National Institute of Diabetes and Digestive and Kidney Diseases

Central Repository

NIDDK-CR Resources for Research

Data Science and Meet the Expert Webinar Series

Building Real-World Data (RWD) Linkages with a Focus on Quality and Reusability

September 25, 2025

National Institute of Diabetes and Digestive and Kidney Diseases

Central Repository

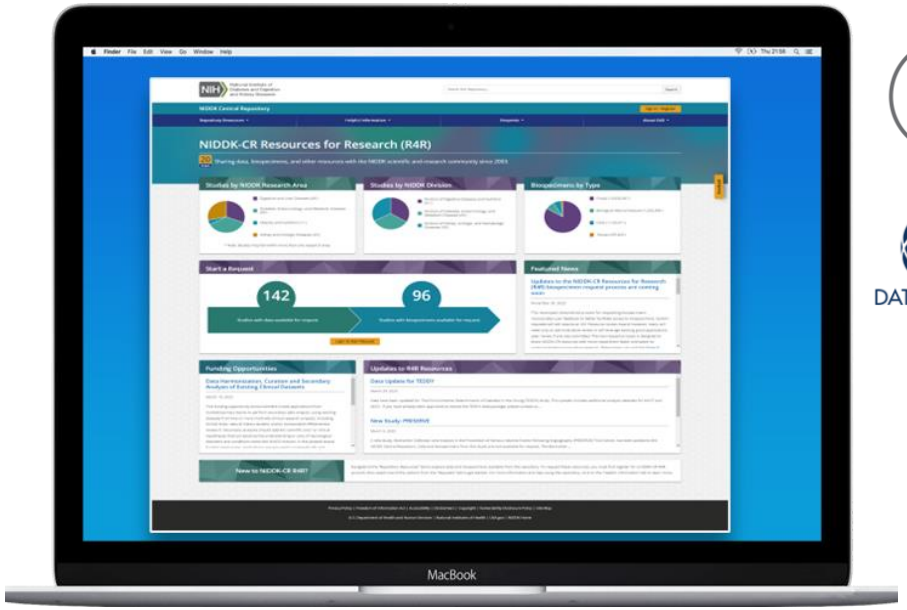
NIDDK Central Repository Overview

Mission

Established in 2003 to **facilitate sharing of data, biospecimens, and other resources** generated from studies supported by NIDDK and within NIDDK's mission by making these **resources available for request to the broader scientific and research community**.

- Supports receipt and distribution of data and biospecimens in a manner that is ethical, equitable, and efficient
- Enables investigators not involved with the original work to test new hypotheses without the need to collect new data or biospecimens
- Promotes FAIR (Findable, Accessible, Interoperable, and Reusable) and TRUST (Transparency, Responsibility, User focus, Sustainability, and Technology) principles

Recorded past tutorials, webinars, and other educational resources can be found on the NIDDK-CR website



CORE TRUST SEAL

WORLD DATA SYSTEM

Imaging Data Files
15.8 M

Clinical Datasets
>8,600
from 194 clinical studies

Biospecimens
>16 M

Registered Users
>7,000

Weekly Users
>5,000

Public Releases
>900

National Institute of Diabetes and Digestive and Kidney Diseases

Central Repository

Secondary Data Science and Meet the Expert Webinar Series

About the Series

- Aims to accelerate data science and AI-driven biomedical research by fostering collaboration between biomedical researchers and experts in the field
- Monthly webinar held on the **last Thursday of each month**

Upcoming Webinars

- Today – Building Real-World Data (RWD) Linkages with a Focus on Quality and Reusability
- October 30th – Data Science Project Pipelines, Workflows, and Tools

Learn more about the webinar series, register for future webinars, and access past webinars materials and recordings

National Institute of Diabetes and Digestive and Kidney Diseases

Central Repository

Webinar Agenda

1. Quality considerations/barriers to using RWD in research
2. Considerations and best practices when generating, stewarding, and using RWD for research
3. Best practices and lessons learned from data linkage implementations across federal health
   - NCI RWD Infrastructure
   - FDA Sentinel
   - National Secure Data Service (NSDS) Demonstration Project: Utilizing PPRL to Link Data from Two Federal Statistical Agencies (CDC and NCSES)
4. Governance framework for data linkage

National Institute of Diabetes and Digestive and Kidney Diseases

Central Repository

Meet the Experts

Jason Meyer is the Vice President of the Government team at HealthVerity, where he leads efforts to deliver innovative real-world data (RWD) and privacy-preserving record linkage (PPRL) solutions to federal health agencies. In this role, he oversees initiatives that expand the use of RWD, advance PPRL adoption, and build partnerships with systems integrators and research organizations to support safety surveillance, population health management, and other research priorities at CDC, NIH, FDA, NCSES, ARPA-H and other agencies. Prior to joining HealthVerity, he was an Associate Partner at IBM Watson Health (formerly Truven Health Analytics) and earlier spent six years at Booz Allen Hamilton supporting federal health clients. He holds a bachelor's degree in Political Science from Dartmouth College.

David Kwasny is a Principal at HealthVerity, where he focuses on privacy-preserving record linkage (PPRL) and data interoperability for government and public-health use cases. Previously he spent 10 years at IQVIA working with Real World Data for federal and life science clients. He is based in the Philadelphia area and is an alum of American University's Kogod School of Business.

