# Patient-Centered Outcomes Research Trust Fund (PCORTF) Pediatric Record Linkage Governance Assessment

# Contents

# Executive Summary

The *Eunice Kennedy Shriver* National Institute of Child Health and Human Development (NICHD) is leading a project to assess and analyze the data governance requirements for individual-level linkage of high-priority U.S. Department of Health and Human Services (HHS) datasets, with funding from the HHS Office of Secretary Patient-Centered Outcomes Research Trust Fund (OS-PCORTF) and support from Essex Management and Booz Allen Hamilton.

Individual-level linkages between datasets from different biomedical studies and HHS administrative and survey datasets provide opportunities to maximize the value of independent datasets by enabling researchers to deduplicate participants across datasets, introduce new variables into analysis plans, reduce costly redundancies in data generation, perform longitudinal analysis, and ask new scientific questions of the enriched dataset. For datasets to be appropriately linked, researchers and data stewards must understand the consent, policy, regulatory, and/or other legal frameworks that apply to each of the original datasets and how the resulting linked dataset inherits rules and controls from these frameworks. They must also understand if and how new limitations arise that impact the sharing and use of the resulting linked dataset, for example implementing new rules and controls to mitigate increased risk of participant identifiability. The collective set of these rules and controls is referred to as data governance, and it defines and enforces appropriate collection, sharing, access, linking, and use of the data, across the data lifecycle. The standardization of data governance information about individual datasets will help researchers and data stewards determine whether multiple datasets can be linked, and, if so, what data governance applies to the linked dataset.

The overall goal of this project is to develop and test a generalizable, scalable, and machine-readable data governance metadata schema that will facilitate decision-making for patient-centered outcomes research dataset linkages and the subsequent research use of linked datasets. This report documents the project team's assessment of data governance information for 11 HHS and other federally funded datasets identified for three theoretical pediatric COVID-19 research use cases and will serve as a foundation for the development the metadata schema.

Based on this dataset governance assessment, the report provides considerations for the development and implementation of a data governance metadata schema, including:

- Publicly sharing the data governance information specified by the schema in a predictable and easy-to-find location will facilitate the ability to create linked datasets

- Publicly shared data governance information, and the associated schema, should:

    - Explicitly describe whether linkage is permissible for a given dataset and, if so, include general guidance for what types of linkages are allowed or prohibited, and what rules and controls the linked data would inherit from the individual dataset

    - Incorporate the provenance of data governance origins including authorizations for data collection, linking, sharing, access, and use as well as applicable laws, regulations, and policies

    - Capture the roles and responsibilities of the multiple stakeholders involved in implementing data governance across the data lifecycle

    - Incorporate information regarding decisions made for previous and new linkages involving a given dataset to communicate appropriate linkage of the data and to inform future linkage involving the same dataset; this information may streamline decision making when linkage governance is not explicitly specified by any dataset governance source.

- The schema should describe data governance in a standard way to facilitate human interpretation and machine-readability, which in turn promotes adherence

- A concerted effort is required to encourage adoption of the schema across federal and other health agencies that generate datasets that could be linked and used by researchers

This work will ultimately promote more thoughtful and appropriate record linkage efforts, build community trust, and yield more discoveries from patient centered outcomes research.

# Introduction and Project Goal

The *Eunice Kennedy Shriver* National Institute of Child Health and Human Development (NICHD) is leading an effort to assess the usage of privacy preserving record linkage (PPRL) for pediatric patient-centered outcomes research. The NICHD Office of Data Science and Sharing (ODSS) has undertaken a project to assess and analyze the data governance and record linkage requirements for high-priority U.S. Department of Health and Human Services (HHS) datasets, with funding from the HHS Office of the Secretary Patient-Centered Outcomes Research Trust Fund (OS-PCORTF) and support from Essex Management and Booz Allen Hamilton.

Individual-level linkages between datasets from different biomedical studies and HHS administrative and survey datasets provide opportunities to maximize the value of independent datasets by enabling researchers to deduplicate participants across datasets, introduce new variables into analysis plans, and reduce costly redundancies, for example in the generation of molecular profiling data. Record linkage across the biomedical and health data ecosystem can also facilitate longitudinal data analysis as well as promote the generation of new research questions driven by the volume and variety of individual-level data. For datasets to be appropriately linked, researchers and data stewards must understand the consent, policy, regulatory, and/or other legal frameworks that apply to each of the original datasets and how the resulting linked dataset inherits these rules and controls from these frameworks. Importantly, they must further understand if and how new limitations arise that impact the sharing and use of the resulting linked dataset, for example implementing new rules and controls to mitigate increased risk of participant identifiability in the linked datasets.

The overall goal of this project is to develop and test a generalizable, scalable, and machine-readable data governance metadata schema that will facilitate decision-making for patient-centered outcomes research dataset linkages and the subsequent research use of linked datasets.

This project is subsequent to a previous initiative, undertaken in 2021, to address the specific need to link pediatric COVID-19 data. This prior project, led by the NICHD ODSS with funding from the National Institutes of Health Office of Data Science Strategy and support from Booz Allen Hamilton, comprehensively assessed 13 existing record linkage implementations and developed technical and governance considerations for appropriately linking data for new efforts. As described in the predecessor project's final report entitled, "Privacy Preserving Record Linkage (PPRL) for Pediatric Covid-19 Studies" (1), the team encountered significant challenges in finding, interpreting, and harmonizing governance information among datasets generated by studies that were part of the

trans-NIH Collaboration to Assess Risk and Identify Long-term Outcomes (CARING) for Children with COVID initiative. The importance of findable, interpretable, and harmonized data governance metadata to enable record linkage became apparent.

The present report, which describes the outcome of the assessment of data governance pertaining to dataset linkage and associated findings, represents the first step to achieving the overall goal of developing and testing a generalizable, scalable, and machine-readable data governance metadata schema. The schema will be informed by a comprehensive assessment of governance information and provenance for a diverse set of HHS and other federally funded patient centered outcomes research-relevant datasets, documented herein, as well as a landscape analysis of existing, relevant standards, terminologies, and ontologies, documented separately.

The intended audience for this report is the NICHD ODSS, NIH ODSS, and HHS agency staff who are currently implementing or plan to implement record linkage.

# Introduction to Data Governance and the Motivation for a Common Record Linkage Metadata Schema

## What is Data Governance?

Data governance is the collective set of rules and controls that define and enforce appropriate collection, sharing, access to, linking, and use of data. Each piece of research data, including de novo clinical research data as well as administrative and public health data used in research, brings with it a complex set of limitations and requirements that dictate how that data is managed across the data lifecycle. Key points in the typical research data lifecycle, including collection, sharing, access, use, and destruction, may contribute regulatory, policy, and/or technical constraints that follow that piece of data into subsequent stages of the data lifecycle. With the emergence of large and disparate research data repositories, federal mandates for data sharing and reuse, and the growing need for multidimensional analysis and the application of data science methods like machine learning, researchers must be aware of and consider these constraints when assembling datasets to power their investigations. Ideally, the data governance associated with each piece of data would be stored with the dataset, standardized, and readily discoverable.

