

Note: All data in the figure are fabricated and for illustrative purposes only.

This step resulted in a dataset of deduplicated, *non-linked* NHIS records (green box in Figures 12 and 13), which we appended to the step 4 dataset containing deduplicated, *linked* NHIS records.

Figure 13. Illustration of step 5: append the deduplicated non-linked NHIS records

Step 4 deduplicated linked dataset (each row is a unique NHIS survey ID):

HVID	NHIS row ID	NHIS survey ID	NHIS covariates	SED row ID	SED survey ID	SED covariates	Row type
123abc	1	1	…	10	10	…	A: Linked – valid HVID
456def	4	2	…	15	13	…	A: Linked – valid HVID

Step 5:
Append deduplicated non-linked NHIS records

Deduplicated non-linked NHIS records (each row is a unique NHIS survey ID):

HVID	NHIS row ID	NHIS survey ID	NHIS covariates	SED row ID	SED survey ID	SED covariates	Row type
789ghi	6	3	…				B: Non-linked – valid HVID
NULL	7	4	…				C: Non-linked – null HVID

Note: All data in the figure are fabricated and for illustrative purposes only.

After all five steps were completed, Mathematica arrived at the linked NHIS-SED file, as illustrated in Figure 5. As a reminder, the linked NHIS-SED file contained both linked (row type A) and non-linked (row types B and C) NHIS records. It was also deduplicated to remove alternative survey records, such that each row of the linked file was a unique NHIS record identified by a NHIS survey ID.

Linkage rate

The linkage rate is the number of linked NHIS records (number of type A rows) out of all NHIS records (total number of rows) in the linked NHIS-SED file. It should be noted that a low linkage rate across the two data sets was expected given that only a small proportion of NHIS participants have doctoral degrees. However, a lower than expected linkage rate resulted from this project. The details are explained below.

5.3. Summary of Results/Output

Assessment of linkage results

In data linkage projects without clear text matching, it is often not possible to directly evaluate the accuracy or quality of the linkage results from the PPRL process (i.e., whether two records that were assigned the same HVID truly belong to the same individual). In this project, Mathematica used the linked NHIS-SED file and NHIS participants' self-reported education as a proxy for truth to assess linkage quality based on the concordance of NHIS participants' reported education level (doctoral or not) and their linkage status to SED (linked or not). Because SED is a census of doctoral degree recipients, Mathematica expected that a large proportion of those NHIS records who reported having a doctoral degree would link to SED. For this analysis, we excluded NHIS records with null HVIDs (row type C), given that these records are by definition not linked to SED.

Confusion matrix and performance metrics

Table 6 is a cross-tabulation that illustrates the four possible concordance outcomes from this analysis. Viewing the linkage to SED as the "prediction" and the NHIS education as the "gold standard", we framed Table 6 as a confusion matrix and labeled each concordance outcome as one of true positive, false positive, false negative, or true negative.

Table 6. Cross-tabulation of concordance between NHIS education and linkage to SED

Linked to SED?	Reported doctoral degree in NHIS	Did not report doctoral degree in NHIS
Linked to SED	Concordant (true positive)	Discordant (false positive)
Not linked to SED	Discordant (false negative)	Concordant (true negative)

True positives refer to NHIS participants who reported a doctoral degree and linked to SED. True negatives refer to NHIS participants who did not report a doctoral degree and did not link to SED. False negatives refer to NHIS participants who reported a doctoral degree but did not link to SED. Finally, false positives refer to NHIS participants who did not report a doctoral degree but are linked to SED.

In addition, Mathematica approximated the quality of the linkage results using standard performance metrics for binary classifiers, including sensitivity (also known as recall), specificity, positive predictive value (also known as precision), negative predictive value, and F1 score. Table 7 summarizes these performance metrics, their formulas, and interpretations in the context of this analysis.

Table 7. Summary of performance metrics used to assess linkage results

Performance metric	Formula	Interpretation
Sensitivity (recall)	TP / (TP+FN)	Proportion of NHIS records linked to SED, out of those reporting doctoral degree
Specificity	TN / (FP+TN)	Proportion of NHIS records not linked to SED, out of those not reporting doctoral degree
Positive predictive value (precision)	TP / (TP+FP)	Proportion of NHIS records reporting doctoral degree, out of those linked to SED
Negative predictive value	TN / (FN+TN)	Proportion of NHIS records not reporting doctoral degree, out of those not linked to SED
F1 score	2*Precision*Recall / (Precision + Recall)	Balance of precision and recall

Note: TP = true positive. FN = false negative. TN = true negative. FP = false positive.