National Institute of Diabetes and Digestive and Kidney Diseases

Central Repository

Meet the Experts

Lisa Mirel is the Statistical Advisor at the National Center for Science and Engineering Statistics (NCSES) within the U.S. National Science Foundation. In this role she serves as the senior technical statistical advisor to support Center priorities related to survey design, statistical standards, privacy and confidentiality, data quality, and demonstration projects to inform the development of a future National Secure Data Service. Prior to coming to NCSES, she served as the Chief of the Data Linkage Methodology and Analysis Branch at the National Center for Health Statistics (NCHS).

National Institute of Diabetes and Digestive and Kidney Diseases

Central Repository

Meet the Experts

Dr. Susan Tenney is a Senior Health Scientist at Booz Allen Hamilton. She holds a Ph.D. from Georgetown University Medical School and advanced certification in Change Management from Georgetown's McDonough School of Business, allowing her to integrate scientific expertise with organizational strategy. At HHS, she has led data-driven projects that strengthen research infrastructure, governance, and advanced analytics. Recently, she led two privacy preserving record linkage (PPRL) projects in support of the NICHD Office of Data Science and Sharing (ODSS) and is currently the Project Manager supporting the Foundation for NIH (FNIH) for the Cross-HHS Patient Health Data Platform where EHR data will be linked using PPRL.

NIDDK-CR Data Science and Meet the Expert Webinar Series:

Building Real-World Data (RWD) Linkages with a Focus on Quality and Reusability

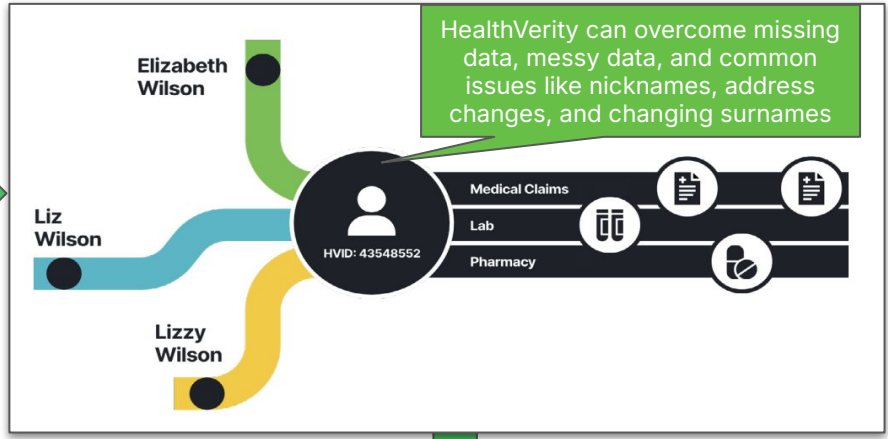
HealthVerity: Jason Meyer & David Kwasny
NSF NCSES: Lisa Mirel, MS
9/25/2025

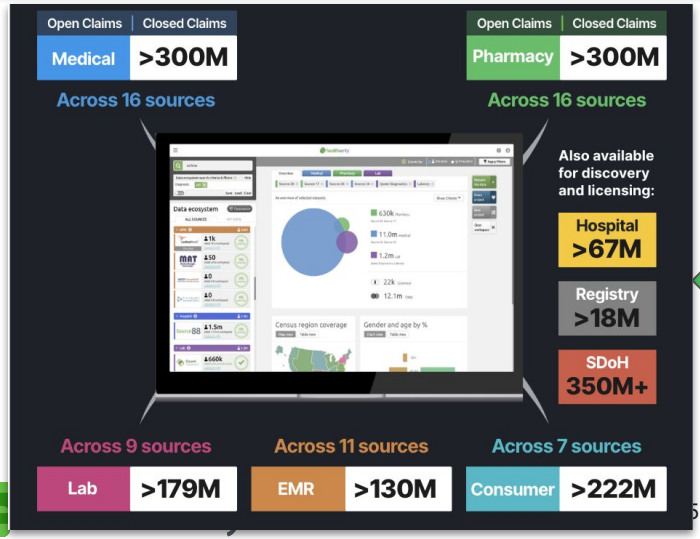
Synchronize the Science℠

HealthVerity: PPRL and RWD Marketplace

HealthVerity Tokenization and Data Linkage

HealthVerity Identity Manager is a tokenization/Privacy Preserving Record Linkage (PPRL) software that accurately links patient identities over time and across data sources by utilizing a continually-updated database of over 200 billion healthcare and consumer transactions, using probabilistic matching with machine learning and AI to ensure the highest accuracy when assigning an HealthVerity ID (HVID).





HealthVerity Tokenization and Data Linkage

Using the HVID, HV has created the largest RWD marketplace in the US, which includes all types of RWD including claims, EMR, Lab, Rx, and SDOH. The data is geographically representative of the U.S., and comes in standardized HIPAA-compliant data models. All data data is centralized - not on a federated network - so data can be contracted for and delivered in a matter of days, not months.