Across the research data lifecycle, instances of data governance origins include:

- Individual participant authorization, e.g., informed consent for adults and assent in the case of children
- Waiver of consent from an Institutional Review Board (IRB)
- Other Determinations of Institutional Review Boards (IRB) or other Privacy Boards
- Data originator or data submission agreements
- Data Use Agreements, Data Sharing Agreements, or Data Use Licenses
- Data repository policies, program or data type specific policies, or other policies
- Contractual or other legal obligations
- Applicable statutes or regulations that may be assigned at the local, state, and/or federal level

- U.S. Tribal and international requirements
- Rules implemented to mitigate risk for a specific situation

Data provenance refers to the retention of information that describes where data originated and where it moves. As a matter of good practice, data provenance is captured in metadata elements and is retained over time to ensure appropriate attribution and use of data.

Data governance information, like data provenance, should be captured in metadata elements that are retained throughout the lifecycle of a dataset. Data governance information also necessitates its own provenance tracking. The provenance of data governance information is critical to ensure appropriate linking, sharing, and use of data.

## What is record linkage and why is it important to patient centered outcome research?

Record linkage, or data linkage, is the process of bringing together data about the same individual or entity from multiple sources to create a new, enriched dataset. Linking records across disparate datasets harnesses the power and maximizes the value of individual research efforts and enables investigators to address innovative questions that require the variety of datasets that may not be generated by a single research project. A multi-modal approach to characterizing a disease or condition – for instance analyzing clinical, imaging, and molecular profiling data collectively – may accelerate the discovery of disease mechanisms and pathways, ultimately most effectively bringing diagnostic strategies and treatment protocols from the laboratory into the clinic.

Individual-level data linkages across multiple data sources provide new opportunities to address research questions. Individual-level dataset linkages are particularly important in the cases of rare disease, such as childhood cancer or multisystem inflammatory syndrome in children (MIS-C) caused by COVID-19, where there are fewer cases repeatedly represented across disparate studies and linkage is needed to avoid working with inflated sample sizes and to co-analyze multiple data modalities from different data sources.

However, record linkage brings data use conditions and/or constraints from all linked data sources to the resulting dataset. Data governance established at each step in the data lifecycle cumulatively determines whether a dataset can be linked with another dataset, and how those linked data can then be used. Figure 1 illustrates the cumulative inheritance of data governance, including participant consent at data collection and data repository policy, and the impact of cumulative data governance on the investigator's ability to address research questions by linking disparate data sources.

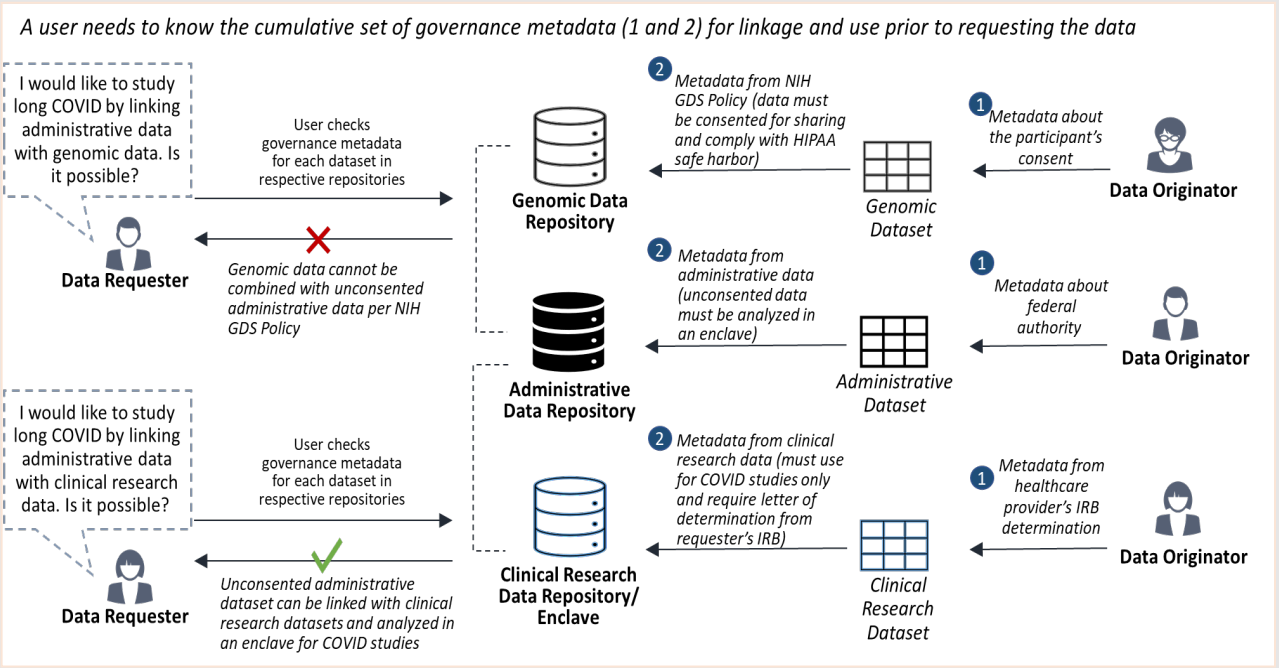


*Figure 1: Illustration of linkage determination based on the cumulative set of governance metadata available for individual datasets. Standardized data governance information about individual datasets helps researchers determine whether two datasets can be linked, and, if so, what rules and controls apply to the linked dataset.*

# The relevance and importance of a common record linkage metadata schema

Although presented in this report in the context of pediatric COVID-19 studies, the benefits and challenges of record linkage are ubiquitous and part of a broader set of data analysis needs. The mission of the HHS is to "enhance the well-being of all Americans, by providing for effective health and human services and by fostering sound, sustained advances in the sciences underlying medicine, public health, and social services" (2). The ability to capture, link, and analyze data is critical to the HHS mission. Likewise, patient-centered outcomes research, which focuses on generating high-quality evidence on the impact of treatments, services, and other health care interventions on patients, relies on data infrastructure that provides high-quality, analysis-ready data. Organizations within HHS, including Office of the Secretary's Patient Centered Outcome Research Trust Fund (OS-PCORTF) and NIH, have developed strategic plans to begin to leverage the opportunities and address the wide range of challenges that impact the implementation of a robust data ecosystem that facilitates their missions.

## OS-PCORTF Strategic Plan

In September 2022, the OS-PCORTF released its strategic plan for building data capacity for patient-centered outcomes research through coordinated, systematic efforts across federal agencies (3). Data capacity, in PCOR context, refers to the availability and

sustainability of data and analytic resources to address national health priorities. The OS-PCORTF strategic plan addresses a broad range of data sources, including clinical, clinical trial, social services, and administrative and claims data, and notes that issues of availability, quality, accessibility, and interoperability are significant hurdles to PCOR research. Varied, multi-modal data sources, data linkage, data analysis, and equitable access are the cornerstones of the PCOR data infrastructure.