Identification of doctoral degree recipients in NHIS

Mathematica classified each NHIS participant as either having a doctoral degree or not having a doctoral degree based on whether their selected education level was a doctoral degree. NHIS records with missing values for the education variable were excluded from this analysis. [6]

Starting in 2021, the NHIS education variable collapsed professional school degrees with doctoral degrees, such that there is one response option for an education level of a “professional school or doctoral degree”, rather than two separate levels. Therefore, it was not possible to determine whether a 2021-2022 NHIS participant who answered having a “professional school or doctoral degree” is a doctoral degree recipient. Given the substantial number of such participants in this study, Mathematica opted to include these records in the analysis and assumed that they are doctoral degree recipients, then conducted a sensitivity analysis to check whether results differ meaningfully when such records were excluded from the analysis.

[6] It should be noted that SED captures all individuals receiving a research doctorate from an accredited U.S. institution in a given academic year while NHIS asks about education attainment without distinguishing the type of doctorate degree and where it was obtained. Therefore, some misalignment between the two sources was expected.

Possible explanations for false positives and false negatives

A false positive or false negative in this concordance analysis does not necessarily imply an inaccurate linkage result, given that there are other plausible reasons for discordance in a NHIS participant's reported education and their linkage status to SED. These scenarios are described below:

- **Pre-2012 doctoral degree recipients.** The range of SED years used in this study was 2012-2022. If a NHIS participant received a doctoral degree prior to 2012, they would report having a doctoral degree in NHIS but would not link to SED, resulting in a false negative (see section 5.4 for a related Lessons Learned).
- **Doctoral degree recipients from non-US institutions.** SED only captures doctoral degree recipients from US institutions. NHIS participants who received a doctoral degree outside of the US would report having a doctoral degree in NHIS but would not link to SED, resulting in a false negative.
- **2021-2022 NHIS participants with professional school degrees.** As discussed previously, the NHIS education variable combined professional school degrees with doctoral degrees starting in 2021. If a NHIS participant reported having a "professional school degree or doctoral degree" but their degree was a professional school and not a doctoral degree (e.g., Doctor of Medicine), they would not link to SED, resulting in a false negative.
- **Participants who obtained their doctoral degree after their NHIS interview.** NHIS participants who completed SED (i.e., obtained their doctoral degree) after NHIS did not yet have a doctoral degree at the time of their NHIS interview. Their records would link to SED but would not indicate a doctoral degree in NHIS, resulting in a false positive.
- **Misreported education.** Accidental or intentional misreporting of education in NHIS can result in either a false negative or false positive.

Given these scenarios where discordance is expected, Mathematica reiterates that this analysis only provides an approximate evaluation of linkage quality. These scenarios, particularly those resulting in false negatives, also provide plausible explanations for the low linkage rate encountered in this project.

For scenarios listed above that could be identified in the data (e.g., Mathematica could identify participants who completed NHIS before SED using the year covariates from each survey), Mathematica conducted the analysis including those records, then repeated the analysis excluding those records to assess whether there are substantial changes to the performance metrics.

Assessment of concordance of shared covariates across surveys

Among NHIS participants that linked to SED ("linked records"), Mathematica assessed the extent to which self-reported data were consistent across surveys for shared variables, where shared variables refer to covariates that were available on both NHIS and SED. For example, if a participant reported being Hispanic in one survey, did they also report being Hispanic in the other survey? This analysis was limited to linked records (row type A) only, and linked records with missingness in the shared variable being analyzed were excluded from the analysis.

Time-invariant covariates

Covariates that generally do not change over time, such as race/ethnicity and sex, are straightforward to analyze. Mathematica compared each participant's response to the covariate in NHIS with their response to the same covariate in SED and evaluated the concordance using a cross-tabulation, percentage agreement, and Cohen's kappa.

Time-variant covariates

For covariates that change over time, such as education level, we considered the timing at which SED and NHIS were completed for a given linked record and temporally aligned the responses to accurately assess the concordance. Below, is a description of the methodology used to temporally align participants' education across surveys based on the timing of surveys. The methodology is also summarized in Table 8.