Federal Project Examples Using Linked RWD

Real World Data and Linkage Project Spotlights

- **National Cancer Institute:** Cancer Real World Data Infrastructure
- **FDA Sentinel:** Linked Claims and EMR
- **National Secure Data Service (NSDS) Demonstration Project:** Utilizing Privacy Preserving Record Linkage to Link Data from Two Federal Statistical Agencies (NCHS and NCSES)
- **Key Considerations:** Key Considerations When Procuring and Using Real-World Data

Federal Program Deep Dive - NCSES National Secure Data Service

National Cancer Institute: Cancer Real World Data Infrastructure Project

Multiple RWD programs were created in the intermediate term to assess the effect of COVID, but each came with various challenges

Real World Data research addressing COVID-19

• **Centers for Disease Control and Prevention (CDC)**	surveillance
• **NIH National COVID Cohort Collaborative (N3C)**	consortium of EMR data
• **Researching COVID to Enhance Recovery (RECOVER)**	consortium of EMR data
• **COVID-19 Global Rheumatology Alliance (GRA)**	consortium of EMR data
• **International Consortium for Characterization of COVID-19 by EHR (4CE)**	consortium of EMR data

Challenges	Unmet Needs
• Uniform structure (or not) among participating organizations • Permissions and Privacy • "Missingness" of data, particularly resulting from patient churn	Unanswered Public Health questions • Long-term impact of COVID-19 on vulnerable populations • Impact of COVID-19 on incidence of cancer, comorbid conditions • Health Care Resource utilization in the post-COVID era • Effectiveness (and risks) of vaccination

Recognizing these gaps, the NCI undertook a series of initiatives to develop a complementary infrastructure aimed at addressing the gaps

The National Cancer Institute (NCI) made efforts to create a platform that filled these gaps, starting with a convening of SMEs

NCI convened a group of subject matter experts (SMEs) to create a set of research questions/research areas of interest not yet able to be answered by other platforms, so that the data and infrastructure could be configured in the appropriate manner

SME Group Participants	
Government Agencies NCI (Lead) NIAID CDC	**Academic Partners** Brown University Johns Hopkins University Louisiana State University
Health Systems* Northwell Health	**Private Industry** Labs: Quest & Labcorp HealthVerity Aetion (Lead)

Research Areas of Interest	
1	Predictors of & Incidence of long COVID in immunocompromised groups
2	Predictors of & Incidence of new-onset chronic diseases (e.g., diabetes, kidney disease, cardiovascular disease) after COVID infection
3	Risk of new-onset cancer or autoimmune conditions post-COVID-19
4	Impact of specific therapies (cancer, rheumatologic conditions) on COVID-19 outcomes
5	Effectiveness and durability of protection of COVID-19 vaccines in immunocompromised populations
6	Incidence of COVID-19 vaccine side effects in immunocompromised populations
7	Recurrence of cancer among patients in (cancer) remission, post-COVID-19

NCI Use Case: COVID-19 Real World Data Infrastructure (CRWDi) with a focus on immunocompromised patients

Platform to be used by IRB-approved academic researchers

Data moved to a secure analytics platform

- SEER Cancer Registry
- Quest & Labcorp Labs
- HealthVerity Medical Claims
- HealthVerity Pharmacy Claims
- Northwell Health System Labs
- Mortality Data

De-identified, linked and harmonized using HealthVerity PPRL software

Stats/Facts

- Distinct patients in each data source:
  - SEER: 6.9M (2010-2023)
  - Commercial Claims: 5.2M (2018 - 2024)
  - Commercial Labs: 5M (2018 - 2024)
  - Northwell: 0.3M (2020 - 2023)
  - Mortality Data linked
- The total number of rows across all tables is n = 25,357,862,985 rows
- 2,000 observations for each of the 10M patients on average
- Databricks hosted environment
- HIPAA compliant

Lessons Learned and Results

Results

- Platform went live in December of 2024 with 4 cohorts created: (1) patients with cancer; (2) patients with rheumatic diseases receiving pharmacotherapy; (3) noncancer solid organ and hematopoietic stem cell transplant recipients; and (4) people from the general population including adults and pediatric patients.
- While real-world data cannot replace clinical trials, this gives the ability to leverage real-world datasets allows investigators and governmental agencies to focus on important subpopulations who are unlikely to be eligible for clinical trials.
- One Published paper*, and one other in progress
- Eight IRB-approved academic research teams onboarded
- *These results cover a 5 month time span as the platform was discontinued in April of 2025 due to the Trump Administration Government Efficiency Reviews*

Lessons Learned

- When linking and harmonizing a diverse collection of healthcare data, it is critical to thoroughly define the research aims/questions prior to procuring the data due to the need to trade fields while ensuring data is HIPAA-compliant.
- The linked data was complex - thorough training prior to use is critical for maximizing value of the data
- DUAs take time to establish
- Privacy and linkage - As noted above, this project required stringent HIPAA risk mitigation activities to ensure the linked data was de-identified. Working HIPAA compliance into the fabric of the project is critical to ensuring patient privacy and participation from all data owners.
- SEER data has a significant data lag, but the linkage of claims data to these patients gives researchers insight into these patients during the lag time - claims + SEER strategy helps overcome missingness/delay of data