The OS-PCORTF articulates four, interrelated goals and desired outcomes.

- Goal 1: Data Capacity for National Health Priorities
    - Outcome 1: Data, tools, and services to improve patient-centered outcomes research relevant to HHS priorities
- Goal 2: Data Standards and Linkages for Longitudinal Research
    - Outcome 2: Accessible, timely, interoperable, linkable, and longitudinal data
- Goal 3: Technology Solutions to Advance Research
    - Outcome 3: Robust real-world data across platforms and systems used to generate real-world evidence and expand data usage that informs patient, clinical, and policy decision making
- Goal 4: Person-Centeredness, Inclusion and Equity
    - Outcome 4: Accurate, relevant, and representative evidence is accessible to individuals; communities; and state, federal and tribal programs when making health decisions

The second goal of the OS-PCORTF strategic plan describes data standards and linkage and includes activities to assess the impact of policies related to privacy, security, and consent on PCOR efforts and to build consensus-based linkage methodology. The aim of this project, to develop and test a generalizable metadata schema that facilitates decision-making for PCOR dataset linkage and the subsequent use of linked datasets, aligns with Goal 2 of this plan. The project's goal to streamline decision-making for record linkage should move the HHS community towards secure and appropriate data linkage methodologies and the responsible usage of linked datasets for PCOR research.

## NIH Strategic Plan for Data Science

In June 2018, the NIH released its first Strategic Plan for Data Science (4) to begin to address the challenges of storing, managing, standardizing, and sharing data generated by NIH-funded research. In the report, it is noted that individual scientists or relatively

small collaborative teams generate most of the biomedical data comprising the NIH data ecosystem. The data tend to be distributed, not optimally integrated (stored in siloes), and generated in a variety of formats. The datasets also tend to lack structured metadata describing appropriate data use, reuse, and requirements/constraints for data sharing. Record linkages that may inspire new research ideas and offer the potential for biomedical innovation, such as the use of machine learning, must be thoughtfully and appropriately implemented in this complex ecosystem.

The NIH data science strategy articulates clear, integrated goals and objectives.

- Goal 1: Support a highly efficient and effective biomedical research data infrastructure
- Goal 2: Promote modernization of the data-resource ecosystem
- Goal 3: Support the development and dissemination of advanced data management, analytics, and visualization tools
- Goal 4: Enhance workforce development for biomedical data science
- Goal 5: Enact appropriate policies to promotes stewardship and sustainability

Collectively, these strategic goals convey the NIH's transformative vision for hardware, software, all stages of the data lifecycle, and the creation of a data science-capable research community. Goal 5 includes the importance of generalizable, consistent, and persistent governance metadata to promote stewardship, sustainability, and appropriate use of research data.

Importantly, across all goals, the NIH data science strategy makes a commitment to ensure that all biomedical research data adhere to FAIR principles (Findable, Accessible, Interoperable and Reusable), and that all processes and tools generate FAIR data (5). The FAIR principles include four key concepts:

- To be findable, data must be assigned a unique, persistent identifier and described with rich metadata
- To be accessible, data (and associated metadata) must be readily retrievable by open, free protocol allowing for authentication and authorization, where necessary
- To be interoperable, data (and associated metadata) must be represented by broadly shared and standardized vocabulary
- To be reusable, data are described by rich metadata that includes data characteristics, clear usage license, and detailed provenance

Data governance models vary greatly within NIH's federated data ecosystem. Reusability, as defined above, includes the important concepts of data usability and requires that metadata describing data governance (including inherited data governance) and provenance is associated with each piece of research data. Further, the metadata are mandated to meet domain-relevant community standards, which steers the community in the direction of common data usability metadata, including a common data governance metadata schema that could inform record linkage, thus aligning this project with NIH's strategic plan for data science.

# Goal, Objectives, and Activities for Record Linkage Governance Assessment

## Goal and Objectives

The goal of this assessment is to analyze governance information for datasets relevant to PCOR pediatric COVID-19 research to inform the development of a generalizable metadata schema that facilitates decision-making for PCOR dataset linkage and use of linked datasets. Figure 2 shows the goal and objectives for this governance assessment, and the sequential activities designed to meet the objectives.

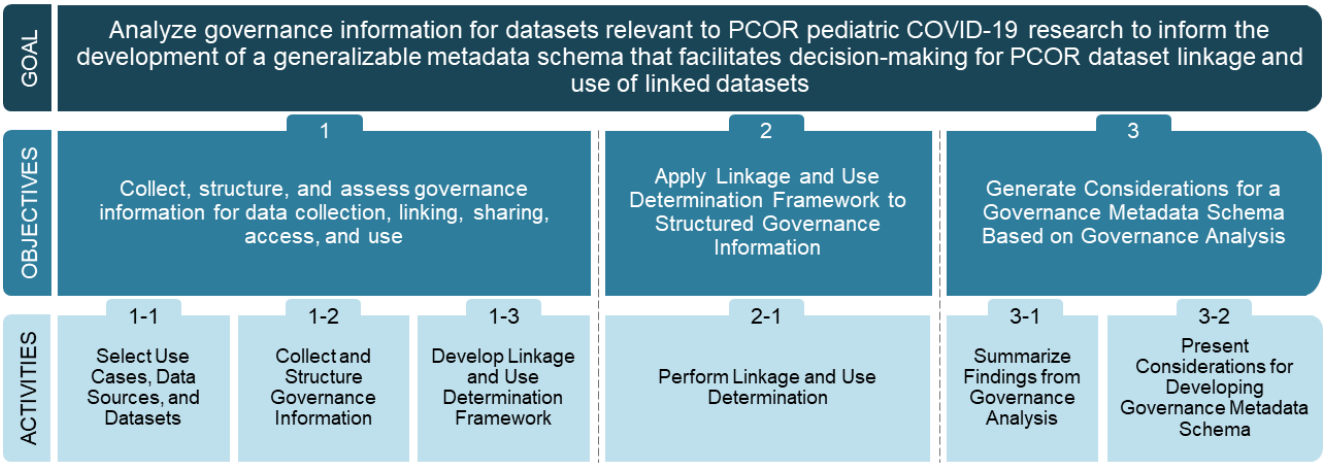


*Figure 2: PCORTF Pediatric Record Linkage Governance Assessment: Overview of the Assessment Goal, Objectives, and Corresponding Activities*

### Objective 1: Collect, structure, and assess governance information for data collection, linkage, sharing, access, and use

#### *1-1: Select Use Cases, Data Sources, and Datasets*

The project team used a two-phased methodology to prepare governance information for the assessment. In the first phase, the team identified applicable scientific use cases and relevant data sources and datasets with potential to provide useful data for record linkage. In the second phase, comprehensive data governance information was collected, organized, interpreted, and analyzed to develop considerations for creating a data governance metadata schema.

## Use Case Selection

The project team chartered a Project Governance Team, comprised of senior Federal data science leadership (review Appendix A for members), to provide overall scientific and subject matter expert guidance to the project team. The Project Governance Team contributed three important and timely scientific questions that served as driving use cases for this project, each providing the opportunity for the evaluation of real-world dataset linkage decisions based on existing data governance information, listed in the next subsection.