Table 8. Summary of concordance assessment of education based on timing of NHIS and SED

Timing of surveys	Assess concordance between NHIS and SED: NHIS	Assess concordance between NHIS and SED: SED
NHIS before SED	Reported education level	Highest reported education obtained before or in the same year as NHIS
NHIS in the same year as SED	Reported education level	Doctoral degree
NHIS after SED	Reported education level	Doctoral degree

There were three possible timings in which a linked participant could have completed NHIS and SED:

1) NHIS before SED (NHIS year is less than SED year)
2) NHIS in the same year as SED (NHIS year is equal to SED year)
3) NHIS after SED (NHIS year is greater than SED year)

For participants who completed NHIS in the same year as or after SED, meaning they had a doctoral degree at the time of their NHIS interview, their reported education in NHIS should be a doctoral degree for their education information to be considered consistent across surveys. An exception is participants who completed NHIS in the same year but an earlier month than SED, in which case they would not technically have a doctoral degree. While it is possible for such participants to not report a doctoral degree in NHIS, it is also possible that they did, especially if they were close to graduating. Given that either situation is possible, we opted to define survey timing using year only. Mathematica also explored a looser survey timing definition, where participants who completed NHIS and SED within one year of one another, rather than strictly in the same year, were considered to have taken the surveys concurrently.

For participants who completed NHIS before SED, meaning they did not have a doctoral degree at the time of the NHIS interview, we used the education history variables from SED to assess whether their reported education in NHIS is consistent with the education that the participant had attained at that time, according to their education history reported in SED. More specifically, a participant's reported education in NHIS should be consistent with the highest degree reported in SED that was obtained before or in the same year as NHIS. For example, if a participant completed NHIS in 2017, SED in 2020, and they indicated having a master's degree in NHIS, we expect that in SED, they reported having obtained a master's degree before or in 2017, with the master's degree being the highest degree that they obtained before or in 2017.

Mathematica used the following education history variables from SED: first associate degree year, first bachelor's degree year, first master's degree year, and professional doctorate year. For a given linked record, we limited their degrees to those obtained before or in the same year as NHIS, and out of those degrees, selected the highest degree. Then, Mathematica compared that selected degree to the degree

reported in NHIS and evaluated the concordance using a cross-tabulation, percentage agreement, and Cohen's kappa.

Covariates with different response options across surveys

Some covariates, such as functional limitations, were available on both surveys but had differing response options across surveys. To assess concordance for any such covariates, Mathematica collapsed certain levels to construct variables that were comparable across surveys. Below, Mathematica discusses the methodology used to construct a binary functional limitations status (with or without functional limitations) variable, per functional domain and overall, in each survey. The methodology is also summarized in Table 9.

Table 9. Definitions of functional limitations status for NHIS and SED

Functional limitations status	NHIS difficulty levels included in this status category	SED difficulty levels included in this status category
Without functional limitations	No difficulty; Some difficulty	No difficulty; Slight difficulty
With functional limitations	A lot of difficulty; Cannot do at all	Moderate difficulty; Severe difficulty; Unable to do

The following functional domains were available in both NHIS and SED: seeing, hearing, walking, and remembering or concentrating.

For NHIS, the response options for each functional domain were *no difficulty*, *some difficulty*, *a lot of difficulty*, and *cannot do at all*. In each domain, Mathematica used the levels *a lot of difficulty* and *cannot do at all* to identify those with functional limitations, and the levels *no difficulty* and *some difficulty* to identify those without functional limitations. This threshold for the determination of functional limitations status is consistent with the recommendation from the Washington Group on Disability Statistics' analytic guidance.[7]

For SED, the response options for each functional domain were *no difficulty*, *slight difficulty*, *moderate difficulty*, *severe difficulty*, and *unable to do*. In each domain, we used the levels *moderate difficulty*, *severe difficulty*, and *unable to do* to identify those with functional limitations, and the levels *no difficulty* and *slight difficulty* to identify those without functional limitations.

In addition to the domain-level functioning limitations status, Mathematica created an overall functioning limitations status calculated across the four shared domains for each survey. Mathematica determined the overall functional limitations status for each survey, consistent with the recommendation from the Washington Group on Disability Statistics, as follows:

- Missing functional limitations status: all domains were missing a functional limitations status
- With functional limitations: functioning status was "with functional limitations" (i.e., *a lot of difficulty* or *cannot do at all* for NHIS; *moderate difficulty*, *severe difficulty* or *unable to do* for SED) in at least one domain
- Without functional limitations: all other instances

Using these functioning limitations status variables that are now aligned to be comparable across surveys, Mathematica evaluated the concordance using a cross-tabulation, percentage agreement, and Cohen's kappa.