*Coronavirus Disease 2019 (COVID-19) Real World Data Infrastructure: A Big-Data Resource for Study of the Impact of COVID-19 in Patient Populations With Immunocompromising Conditions; Crawford, et al

FDA Sentinel: Linked Claims and EMR data for the Sentinel Innovation Center

FDA Sentinel System & Innovation Center: RWD and PPRL

FDA Sentinel Background

FDA's Drug surveillance system

- Established 2008 (FDAAA)
- >360M unique patients, >700M person-years
- Distributed/Federated claims data network to monitor drug safety

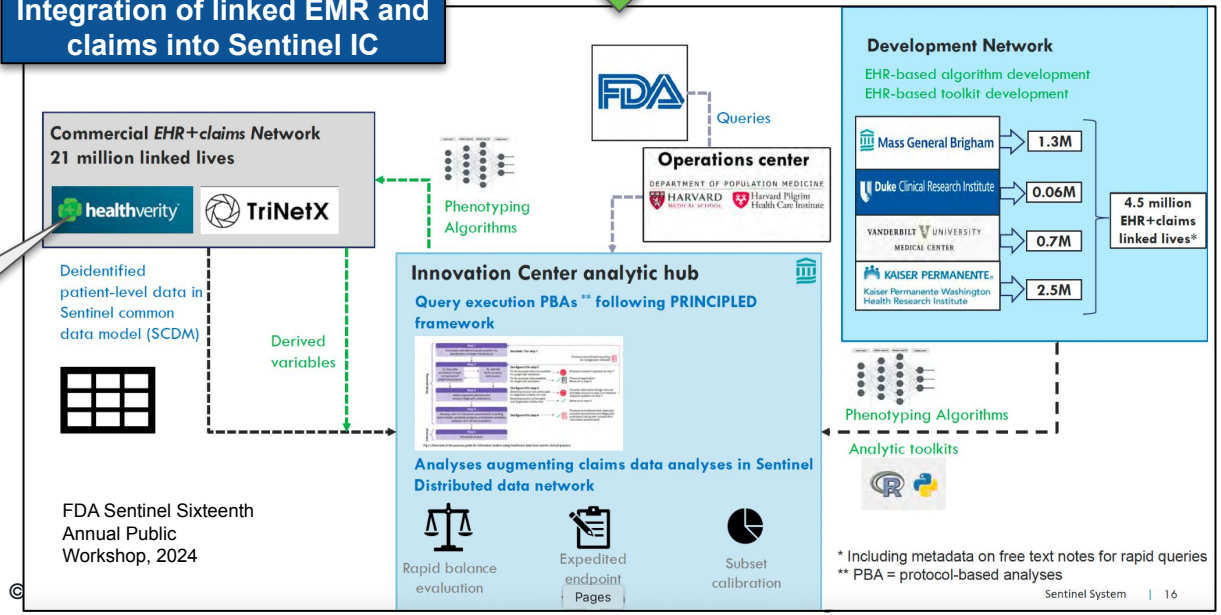
Results

- Hundreds of FDA-led safety analyses and studies since Sentinel's launch.
- Harmonization of multiple payer claims databases - dozens of data partners
- Establishment of extensive tools and data models to facilitate studies

Sentinel Innovation Center (IC) - Established 2019

In 2019 FDA established the Sentinel Innovation Center to be a test bed to identify, develop and evaluate innovative methods to study drug safety and analysis using linked Claims and EMR real-world data.The IC was created in response to FDA's Medical Data Enterprise Initiative to build a new system containing electronic health records from 10 million lives and develop methods for causal inference, NLP, and feature engineering

Integration of linked EMR and claims into Sentinel IC



Linked claims and EMR enabled by PPRL

healthverity

FDA Sentinel System & Innovation Center: RWD and PPRL

"The FDA Sentinel Real World Evidence Data Enterprise (RWE-DE)"

(Pharmacoepidemiology and Drug Safety, 2024; Desai et al.)

Historically, Sentinel relied mostly on insurance claims data, which are strong for longitudinal coverage but lack granular clinical detail (labs, vitals, indications, outcomes). EHRs provide that clinical richness, but usually capture only fragments of a patient's care ("information leakage"). **Linking EHRs with claims fills gaps — enabling more complete longitudinal evaluation of medication safety**

Two components of data providers for Study:

- Commercial Network (TriNetX + HealthVerity) → 21M lives.
- HealthVerity = mostly ambulatory EHRs + >150 payers' closed claims (2018–2019)
- TriNetX = hospital-centric EHRs (20 orgs) + >150 payers' claims (2010–2023).
- Development Network (MGB, Duke, Vanderbilt, Kaiser WA) → 4.5M lives.
- Rich access to both structured & unstructured EHRs, linked to Medicare, Medicaid, or integrated system claims

Findings

- All sites converted their data to the Sentinel Common Data Model (SCDM)
- Completeness varied by site: claims-based tables were nearly universal; EHR-based tables (labs, vitals) were less complete, reflecting tests done outside contributing systems.
- This means that while linked data are available, the depth and overlap of claims/EHR vary significantly by network partner.
- Primary analyses: conducted in RWE-DE when claims-only analyses are insufficient (e.g., acute pancreatitis — PPV improved from 61% with claims codes → 92% when lipase labs from EHR added).
- Claims ensure continuity across care settings; EHRs add richness for confounders, labs, vitals, and nuanced outcomes; together, they enable studies that were previously impossible in Sentinel.