The use case requirements included:

- Relevance to pediatric COVID-19 and related conditions, and impact of COVID-19 on all facets of the health, development, and well-being of children
- Alignment with PCORTF goals[1], specifically building data capacity for national priorities and data standards and linkages for longitudinal research
- Diversity of data types, federal agencies, and applicable governance constraints to ensure that the outcomes and outputs from this governance assessment have broad utility across HHS agencies

Accordingly, each of the use cases focused on the impact of COVID-19 (either pandemic or infection) on children's health and included aspects such as mental health, foster care, cancer, and education. The use cases drove the assessment of a broad set of data types (for instance clinical, survey, incidence/survival, and claims) from a variety of HHS agencies, including the NIH, the Centers for Disease Control and Prevention (CDC), Substance Abuse and Mental Health Services Administration (SAMHSA), the Administration for Children and Families (ACF), and Centers for Medicare and Medicaid Services (CMS), as well as the Patient-Centered Outcome Research Institute (PCORI) non-profit institute funded through PCORTF.

## Final Use Cases Developed with the Project Governance Team

- Use Case 1: Effects of the COVID-19 pandemic on mental health of children. Are related outcomes more severe for children in foster care?
- Use Case 2: What is the impact of COVID-19 infection on pediatric cancer survivors? Or what is the impact of COVID-19 infection on future pediatric cancer outcomes?

[1] OS-PCORTF Strategic Plan for 2020-2029: https://aspe.hhs.gov/reports/os-pcortf-strategic-plan-2020-2029

- Use Case 3: Does SARS-CoV-2 vaccination result in reduced asthma-related school absences at 3/6/12+ months post-vaccination?

## Data Source and Dataset Selection

The Project Governance Team identified for each use case potential data sources that likely maintained datasets relevant to the scientific questions asked in the use cases. The data sources included large governmental initiatives, for instance the National Health and Nutrition Examination Survey (NHANES) and the National Survey on Drug Use and Health (NSDUH), with potential to contribute data from expansive populations that include children.

The project team researched potentially relevant datasets available from the selected data sources by gathering key information on each dataset to determine its usefulness for this assessment. These criteria included:

- Accessibility of the dataset to researchers
- Public availability of essential data types such as mental health, foster care, cancer survivorship, and school attendance variables
- Level of data captured (individual vs. aggregate)
- Capture and availability of Personally Identifiable Information (PII), data elements that could facilitate record linkage across datasets

Based on this preliminary evaluation, the project team identified the datasets that were relevant to the use case research questions therefore representing realistic researcher needs and associated governance requirements. The following tables show the list of data sources and final datasets selected for the governance assessment.

**Table 1a: Use Case 1 - Effects of the COVID-19 pandemic on mental health of children. Are related outcomes more severe for children in foster care?**

| # | Data Source | Data Source Description | Dataset Selected |
|---|---|---|---|
| 1 | Centers for Disease Control and Prevention (CDC)/National Health and Nutrition Examination Survey (NHANES) | The NHANES interview, publicly available and shared at the individual level, includes demographic, socioeconomic, dietary, and health-related questions. The examination component consists of medical, dental, and physiological measurements, as well as laboratory tests administered by highly trained medical personnel. | Nine-item depression screening instrument administered to participants aged 12 to 17 years; includes NHANES 2017-March 2020 pre-pandemic data |

| # | Data Source | Data Source Description | Dataset Selected |
|---|---|---|---|
| | | In addition to the publicly available data, limited access data, which includes data collected on drug use, sexually transmitted diseases, and alcohol use, are available through NCHS's Research Data Center (RDC) for the years 1999-2020.<br><br>Source: https://www.cdc.gov/nchs/nhanes/index.htm | |
| 2 | Substance Abuse and Mental Health Services Administration (SAMHSA)/ National Survey on Drug Use and Health (NSDUH) | Nationwide study that provides up-to-date information on tobacco, alcohol, and drug use, mental health and other health-related issues in the United States on persons aged 12 and older.<br><br>Demographic data include gender, race, age, ethnicity, educational level, employment status, income level, veteran status, household composition, and population density, personal and family income, health care access and coverage, illegal activities and arrest records, problems resulting from the use of drugs, and perceptions of risks.<br><br>Dataset includes age at first use and lifetime, annual, and past-month use of alcohol, marijuana, cocaine (including crack), hallucinogens, heroin, inhalants, tobacco, pain relievers, tranquilizers, stimulants, and sedatives, as well as substance use treatment history and perceived need for treatment.<br><br>Source: https://www.datafiles.samhsa.gov/dataset/national-survey-drug-use-and-health-2020-nsduh-2020-ds0001 | Population health survey (2017 – 2020) that includes data captured on mental health issues and the use of mental health services*<br><br>* SAMHSA discourages comparison of the 2020 survey to other years due to changes in data collection during 2020. This may limit the usefulness of NSDUH for answering this hypothetical research question; however, the project team has chosen to include NSDUH in this analysis because the governance for linking and using the data is unaffected by technical features of the data, and linkages with NSDUH remains viable for other research questions. |
| 3 | National Institute on Drug Abuse (NIDA)/ Monitoring the Future (MTF): A Continuing Study of American Youth, 2017 – 2021 | The Monitoring the Future (MTF) project, also widely known for some years as the National High School Senior Survey, is a repeated series of surveys in which the same segments of the population (8th, 10th, and 12th graders; college students; and young adults) are presented with the same set of questions over a period of years to track how answers change over time.<br>The survey includes questions about drug use, demographic characteristics, attitudes toward religion, parental influences, changing roles of women, educational aspirations, self-esteem, | DS0 Study-level files:<br>DS1 Core Data<br>DS2 Form 1 Data<br>DS3 Form 2 Data<br>DS4 Form 3 Data<br>DS5 Form 4 Data<br>DS6 Form 5 Data<br>DS7 Form 6 Data |

| # | Data Source | Data Source Description | Dataset Selected |
|---|---|---|---|
| | | exposure to sex and drug education, violence, and crime (both in and out of school).<br><br>Source: https://www.icpsr.umich.edu/web/NAHDAP/series/35# | |
| 4 | Administration for Children and Families (ACF)/ The Adoption and Foster Care Analysis and Reporting System (AFCARS) | The Department of Health and Human Services (HHS), Administration for Children and Families (ACF), Children's Bureau (CB) is responsible for the implementation and management of the Adoption and Foster Care Analysis and Reporting System (AFCARS). State and Tribal title IV-E agencies are required to report AFCARS case-level information on all children in foster care and children who have been adopted with title IV-E agency involvement (per §479 of the Social Security Act).<br><br>Data set includes general information (e.g., Agency, reporting period, record number), child demographic Information (e.g., date of birth, sex, race, ethnicity, disabilities, adoption status/age), removal/placement indicators, circumstances of removal, current placement setting, case plan goals, caretaker information, parental rights, foster family home outcome data, source of assistance/support, adoption elements, and information about parents (birth and adoptive).<br><br>Source: https://www.acf.hhs.gov/cb/research-data-technology/statistics-research/afcars | Foster Care Files (2017 – 2020):<br>Data files contain 37 elements that provide information on children covered by the protections of Title IV-B/E of the Social Security Act (Section 427) |