[7] See https://www.washingtongroup-disability.com/fileadmin/uploads/wg/Documents/WG_Document__5B_-_Analytic_Guidelines_for_the_WG-SS__SAS_.pdf for additional information.

Reporting sample characteristics

To understand the characteristics of survey participants in comparison to the subset of participants in the study who linked, Mathematica examined the distribution of key demographic variables (e.g., age, sex, race/ethnicity, education) within each sample.

Linkage eligibility-adjusted weights

For NHIS, Mathematica used sample weights that have been adjusted for linkage eligibility when reporting weighted survey estimates. The original NHIS sample weights account for differential non-response and selection probabilities across subgroups and are post-stratified, such that the sample matches known population totals for key demographic characteristics to ensure estimates are nationally representative. Because the study sample is limited to NHIS participants who are linkage eligible, which may not be a random sample of NHIS participants (i.e., linkage eligible participants differ in key demographic characteristics from linkage ineligible participants), the sample weights need to be further adjusted to account for this potential bias. Additional details on the weighting adjustment methodology are provided in Appendix III of NCHS' report on the Linkage of NCHS Population Health Surveys to Administrative Records from Social Security Administration and Centers for Medicare & Medicaid Services.[8]

Further, Mathematica reported weighted survey estimates separately for years prior to 2019 from years 2019 and onwards, given the NHIS redesign in 2019 which introduced substantial changes to the survey weighting and design methodology.[9]

For SED, sample weights are not available given that SED is a census.

Additional analyses and covariates

The analyses described above are not an exhaustive list of analyses that can be conducted on a linked file. While the scope of analysis for this project was limited by a low linkage rate, with more linked records, additional analyses may be possible. For example, the project team may consider utilizing covariates that are not shared across surveys to generate analytic insights for linked records that are not otherwise possible with one survey alone. In the context of NHIS and SED, this could include statistical modeling of the relationship between educational experiences and health outcomes.

When selecting covariates from each survey for inclusion in the covariate file, it is also important to ensure that those variables are available and comparable across the years of the survey that are included in the study. Otherwise, the end results may be a large amount of missingness or may have to separate analyses by survey year if the covariates have changed substantially (e.g., in the response options or in the question prompt) over the years.

5.4. Lessons Learned

As stated previously, this project was purposely defined as a demonstration project, with the goal of using it as a template and to inform future National Secure Data Service (NSDS) projects and initiatives. Therefore, the documentation of pertinent Lessons Learned was especially vital to the overall success of this project. As the project team worked through the four Task Areas of the project, Lessons Learned

[8] Golden C, Driscoll AK, Simon AE, et al. Linkage of NCHS population health surveys to administrative records from Social Security Administration and Centers for Medicare & Medicaid Services. National Center for Health Statistics. Vital Health Stat 1(58). 2015

. Aram J, Zhang C, Golden C, Zelaya CE, Cox CS, Ye Y, Mirel LB. Assessing linkage eligibility bias in the National Health Interview Survey. National Center for Health Statistics. Vital Health Stat 2(186). 2021. DOI: https://dx.doi.org/10.15620/cdc:100468.

[9] See https://www.cdc.gov/nchs/hus/sources-definitions/nhis.htm for additional information.

were discussed and documented by the core project team. On a quarterly basis, the Lessons Learned were reviewed and approved by the AOR.

A complete list of all of the approved Lessons Learned from this project can also be accessed and reviewed at: https://www.americasdatahub.org/adc-lessons-learned-pprl1-23-n03/.

During the completion of Task Area 3, the following Lessons Learned were documented and approved:

1. Prior to the linkage process all parties should discuss the types of questions that can be addressed once the files are linked and identify the key covariates that will be used in the analyses. It is advantageous to share data dictionaries early in the project to allow project team members to better understand the scope of the data. This allows for brainstorming and identifying research questions and analyses that can be conducted with the available data and facilitates discussions on covariate selection prior to starting the analytics work which is particularly important with PPRL.

2. Working in a secure compute environment limits the amount of output that can be shared without a disclosure review assessment. Researchers should plan to have clear analytic plans to ensure transparency in approaches and aid in coordination of output.