Use of Linked EMR and Claims for Sentinel Study

- ***Study:*** Risk of Non-arteritic Anterior Ischemic Optic Neuropathy (NAION) with GLP-1 receptor agonist use in Type 2 Diabetes patients
- ***Data:*** *HealthVerity claims + EHR linkages (incl. Mass General Brigham), Jan 2018–June 2024*
- ***Method:*** *PRINCIPLED causal framework with enhanced outcome validation*
- ***Impact:*** *Demonstrates how linked EHR + claims improve FDA's ability to investigate emerging drug safety concerns*

https://sentinelinitiative.org/studies/drugs/individual-drug-analyses/risk-non-arteritic-anterior-ischemic-optic-neuropathy-naion

National Secure Data Service (NSDS) Demonstration Project: Utilizing Privacy Preserving Record Linkage to Link Data from Two Federal Statistical Agencies (NCHS and NCSES)

NSDS Demonstration Project: Utilizing Privacy Preserving Record Linkage to Link Data from Two Federal Statistical Agencies (NCHS and NCSES)

Project Objectives

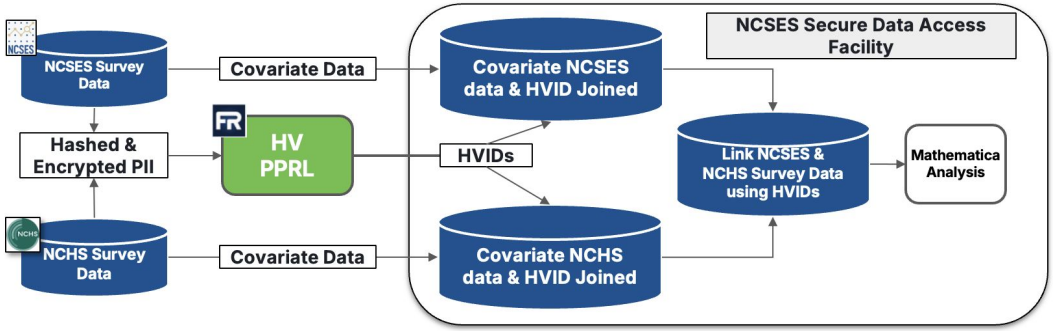
Running from August 2023 to August 2025, this project aimed to:

- Establish data sharing agreement between two federal agencies
- Deploy FedRAMP authorized PPRL technology to securely link de-identified records
- Enable analysis in a secure government analytics environment
- Test feasibility of cross-agency PPRL and gather lessons learned to inform the National Secure Data Service

Project Stakeholders

- **NSF NCSES** – Government Project Lead
- **CDC NCHS** – Government Project Tech Partner
- **HealthVerity** – Prime Contractor (Linkage)
- **Mathematica** – Subcontractor (Analysis)
- Awarded via America's Datahub Consortium, managed by **Advanced Technology Int'l**

Linkage Workflow



Key Results/Lessons Learned

FedRAMP Moderate ATO

Established a data sharing agreement (DSA) between NCSES and NCHS

820k Records Processed in 6 Minutes

High tokenization rate; low overlap between data sets

FINAL REPORT

Extensive Lessons Learned including capturing the DSA steps for faster execution in the future, and understanding data layouts to optimize linkage success

Key Considerations When Procuring and Linking Real-World Data

Area of Focus	Discussion
Data Governance/ DUAs	Data owners are increasingly sensitive to risks: (1) Unauthorized re-identification of patients; (2) Misuse of data outside the agreed DUA; (3) Privacy breaches due to downstream use (e.g., token profiling). High-trust governance is essential: establish strong, enforceable DUAs - which includes clearly articulating the use case before procurement
Security and Privacy	Tokenization ≠ automatic HIPAA compliance. Plan for HIPAA certification, or adopt HIPAA-compliant data models early. Be aware that safe harbor approaches may reduce data utility. Consider secure enclaves, audit trails of data use, and HIPAA training.
Tokenization and Linkage	Evaluate trade-offs across accuracy, speed, cost, resources, and data value when assessing commercial, open source, and “home-grown” linkage approaches.
Working with Complex and Messy/Missing Data	Linking and analyzing multiple datasets is technically challenging. Add in the fact that many datasets are missing fields, or are “messy”, and your team will need a core of analytic experts in order to overcome these issues. Success requires: (1) Skilled programmers or advanced analytics platforms; (2) Clearly defined use cases, field requirements, and (3) thorough vetting of data sources to form a clear understanding of data completeness/cleanliness up front to avoid costly rework or procurement of new or supplemental data.
Social Determinants of Health (SDOH)	Race/ethnicity data is inconsistently captured (OMB, hospital forms, NIH standards, private-sector variations), and has historically low completion rates in RWD sources. Adding SDOH fields (e.g., income, language, education) may disrupt HIPAA certification—plan for field substitutions if linking in SDOH data at the patient level.
Artificial Intelligence	Many data owners are cautious about AI use cases. Guardrails and clear restrictions on AI-related applications are often required before data is shared.
Procurement Timelines	Due to governance, privacy, linkage, and contracting complexities, data procurement takes time. Build these extended timelines into project plans and procurement cycles.