**Table 1b: Use Case 2 - What is the impact of COVID-19 infection on pediatric cancer survivors? Or what is the impact of COVID-19 infection on future pediatric cancer outcomes?**

| # | Data Source | Data Source Description | Dataset Selected |
|---|---|---|---|
| 1 | National Cancer Institute (NCI)/ National Childhood Cancer Registry (NCCR) | Registry that contains pediatric data from participating NCI Surveillance, Epidemiology, and End Results (SEER) and CDC National Program of Cancer Registries (NPCR) registries. The NCCR was developed under the NCI Childhood Cancer Data Initiative (CCDI) to leverage the nation's existing, primarily adult, cancer registries to identify and follow childhood cancer cases in the United States. | CiNA datasets, including individual-level CiNA research dataset, 1995 – 2019, and the CiNA public use dataset |

| # | Data Source | Data Source Description | Dataset Selected |
|---|---|---|---|
| | | NCCR contributes to the CCDI data ecosystem by serving as a linked infrastructure of central cancer registry data that will integrate various other childhood cancer data – from hospitals, research centers, health care administrations, and other sources – to enhance access to and utilization of childhood cancer and survivorship data.<br><br>The NCCR Explorer includes data from Cancer in North America (CiNA), which includes North American Association of Central Cancer Registries (NAACCR) for the years 1995 - 2018, and SEER. Contributing state registries, representing 66% of all U.S. children, adolescents, and young adults of age 0 – 39 years based on 2018 populations, include: California, Connecticut, Florida, Georgia, Hawaii, Idaho, Illinois, Iowa, Kentucky, Louisianna, Massachusetts, New Jersey, New Mexico, New York, Ohio, Pennsylvania, Washington – Seattle/ Puget Sound, Tennessee, Texas, Utah, and Wisconsin.<br><br>The NCCR uses the Virtual Pooled Registry Cancer Linkage System to link multiple cancer registries and generate an accurate count of childhood cancer cases by combining information that appears in more than one registry.<br>Source: https://nccrexplorer.ccdi.cancer.gov/ | |
| 2 | Center for Medicaid and Medicare Services (CMS)/ Transformed Medicaid Statistical Information System (T-MSIS) | T-MSIS collects Medicaid and Children's Health Insurance Program (CHIP) data from U.S. states, territories, and the District of Columbia into the largest national resource of beneficiary information.<br><br>CHIP, jointly funded by states and the federal government, provides health coverage to eligible children through both Medicaid and separate CHIP programs. CHIP is administered by states, according to federal requirements. Data include state-level Medicaid data, eligibility and enrollment data, Medical/Health Services data (e.g., vaccinations, contraceptive care, telehealth, and dental services), and financial | T-MSIS Analytical Files |

| # | Data Source | Data Source Description | Dataset Selected |
|---|---|---|---|
| | | data (i.e., state-by-state total expenditures by program).<br><br>Source: https://www.medicaid.gov/medicaid/data-systems/macbis/transformed-medicaid-statistical-information-system-t-msis/index.html | |
| 3 | CDC/ COVID-19 Case Surveillance Data | Data collected by the CDC related to COVID-19 case, death, and testing data (national and regional), case and death demographic data, vaccination distribution and coverage data, vaccine effectiveness and breakthrough surveillance, health equity data, pediatric data (MIS-c, demographics, regional data, hospital admissions, ER visits), social impact data, pregnancy data, variant and genomic surveillance data, antibody seroprevalance, post-COVID conditions, and COVID-19 therapeutics data.<br>Source: https://data.cdc.gov/Case-Surveillance/COVID-19-Case-Surveillance-Restricted-Access-Detai/mbd7-r32t | COVID-19 Case Surveillance Restricted Access Detailed Data: Restricted access, patient-level dataset with clinical and symptoms data, demographics, and state and county of residence (33 data elements) |

**Table 1c: Use Case 3 - Does SARS-CoV-2 vaccination result in reduced asthma-related school absences at 3/6/12+ months post-vaccination?**

| # | Data Source | Data Source Description | Dataset Selected |
|---|---|---|---|
| 1 | The National Center for Advancing Translational Sciences (NCATS)/ National COVID Cohort Collaborative (N3C) | NCATS N3C Data Enclave is a centralized, secure, national clinical data resource with powerful analytics capabilities that the research community can use to study COVID-19, including potential risk factors and long-term health consequences.<br><br>N3C systematically and regularly collects data derived from the electronic health records of people who were tested for COVID-19 or who had related symptoms, as well as data from individuals infected with pathogens that can support comparative studies, such as SARS 1, MERS and H1N1. The data set includes such information as demographics, symptoms, lab test results, procedures, medications, medical conditions, and physical measurements.<br>Source: https://covid.cd2h.org | N3C "limited" dataset |

| # | Data Source | Data Source Description | Dataset Selected |
|---|---|---|---|
| 2 | The National Patient-Centered Clinical Network (PCORNet)/ PEDSnet | PCORnet, a national resource where health data, research expertise, and patient insights are available to deliver fast, trustworthy answers that advance health outcomes, is an integrated partnership of large Clinical Research Networks and a Coordinating Center that represents a diverse set of patients and institutions, ranging from academic medical centers to local community health clinics.<br><br>PEDSnet is a Clinical Research Network in PCORnet that collects EHR data from the 11 participating PEDSnet institutions, including demographic data, outpatient encounters, inpatient admissions, ER encounters, anthropometrics, vital signs, providers, diagnoses, treatments, visit payer, lab test results, and medications.<br><br>Source: https://pcornet.org; https://pedsnet.org/ | PEDSnet: Dataset contains more than 200 data elements, including clinical and demographics data, for over 3 million pediatric patients from 12 states |
| 3 | NIH/Rapid Acceleration of Diagnostics (RADx) | The NIH Rapid Acceleration of Diagnostics Data Hub (RADx Data Hub) supports researchers in accessing curated and de-identified COVID-19 data, allowing them to find, aggregate and perform data analyses in a cloud-enabled platform.<br><br>The RADx Data Hub supports efforts to understand COVID-19 and factors associated with disparities in COVID-19 morbidity and mortality in underserved and vulnerable populations. The RADx Data Hub seeks to accelerate scientific solutions and innovations in the development, commercialization, and implementation of technologies for COVID-19 testing by providing de-identified COVID-10-related data, algorithms, and other capabilities generated by various digital health solutions and technologies.<br><br>RADx Underserved Populations (RADx-UP), one of four RADx Coordinating and Data Collection Centers, aims to understand the factors associated with disparities in COVID-19 morbidity and mortality to reduce health disparities for underserved and vulnerable | RADx Underserved Populations (RADx-UP) Return to School Initiative |

| # | Data Source | Data Source Description | Dataset Selected |
|---|---|---|---|
| | | populations who are disproportionately affected by the highest infection rates of COVID-19 and/or are most at risk for complications or poor outcomes from the pandemic.<br><br>Source: https://radx-hub.nih.gov/ | |
| 4 | Environmental Protection Agency (EPA)/ Air Quality Systems (AQS) | The AQS provides access to air quality data collected at outdoor monitors across the U.S., Puerto Rico, and the U.S. Virgin Islands. The ambient air quality data is collected at the site, county, state, and core-based statistical areas (CBSAs) defined by the Census Bureau.<br><br>The AQS Database includes pollutant data (including CO, Pb, NO2, Ozone, PM10, PM2.5, and SO2), Remote Sensing Data (including satellite data and NASA Moderate Resolution Imaging Spectroradiometer), Community Multi-scale Air Quality (CMAQ), and National Environmental Satellite, Data, and Information Service (NESDIS) biomass burning data.<br><br>Source: https://www.epa.gov/outdoor-air-quality-data | AQS Database |