3. PPRL requires some a priori planning of data quality metrics for the data sources being linked. These may include distinguishing reasons for non-links (e.g., insufficient or low quality personally identifiable information (PII) or not meeting a linkage threshold cutoff). As practitioners embark on data sharing agreements for PPRL with security/privacy protocols it is worth considering the need to assess data quality metrics prior to any data file destruction and allow sufficient time and space for those assessments to occur. In addition, clear metrics for data quality assessments should be defined and reported as part of the project plan and final report.

4. Documenting the results of linked/blended data is critical to determining the files fitness for use. In the federal statistical system, there may be examples of documentation for linked files. These examples should be used to establish templates for reporting linkage methodology and analytic considerations when working with linked data.

5. One aspect of HealthVerity's standard PPRL implementation and workflow that proved to be very beneficial in this project, was the two-step PPRL Deidentification (DeID) Engine process. The first step involved processing just a Partial Dataset File (10k records) through the DeID Engine for each agency, before proceeding to processing each agency's Full Dataset File through the DeID Engine. Once the Partial Dataset Files were run through the DeID Engine and submitted to HealthVerity for matching and QA, some data quality issues were identified. This provided each agency with the ability to address any data quality issues, before processing their Full Dataset Files through the DeID Engine. followed by HealthVerity conducting a QA

6. When using privacy enhancing technology tools for linkages, researchers should consider conducting pre-linkage assessments of key demographic variable statistics (such as age), in order to assess the alignment of years or other denominators to better inform linkages. Quality control and quality assurance assessments can be difficult when working with encrypted data. Adding initial assessments of summary statistics of the files to be linked prior to linking can shed light on benchmarks which may inform the quality assessments once the files are linked.

   Specifically in this project, a misalignment of the timing of degree attainment and years included in the linked files may have increased the number of false negatives which will impact the utility of the file. This is an additional quality assurance check that was noted but could not be addressed by the standard HealthVerity PPRL implementation process or workflow.

6. Task Area 4 - Project Management (PM)

6.1. Overview

Per the terms of the Statement of Work, HealthVerity provided all project management support for the project. In order to accomplish this, HealthVerity leveraged the following software tools:

- Smartsheet - for the creation and ongoing management of the project plan and timeline, as well as other project management related artifacts such as Risks, Assumptions, Issues and Dependencies (RAID) Log, Stakeholder Register, and Lessons Learned Log.
- Google Suite (Google Mail, Google Calendar, Google Drive, Google Docs, Google Sheets, Google Slides) - for communication, meeting scheduling, and project related documentation (agendas, meeting notes, presentations, CIPSEA compliance tracking, etc.)
- Zoom - for all virtual project-related meetings
- Other software tools as needed, including MS Office, Jira, Slack, and Confluence

6.2. Summary of Steps Taken

Below is a summary of the project management related activities that were provided and supported during the life of the project:

1. The HealthVerity PM developed a project plan and timeline, which was reviewed and approved by all project team members and key stakeholders. Managed and updated the project plan and timeline throughout the project, and distributed it monthly to NCSES, NCHS, and Mathematica.
2. The HealthVerity PM developed, managed, and maintained a RAID Log, Stakeholders Register, and a Lessons Learned Log throughout the project.
3. Project and Task Area Kick-Off Meetings - scheduled, facilitated, agenda, meeting notes/follow-up
    - Project Kick-Off Meeting
    - Task Area 1 Kick-Off Meeting
    - Task Area 2 Kick-Off Meeting
    - Task Area 3 Kick-Off Meeting
4. Bi-Weekly Project Status Meeting - scheduled, facilitated, agenda, meeting notes/follow-up

    Bi-Weekly Project Status Meetings (with the core project team members) were held throughout the duration of the project.
5. Monthly Project Status Reports - as required, and per the format and schedule defined in the Statement of Work, the HealthVerity PM prepared and submitted monthly project status reports to ATI and the NCSES AOR. These reports provided the necessary documentation to support timely invoicing of all completed deliverables and milestones.
6. CIPSEA Compliance - In order to fully comply with the CIPSEA requirements for the combined HealthVerity and Mathematica project team members, the HealthVerity PM coordinated the timely

completion of the CIPSEA requirements as well as the submission of the required documentation to NCSES and NCHS.

7. CPDSP Compliance - A Confidentiality Plan and Data Security Procedures (CPDSP) was required for protecting privacy and other sensitive information. A CPDSP involves implementing measures to prevent unauthorized access, use, or disclosure of confidential data. The CPDSP includes procedures that encompass physical security, access controls, data encryption, incident response plan, data backup and recovery, secure disposal of data, and employee training. In order to comply with the CPDSP requirement, HealthVerity and Mathematica worked with NCSES for approval of their respective CPDSPs.