NSDS - Infrastructure to support data sharing and linkages across government

NSDS Infrastructure Build

The CHIPS and Science Act of 2022 authorized the National Science Foundation (NSF) to establish a National Secure Data Service (NSDS) Demonstration project

An NSDS aims to strengthen data informed decision making across government agencies by supporting:

- ✔ Privacy-enhancing technologies like PPRL
- ✔ Interoperable and secure multi-agency data sharing
- ✔ Coordinated governance and technical capacity-building
- ✔ Support innovation and fuel progress

NSDS's Vision and Mission

Interagency Council on Statistical Policy
Leaders of the United States Federal Statistical System

National Secure Data Service

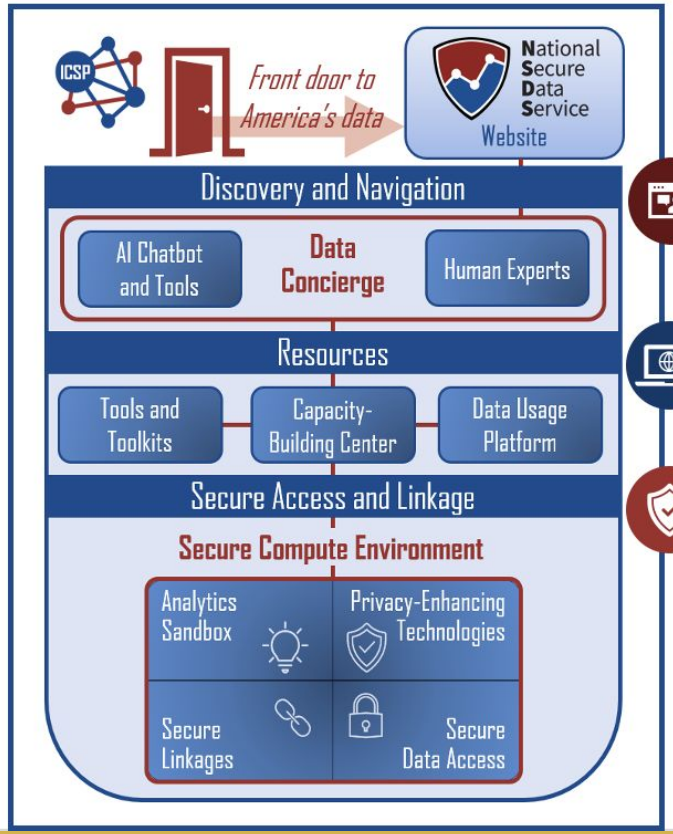
VISION

The NSDS is building a secure, scalable service – using innovative tools and powerful privacy protections – to further connect people with trusted data and solutions to make smarter decisions and solve real-world problems.

MISSION

Enabling accessible, secure, and seamless data solutions to spark innovation, strengthen communities, and build a better future —
Your Federal statistical system at work.

NSDS 1.0



THE NSDS WILL INCORPORATE AI IN MULTIPLE AREAS OF SHARED SERVICES:

Discovery and Navigation
- AI Chatbot

Resources
- AI Readiness Scorecard
- Data standardization tools

Secure Access and Linkages
- Privacy-Preserving Record Linkages
- Synthetic data generation

From Insights to Innovation

Interagency Council on Statistical Policy
Leaders of the United States Federal Statistical System

National Secure Data Service

SECURE COMPUTE ENVIRONMENT

Secure compute environment scan feasibility study included conversations with Federal statistical system partners to ensure security requirements could be met. Projects being conducted in testbed will inform production.

DATA CONCIERGE

Models for a Data Concierge feasibility study began by gathering needs and requirements from agencies and culminated in proposed models for a Data Concierge service.

AI CHATBOT

An AI Chatbot trained exclusively on data and resources pulled from the Federal statistical system began as an uncertain vision and has matured into a fully functioning proof of concept.

PRIVACY-PRESERVING RECORD LINKAGES

A scan of available technologies resulted in the development of a suite of privacy-preserving tools spanning open source and proprietary solutions.

DATA USAGE PLATFORM

Lessons learned from the NCSES User Analytics Platform and the Democratizing Data: A Search and Discovery Pilot program have refined NSDS's approach to its Data Usage Platform.

SYNTHETIC DATA

Pilot projects conducted in the secure compute environment testbed have led to a variety of tools for creating and working with synthetic data.