### *1-2: Collect and Structure Governance Information*

The project team leveraged the findings and considerations from the predecessor PPRL report to inform the methodology for collecting and structuring the governance information for this assessment.

### Findings from the "Privacy Preserving Record Linkage (PPRL) for Pediatric COVID-19 Studies" report for Governance Information

"Privacy Preserving Record Linkage (PPRL) for Pediatric COVID-19 Studies," the final report of the predecessor project, describes the assessment of thirteen record linkage implementations across NIH, other federal agencies, and non-governmental organizations (1). The report then proposes a set of governance and technical considerations that could inform the design of any PPRL implementation in a federated data ecosystem. The report provides insight into the types of governance information associated with data across the research landscape, and the report findings form the basis for the information collection process, Governance Information Framework, and Linkage and Use Determination Framework used in this assessment.

The record linkage implementations assessed in the report demonstrated a wide range of mechanisms for authorizing dataset linkage. These authorization sources include:

- Explicit consent from participants, and assent in the case of children
- Waiver of consent from the data originator's institutional review board (IRB)
- Approval from an IRB or equivalent Privacy Board
- Authorization from the data originators and their institutions, for example via Data Submission Agreements
- Federal law, e.g., Titles 13 and 26 of the United States Code, Office of Management and Budget (OMB) Guidance M-14-06, Section 308(b) of the Public Health Service Act (42 United States Code 242m(d), the Confidential Information Protection and Statistical Efficiency Act (Title V of Public Law 107-347), and the Privacy Act of 1974

The record linkage implementations demonstrated other sources of dataset governance that imposed rules and controls either inherited by individual datasets contributing to the linkage or to added to mitigate potential risk introduced by linkage. These controls included:

- IRB/equivalent Privacy Board determination or exemption for linking or accessing linked data
- An agreement from the data originator's or data submitter's institution attesting that the data can be linked and shared
- Using a standard definition of de-identified for all datasets contributed to the linkage
- Instituting risk mitigation controls, such as committee review or data transformations, prior to including a given dataset in a linkage implementation or after linking but prior to sharing
- Establishing governance bodies to review and approve requests for data linkage and access to linked data
- Providing access to the linked data through an enclave and/or controlled access

- Requiring a data use agreement that specifies compliance with data use limitations imposed by consent or other authorizations
- Providing individual datasets alongside linkage information so researchers can navigate dataset-level data use limitations, if needed
- Linking the data for a specific study but destroying the linked data after the study is completed
- Prohibiting re-identification as a term of access

Finally, the scope of authorized data linkage for a given record linkage implementation could encompass data part of a specific study or multiple datasets in a data repository or specific network.

**Governance Information Collection**

Drawing on the findings from the predecessor report, the project team identified the broad categories of governance-relevant information that inform determinations about linkages between datasets, including:

- Authorizations for each step of the data lifecycle (data collection, linkage, sharing, access, and use)
    - Assent, consent, IRB/equivalent Privacy Board determination, local/state/federal regulation/s, data originator agreement, repository agreements/policies, and Other (if any)
- Specific local/tribal/state/federal regulations and/or policies for each step of the data lifecycle (data collection, linkage, sharing, access, and use) for the chosen dataset
    - Such as the NIH Genomic Data Sharing policy, Family Education Rights and Privacy Act (FERPA), contractual obligations, etc.
- Restrictions/controls on de-identification status of the dataset for sharing through the repository
- Stipulations on how re-identification risk is managed for the dataset prior to sharing
- Data access and use requirements
- PII elements collected and the organization/party that holds the PII

- Information about prior data linkages that included the dataset

The project team researched and captured the governance-relevant information mentioned above for each of the datasets in each of the use cases.

This work included researching whether PII was collected as part of each dataset, and if so, what specific PII elements are included in the dataset, and which party holds the PII. The team also researched which common data model, if any, was used, e.g., Observational Medical Outcomes Partnership (OMOP). The project team also collected information on any prior record linkages involving each dataset (e.g., PII elements used in linkage, party that performed entity resolution, party that performed linkage) and information about the other datasets involved in the linkage (e.g., data type, data source) to clarify how the identified governance information had informed past record linkage implementation.

For each of the 11 datasets, the project team first attempted to collect all data governance and related information from publicly available websites and documents. The project team performed Google searches using "<data source/dataset>" and "<agency name>" to locate publicly available information. The team then reviewed web pages with overviews of the data source/dataset and those specifically related to data governance, such as pages with the titles "Data Governance", "Policies", "Linkage", etc. Data source websites offered high level governance information and typically linked to repository or program websites that provided more specific information describing requirements for the use of the dataset. Where available, the project team reviewed documents such as informed consent (and/or assent) forms, data dictionaries, case report forms, data originator agreements, data use agreements, project brochures, and user guides, all of which provide some insight on the conditions for data use, linkage, and sharing. The project team captured the raw governance language from these public resources and documents, citing the source for each instance.

Once public resources were exhausted, the project team reviewed the collected information and identified gaps in governance information as well as any raw language with unclear meaning or ambiguous implications for governance. The team then developed structured interview questions for each dataset to fill any gaps in governance information and to validate all information collected from public sources. The project team collected responses to the questions from relevant points of contact (e.g., data stewards, principal investigators, data source leaders) through structured interviews, offline communication, or both. The structured interviews and offline communication also clarified ambiguous information, provided granular information about authorizations, data governance, PII captured, and prior dataset linkages, and validated information collected from public sources. The project team added any new information

collected during the interview or from other communications to internal documentation. The team also updated any previously collected information with additional details obtained from POCs and archived any previously collected publicly available information deemed inaccurate by POCs. Table 2 lists all resources used to gather governance information for each of the 11 datasets.