6.3. Summary of Results/Output

Although flexibility was required throughout this demonstration project - largely due to its “proof of concept” nature - the combined project team was able to complete all deliverables and milestones to support the successful completion of this project within the defined period of performance. When one considers that the two agencies involved had never previously shared or linked data with one another - nor worked together to negotiate and fully execute a joint DSA - this was a remarkable accomplishment.

6.4. Lessons Learned

As stated previously, this project was purposely defined as a demonstration project, with the goal of using it as a template and to inform future National Secure Data Service (NSDS) projects and initiatives. Therefore, the documentation of pertinent Lessons Learned was especially vital to the overall success of this project. As the project team worked through the four Task Areas of the project, Lessons Learned were discussed and documented by the core project team. On a quarterly basis, the Lessons Learned were reviewed and approved by the AOR.

A complete list of all of the approved Lessons Learned from this project can also be accessed and reviewed at: https://www.americasdatahub.org/adc-lessons-learned-pprl1-23-n03/.

During the completion of Task Area 4, the following Lessons Learned was documented and approved:

1. Flexibility was expected and required given the demonstration nature of this project. However, it was also important for the project to be successful in the timely completion of all project deliverables and milestones. By leveraging industry standard project management processes, methodologies, and tools, the project team was able to strike an appropriate balance of flexibility and timely project execution.

7. Conclusion and Recommendations

This project served as a critical proof-of-concept for establishing a data sharing agreement between two agencies and leveraging PPRL to facilitate secure, privacy-compliant interagency data sharing and analysis. With a focus on linking data from the NCHS and the NCSES, the work successfully demonstrated the viability of two agencies working together to deploy PPRL technologies to advance evidence-building across the federal government. In doing so, it laid foundational infrastructure and guidance for future efforts under the NSDS vision.

Demonstrating Secure and Effective Interagency Linkage

At its core, the project sought to answer whether it is possible to securely link data from two federal statistical agencies—without exchanging clear text personally identifiable information (PII)—and use the resulting linked data for analyses. The answer, resoundingly, is yes. Through a close partnership with HealthVerity and Mathematica, the team:

- **Established a Data Sharing Agreement (DSA)** between NCSES and NCHS—two agencies with no prior history of shared agreements—successfully navigating legal, confidentiality, and information security review processes.
- **Utilized a FedRAMP-authorized PPRL tool** that allowed de-identified linkage using encrypted tokens, ensuring strict adherence to CIPSEA compliance requirements.
- **Built and validated secure data workflows** within a trusted compute environment, enabling linked datasets to be analyzed for statistical purposes (with the data remaining secure and privacy compliant after being de-identified and linked and placed into the secure data access facility)
- **Provided a path forward for future similar studies**, such as capturing the stakeholders required, timelines needed, technical specifications that need to be addressed, and other items that are critical of using linked cross-agency data for future studies.
- **Delivered these outcomes on an accelerated timeline**, transforming an untested plan into a proven approach ready for future use. The experience gained has positioned NCSES to move even faster in the future, with this project's timeline serving as a performance baseline and Key Performance Indicator (KPI) against which to measure and improve similar efforts going forward.

Each objective of the original contract was met, including generating the technical and procedural insights needed to inform future NSDS linkages and ensuring that the resulting data products remain analytically robust and privacy-compliant.

Lessons Learned: A Roadmap for Future Linkages

Beyond simply demonstrating feasibility, this project captured a rich set of **Lessons Learned** that can guide future interagency data sharing and linkage initiatives. The detailed Lessons Learned are organized by Task Area and can be found at the end of each of the four Task Area sections of this report, The Lessons Learned have also been summarized below:

1. **Build Ample Time for Security and Governance**

Timeframe planning is critical. Processes like obtaining an ATO under FedRAMP standards and aligning calendars of critical leaders across agencies require extended lead times. Projects of this nature must

incorporate slack into their timelines to accommodate necessary security reviews, policy alignments, data governance approvals; and many times the re-work that needs to be accomplished when reconciling requirements that may differ between the two agencies that are working together.