NSDS: Problem and Solution Approach

SECURE ACCESS AND LINKAGE PROBLEM

Linking data with attribute records can increase disclosure risk

SOLUTION

Privacy Preserving Record Linkage: AI/ML techniques enable the ability to streamline and innovate data linkages across sources using technology like PPRL

NSDS Demonstration: PPRL Projects

Objectives:

- Develop data sharing and governance prototype agreement
- Utilize PPRL tool when linking person level data from disparate sources
- Create analytic datasets that can be used to inform questions that could not be assessed with either source alone

Linking Data: Two Federal Statistical Agencies

Data sources

National Center for Health Statistics (NCHS): National Health Interview Survey (NHIS)

NCSES: Survey of Earned Doctorates (SED)

Output

- Developed data sharing agreement to link data from two federal statistical agencies
- Established processes for utilizing a commercial PPRL tool: HealthVerity
- Informed analytic initiatives

Conclusion

1. Having an analytic sandbox to deploy tools supports innovative methods to link data
2. AI/ML tools improve linked data quality while maintaining public trust in how data are handled
3. Resulting linked data can be used for analytic initiatives
4. This work supports national leadership in AI development through innovative models within a shared data ecosystem

National Institute of Diabetes and Digestive and Kidney Diseases

Central Repository

NIDDK-CR Resources for Research

Data Science and Meet the Expert Webinar Series

DATA LINKAGE GOVERNANCE

Booz Allen Hamilton: Susan Tenney, Ph.D.

September 25, 2025

National Institute of Diabetes and Digestive and Kidney Diseases

Central Repository

Linkage Governance - Topics

- Governance Definition in the Context of Health Data
- The Journey of a Dataset from a Governance Perspective
- Governance Requirements – To Link or Not to Link?
- Potential Sources for Data Linkage Governance
- Linkage Governance Compliance – Key Considerations
- Linkage Governance Resources

National Institute of Diabetes and Digestive and Kidney Diseases

Central Repository

Governance Definition in the Context of Health Data

Governance refers to the set of policies (rules), processes, and institutional roles, standards, and technologies that ensure health data is collected, stored, shared, and used in a manner that is secure, ethical, compliant, and trustworthy.

Goal
• **Balance data utility with responsibility**—making sure data can be used effectively to improve patient outcomes, while also protecting patient rights, confidentiality, and trust.

Benefits	
• **Protects individuals' rights** • Ensures quality and **trustworthiness** • Enables **beneficial uses** (for care, public health, research, innovation) • Maintain **accountability**	• Offers **protection** (privacy, security) but also **enablement** (legitimate access, interoperability, use for public health, research, quality improvement)

"Governance is a runway, not a speed bump."

- Zafar Chaudry, MD, Senior Vice President and Chief Digital, AI, and Information Officer, Seattle Children's Hospital

National Institute of Diabetes and Digestive and Kidney Diseases

Central Repository

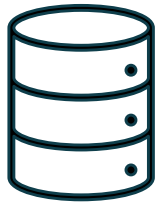
The Journey of a Dataset from a Governance Perspective

Patient provides data

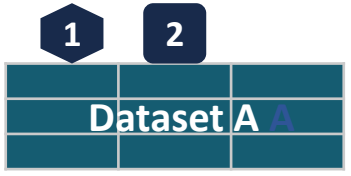
1. Informed Consent

Institution collects data

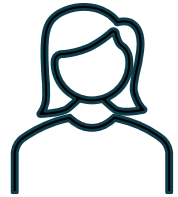
1. Informed Consent
2. Institutional (IRB/ Privacy Board, etc.)



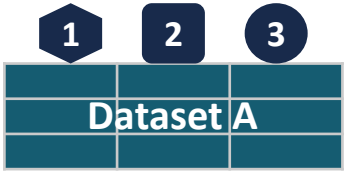
Repository links & shares data



1. Informed Consent
2. Institutional (IRB/ Privacy Board, etc.)
3. Repository (De-identified data, Enclave, etc.)



User uses data



1. Informed Consent
2. Institutional (IRB/ Privacy Board, etc.)
3. Repository (Data Use Agreement, Data Access Committee approval, etc.)

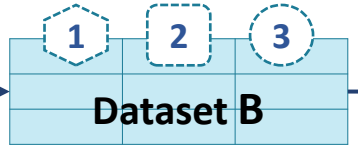
Governance Sources

National Institute of Diabetes and Digestive and Kidney Diseases

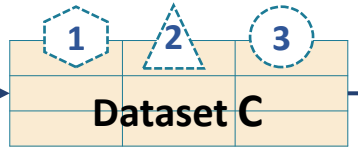
Central Repository

Governance Requirements – To Link or Not to Link?

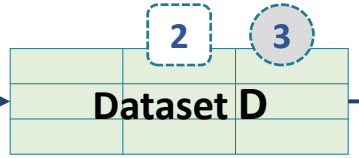
- ✓ Linkage authorized
- ✓ Must access in Enclave A
- ✓ No limitations to data use



- ✓ Linkage authorized
- ✓ Can access virtually (via isolated machine)
- ✓ No limitation to data use

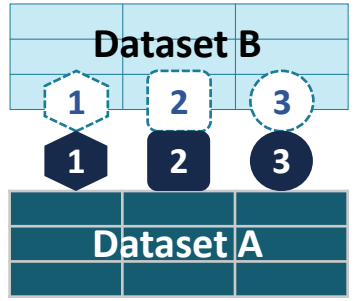


- ✓ Linkage authorized
- ✓ Must access in Enclave C
- ✓ No limitation to data use



- ✓ Linkage authorization missing
- ✓ Can access virtually (via isolated machine)
- ✓ Use limited to statistical purposes