**Table 2: Data governance information resources used to gather information for each of the 11 datasets**

| Data Source | Data source/ Dataset website | Documentation provided by dataset Stakeholders | Consent Documents | Brochure | Stakeholder Interviews via Email | Stakeholder Interviews in Person |
|---|---|---|---|---|---|---|
| NHANES | X | X | X | X | X | X |
| NSDUH | X | X | X | | | |
| MTF | X | X | X | | X | X |
| AFCARS | X | X | | | X | X |
| NCCR | X | X | | | X | X |
| T-MSIS | X | X | | | X | X |
| CDC COVID | X | | | | | |
| N3C | X | X | | | X | |
| PEDSnet | X | X | X | | X | |
| RADx | X | X | X | | X | X |
| EPA Air Data | X | | | | | |

## Governance Information Structuring

After collecting the governance information for the 11 datasets, the project team structured the data governance information and associated provenance for analysis using a Governance Information Framework (Table 3). The team developed the framework to efficiently organize the governance information for each of the data lifecycle stages based on the fundamental questions at hand: what rules apply to the data and from where those rules originate. The development of the framework also enabled the project team to begin to generate, through the real-world examples, an initial structure for a standardized set of metadata to effectively describe the governance information for each dataset. The project team transformed the information collected above into this framework to generate the governance information data sheets included in Appendix D.

For each stage of the data lifecycle (collection, linkage, sharing, access, data use), the framework organizes the governance elements into two sections:

- Governance origins: Authorization(s) and applicable regulations and/or policies from which the governance originates
- Governance variables: Governance for data linkage, sharing, access and use based on governance origin (i.e., authorization(s) and applicable regulations and/or policies or repository policies), specifically:
    - Whether data can be linked
    - With what other data can it be linked or can it not be linked (scope of linkage)
    - Whether data can be shared
    - How data can be shared (de-identification status, disclosure review)
    - How data can be accessed (access type, data use agreement, data access committee/group approval, IRB letter of determination, etc.)
    - How data can be used (data use limitations)
    - Other

Importantly, the project team acknowledges that the act of data sharing, which we generally define as making data accessible to the broader data use community, often encompasses multiple steps and parties. For example, sharing sometimes involves a data originator providing data to a central resource (i.e., repository) which then distributes the data to secondary users through standard access procedures. For the purposes of this framework, sharing authorizations and applicable regulations/policies refer primarily to the step where a data originator or collector provides or submits data to a central resource or repository, whereas data access authorizations and applicable regulations/policies refer to secondary users' access to the data through the central resource.

The governance variables, "Whether data can be shared" and "How data can be shared", more broadly encompass data originator and/or repository processes for making the data accessible to secondary users.

As part of structuring information in the framework, the project team recorded the source of information for each governance origin and variable to ensure tracking of provenance. The team then interpreted the raw information gathered from public sources and stakeholder interviews to standardize the assessment of origins and variables in succinct and consistent language in preparation for the subsequent exercise

to determine if datasets for a given use case could be linked and how resulting linked datasets could be used (dataset linkage and use determination).

For governance origins (authorization and applicable regulations/policies), the project team determined whether each governance authorization origin (e.g., assent, consent, IRB approval) for the dataset specifically authorizes or impacts each data lifecycle activity (data collection, linkage, sharing, access, and use). The team used templated language to ensure uniformity, i.e., "[Authorization origin] authorizes [data life cycle stage] or [Applicable regulation origin] applies to [data life cycle stage]." When multiple origins existed for the same variable, each authorization or applicable regulation/policy was numbered for clarity.

For governance for data linkage, sharing, access, and use based on authorization or applicable policies and regulations, the project team determined whether each authorization or applicable policy/regulation (i.e., origin of the governance) specifies whether and how the dataset can be linked, shared, accessed, and/or used. Again, the team used templated language to ensure standardization of governance variables, i.e., "[Authorization origin] specifies that the data will/can be [linked/shared/accessed via [location] or used for [specifically described] purposes]."

When governance information was not available despite the project team's diligent research, "Information not available/found" was specifically noted. The team indicated that a particular authorization or regulation/policy type did not apply to the dataset for the given lifecycle stage with "Not Applicable".

The standardized interpretation of each variable was structured as shown in Table 3 for analysis. The restructured data enabled the project team to efficiently make dataset linkage and use determinations.

**Table 3: The final simplified Governance Information Framework used in the project team's analysis of each dataset, representing standardized information about governance origins and governance variables across the data lifecycle. The full governance information data sheets included three columns for each governance origin and variable: Raw Language, Interpretation, and Source. Only Interpretation was included in the final, simplified version. The full framework also included other variables about PII elements and prior data linkages (where applicable).**

| Governance Information Framework | | Dataset | | | | |
|---|---|---|---|---|---|---|
| | | Data Collection | Data Linking | Data Sharing | Data Access | Data Use |
| **1** | **Governance Origins: Authorizations and Applicable Regulations/Policies** | | | | | |
| **1.1** | **Authorizations** | | | | | |
| 1.1.1 | Assent | | | | | |

| Governance Information Framework | | Dataset | | | | |
|---|---|---|---|---|---|---|
| | | Data Collection | Data Linking | Data Sharing | Data Access | Data Use |
| **1** | **Governance Origins: Authorizations and Applicable Regulations/Policies** | | | | | |
| **1.1** | **Authorizations** | | | | | |
| 1.1.2 | Consent | | | | | |
| 1.1.3 | IRB/equivalent Privacy Board determination | | | | | |
| 1.1.4 | Local/tribal/state/federal/ international/contractual regulation(s)/policies | | | | | |
| 1.1.5 | Institutional Certification | | | | | |
| 1.1.6 | Data originator agreement | | | | | |
| 1.1.7 | Repository agreements/policies | | | | | |
| 1.1.8 | Other (specify) | | | | | |
| **1.2** | **Applicable Regulations/Policies** | | | | | |
| 1.2.1 | Local regulations/policies | | | | | |
| 1.2.2 | Tribal regulations/policies | | | | | |
| 1.2.3 | State regulations/policies | | | | | |
| 1.2.4 | Federal regulations/policies | | | | | |
| 1.2.5 | International regulations/policies | | | | | |
| 1.2.6 | Contractual obligations | | | | | |
| 1.2.7 | Repository agreements/policies | | | | | |
| **2** | **Data Linking/Sharing/Access/Use Governance Based on Governance Origin** | | | | | |
| 2.1 | Whether the data can be linked | | | | | |
| 2.2 | With what other data it can or cannot be linked (scope of linkage) | | | | | |
| 2.3 | Whether data can be shared | | | | | |
| 2.4 | How data can be shared (de-identification status, disclosure review) | | | | | |
| 2.5 | How data can be accessed (access type, data use agreement, data access committee/group approval, IRB LOD, etc.) | | | | | |
| 2.6 | How data can be used (including data use limitations) | | | | | |
| 2.7 | Other (specify) | | | | | |

### *1-3: Develop Linkage and Use Determination Framework*

## Linkage and Use Determination Framework

Using the lessons learned from the predecessor report and the information gathered for the current analysis, the project team developed a Linkage and Use Determination Framework for analyzing whether and how the datasets for each use case can be linked and used based on the governance information of the individual datasets. This framework was designed to capture the key governance elements that allow or prevent linkage between datasets and that specify the scope or means of allowable linkage and use.