2. **Engage the Right Stakeholders Early and Often**

Project success hinged on sustained engagement across multiple stakeholder groups: government project team, contractor project team, data owners, analytics teams, legal advisors, information security officials, privacy experts, technical experts, and other vendors (e.g., secure analytics environment admin). Decisions needed to be made jointly and often on short timelines, underscoring the importance of a collaborative structure that is established during the planning phases of the project, not while the project is in mid-flight. A predefined stakeholder map and sequencing plan (who is involved, when, and how) should be a standard component of future initiatives. Ultimately this project laid the foundation for how to engage with critical stakeholders for future projects.

3. **Understand Technical Requirements Upfront**

Successful PPRL implementation requires deep knowledge of:

- **Data environments** (cloud vs. on-premises, ownership, and access)
- **Data structure and quality**, especially completeness of key PII fields
- **Potential naming inconsistencies or formatting issues** for linkage fields that can impact token generation and matching
- **Linkage-specific issues**, such as handling of nicknames or alternate identifiers
- **Performing linkage quality assessments**, such as investigating reasons for match failure, prior to the destruction of de-identified data

4. **Plan Analytics with Data Constraints in Mind**

Working in secure compute environments introduces limitations on output sharing and analytic flexibility. The team learned early development of analytic plans and clear understanding of covariates, ideally through collaborative review of data dictionaries, streamlined modeling and interpretation of the linked file. Pre-linkage brainstorming about analytic goals and variables of interest ensures the resulting data structure supports meaningful analysis.

5. **Invest in Accessible Support Materials**

As NSDS evolves into a shared service model, it must serve users with varying levels of technical expertise. Documentation related to linkage, privacy protocols, and analytic considerations must be crafted for both technical and non-technical audiences. The team worked with NCSES to enhance HealthVerity's documentation to meet these needs and recommends that future efforts formalize the development of user-friendly technical assistance materials.

6. **Feedback regarding commercial vs. open-source vs. custom-developed PPRL solutions**

While the scope of this project did not include a head-to-head evaluation of PPRL technology types, the implementation offered valuable insights into benefits of using a commercial solution; insights that are especially relevant as the NSDS moves toward broader operationalization and scaling.

Although PPRL as a concept is no longer novel, its real-world application remains complex, particularly in environments that require secure, accurate, and scalable linkage across multiple data sources. As the NSDS evolves and continues to integrate disparate federal datasets, the lessons from this project suggest that commercial PPRL solutions can offer several distinct advantages:

- **Accuracy and Coverage**: HealthVerity's PPRL approach goes beyond tokenization as it leverages a robust, continuously maintained master patient list that spans a majority of the U.S. population. This ecosystem enhances matching accuracy across datasets that may vary in structure and completeness.
- **Dedicated Support and Responsiveness**: Unlike open source or custom-built tools, commercial providers typically offer dedicated technical support and customer service. This proved especially valuable in this project for troubleshooting, onboarding, and resolving software deployment and implementation challenges efficiently and quickly.
- **Reduced Staffing and Skill Dependencies**: Open source and custom PPRL implementations often require in-house expertise in languages such as R or Python, as well as sustained institutional knowledge. By contrast, the commercial solution mitigated these demands and reduced the dependency on specific technical personnel.
- **Speed and Scalability**: HealthVerity's solution demonstrated the ability to process large volumes of records quickly, with built-in mechanisms for quality assurance, secure data handling, and post-processing—all of which are critical for time-sensitive projects operating at scale.

That said, these benefits must be weighed against two important considerations: **cost** and **transparency**. Commercial solutions typically involve higher financial investment and may not always allow full visibility into the underlying algorithms or processes used for matching. Agencies evaluating future PPRL strategies under NSDS should carefully consider these trade-offs, balancing performance and support needs with budget constraints and the importance of interpretability in statistical work.

A Foundation for the National Secure Data Service (NSDS)

Ultimately, this project demonstrated that linking sensitive datasets across federal agencies is not only possible, but it also can be done in a way that is secure, reproducible, and scalable. The project generated a blueprint for cross-agency collaboration that can be replicated and streamlined.

Importantly, all data and metadata remain accessible only within a secure, governed environment and may be made available under the appropriate terms of an updated or new DSA. In this way, the project not only met its immediate goals but also ensured that its value will persist well beyond its conclusion.

As the federal government advances toward a more integrated, secure, and privacy-aware data infrastructure, this work marks a meaningful step forward. With its combination of technical execution, governance innovation, and deep documentation, the project stands as a model for how the NSDS can operate in practice—balancing innovation, privacy, and analytical power to inform policy in a complex data landscape.