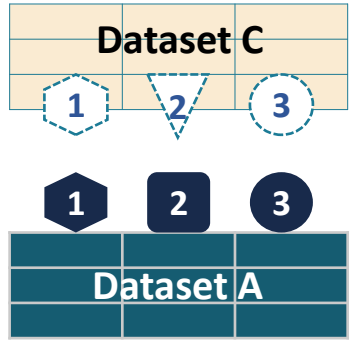
✓



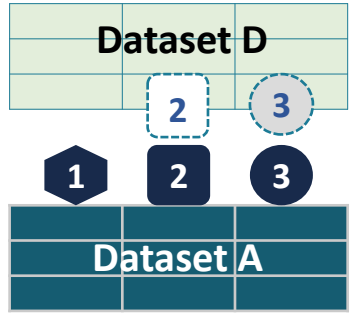
Dataset A and B can be linked:

- ✓ Both are authorized to link
- ✓ Dataset B can be accessed virtually through Enclave A
- ✓ Can be used for general purposes

X



?



Dataset A and D can be linked provided:

- ✓ Linkage authorization for Dataset D can be obtained
- ✓ Dataset D is accessed virtually through Enclave A
- ✓ Use is limited to statistical purposes

LEGEND

1 – Linkage authorization
2 – Access requirements
3 – Data use limitations

National Institute of Diabetes and Digestive and Kidney Diseases

Central Repository

Potential Sources for Data Linkage Governance (Examples)

Governance Sources	Data Collection	Data Linking	Data Sharing	Data Access	Data Use
Local/State/Federal/Tribal Regulations/Policies	✓	✓	✓		✓
Assent	✓	✓	✓	✓	✓
Informed Consent	✓	✓	✓	✓	✓
Waiver of Consent from IRB		✓	✓		✓
IRB/Privacy Board Determination	✓	✓	✓	✓	✓
Institutional Certification		✓	✓	✓	✓
Data Originator Agreement		✓	✓	✓	✓
Repository Policies/Agreements (DUA, DTA, etc.)		✓	✓	✓	✓

National Institute of Diabetes and Digestive and Kidney Diseases

Central Repository

Linkage Governance Compliance – Key Considerations

Governance Categories	Compliance Considerations
❑ Scope of data linkage	• Determine type of datasets or longitudinal datasets required for the linkage and study • Identify dataset characteristics: modalities, sources, de-identification levels, minimal data necessary

❑ Authorization for linkage	• Determine approval authorities (participants via consent, IRBs/Privacy board, data contributors, institutional bodies, repository governance, etc.)
	  ○ Best practice for consent/assent:
	  • *Prospective studies*: Incorporate explicit language in the consent for data linkage; allow flexibility for future addition of datasets
	  • *Existing studies* where linkage is not explicit in the consent: Re-consent or obtain IRB/equivalent Privacy Board determination for linking or a waiver of consent

❑ End-to-end governance	• Categorize existing governance for full data lifecycle – data collection, linkage, sharing, access, and use – from various dataset sources, including governmental policies, consent/assent, IRB determinations or waiver of consent, institutional policies and governance, data repository policies and governance • Determine whether the datasets can be linked based on existing governance; identify and address conflicts or gaps with data owners • Ascertain which dataset-level governance that exists prior to linkage (e.g., for collection and sharing) will be inherited by the linked dataset for sharing, accessing, and using
❑ Deductive disclosure review of the linked data	• Perform deductive disclosure review of the linked datasets to mitigate re-identification risk (perturb, suppress variables, collapse small cell sizes, etc.) • Consider privacy enhancing techniques (PET) such as differential privacy for added protections

Sources: NICHD ODSS Reports on Privacy Preserving Record Linkage (PPRL) for Pediatric COVID-19 Studies (2022) and Patient-Centered Outcomes Research Trust Fund Pediatric Record Linkage Governance Assessment (2023)

National Institute of Diabetes and Digestive and Kidney Diseases

Central Repository

Linkage Governance Resources

- NICHD ODSS Report on Privacy Preserving Record Linkage (PPRL) for Pediatric COVID-19 Studies (2022)
  - NICHD ODSS Record Linkage Implementation Checklist
- NICHD ODSS Report on Patient-Centered Outcomes Research Trust Fund Pediatric Record Linkage Governance Assessment (2023)
  - NICHD ODSS Linkage Determination Examples

National Institute of Diabetes and Digestive and Kidney Diseases

Central Repository

Contacts

- Lisa Mirel – lbmirel@nsf.gov
- David Kwasny – dkwasny@healthverity.com
- Jason Meyer – jmeyer@healthverity.com
- Susan Tenney – Tenney_Susan@bah.com

Upcoming Webinar: Advancing Collaborative Data Science with Texera and the NIDDK-CR Analytics Platform

- **Date:** October 30th from 2-3:30pm ET
- **Agenda:** This session is geared towards researchers interested in learning about how data science pipelines and workflow tools can enhance reproducibility, collaboration, and efficiency in research.
  - Introduction to data science pipelines, workflows, and challenges in research
  - Demonstration of TEXERA for workflow-driven analytics
  - Overview and demo of the NIDDK-CR Data Challenge Management platform
  - Demonstration of the NIDDK-CR Analytics Workbench
  - Summary of tools to support end-to-end research workflows
- **Scan the QR code register**

Thank You!