The Linkage and Use Determination Framework comprises the following four questions:

- Can the datasets be linked?
- What limitations does the linked dataset inherit?
- What controls does the linked dataset require?
- What authorization gaps exist?

In the framework, limitations are defined as requirements for how a dataset must or must not be used and/or linked. Controls are defined as technical or administrative processes that enforce alignment to limitations. Authorization gaps are defined as missing authorizations for particular data lifecycle activities that may prevent the proposed linkage or use of the linked data if not addressed. Together, limitations, controls, and gaps specify whether and how datasets can be linked and how linked datasets can be used (Table 4).

**Table 4: The four questions that form the Linkage and Use Determination Framework applied to each use case.**

| Use Case Dataset Linkages | Can the datasets be linked? | What limitations does the linked dataset inherit? | What controls does the linked dataset require? | What authorization gaps exist? |
|---|---|---|---|---|
| Dataset 1 and 2 linkage | Yes, provided limitations 1, 2, etc. are respected and controls 1, 2, etc. are implemented | 1. ....<br>2. ....<br>3. ....<br>Etc. | 1. ....<br>2. ....<br>3. ....<br>Etc. | 1. ....<br>2. ....<br>Etc. |
| Dataset 1 and 3 linkage | | | | |
| Dataset 1 and 4 linkage | | | | |
| Dataset 2 and 3 linkage | | | | |
| Dataset 2 and 4 linkage | | | | |
| Dataset 3 and 4 linkage | | | | |

## Objective 2: Apply Linkage and Use Determination Framework to Structured Governance Information

### *2-1: Perform Linkage and Use Determination*

#### Application of Linkage and Use Determination Framework

The project team performed a systematic pairwise analysis of the structured governance metadata for each of the datasets within each of the use cases, using the dataset Linkage and Use Determination Framework. Figure 3 demonstrates an application of the framework for a pairwise linkage and use determination for two datasets from Use Case 1, NHANES and MTF. This is an illustrative example, and only a subset of governance information is shown. For each of the two datasets, the project team first summarized the limitations related to data linkage and use. In this case, NHANES must only be linked to vital statistics, health, nutrition, and other related records. The project team also summarized the controls required for each dataset, and in this case, NHANES and MTF each must be accessed through a particular enclave. Both datasets also require researchers to sign Data Use Agreements, obtain approval from data source staff, and provide a letter of determination from their IRB. The project team then summarized the authorization gaps for linking and using each dataset, and in this case, MTF does not have an explicit authorization for data linkage, but there is also no explicit prohibition on linkage.

The project team then generated a combined list of gaps that should be addressed for the two datasets to be linked, the controls inherited from each dataset by the new linked dataset, and the data use limitations that must be observed for the linked dataset. This combined set of governance allowed the project team to answer the question, “Can these datasets be linked?” For the present example, the project team determined that NHANES and MTF can be linked provided the MTF authorization gap is addressed, and the inherited limitations and controls shown in Figure 3 that govern the appropriate use of the linked dataset are respected. In this example, NHANES and MTF have conflicting controls: users are required to access the datasets in two different enclaves. Therefore, to respect the controls inherited from each of the datasets, a user wishing to link NHANES and MTF would have to reach an agreement with both data sources about the enclave used for data access.

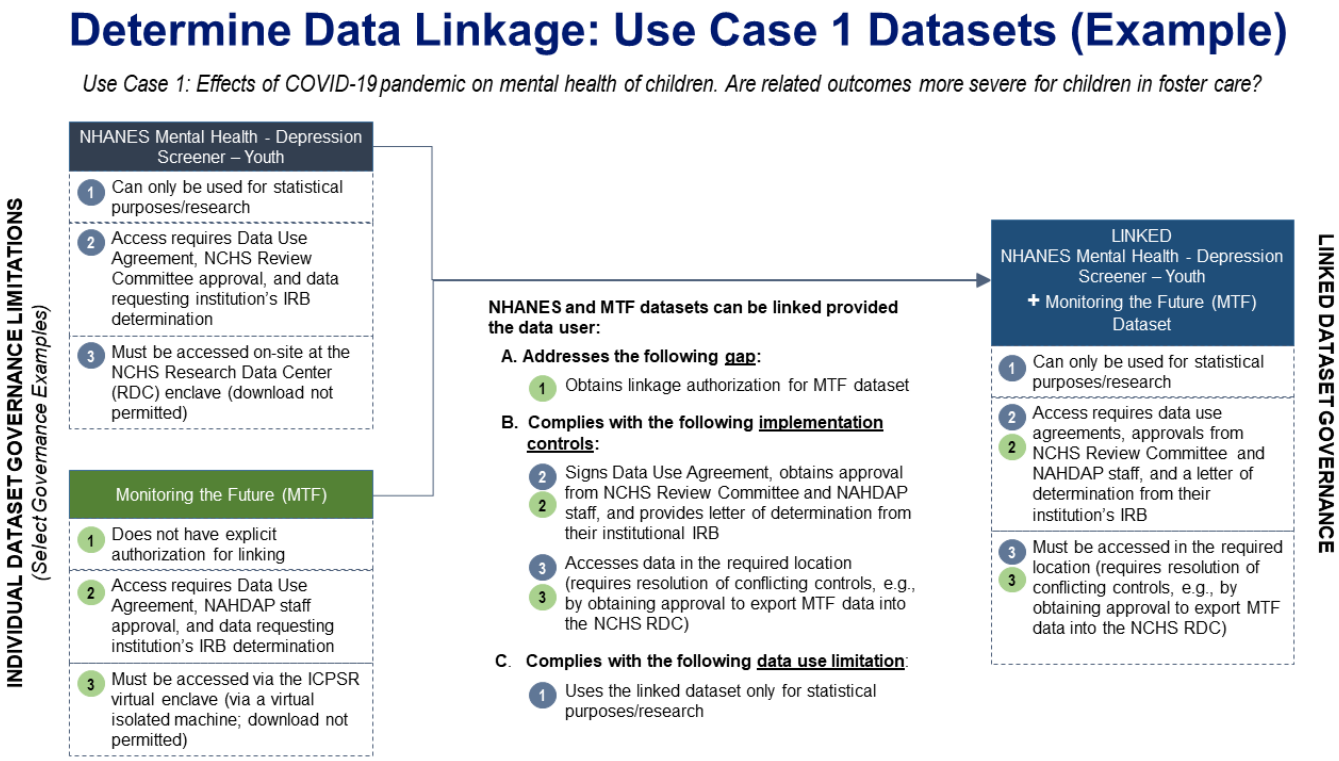


*Figure 3. Process of dataset linkage determination for two of the four datasets (NHANES and MTF) included for Use Case 1. For clarity in this illustration, a subset of data governance requirements is shown.*

The project team applied the Linkage and Use Determination Framework to each of the pairwise combinations of dataset for each use case. Finally, the project team compared the limitations, controls, and authorization gaps for all the datasets in a given use case and used the combined governance to determine whether and under what conditions all the datasets could be linked to address the use case.

### Outcomes from Linkage and Use Determination

The project team reviewed the results of the linkage and use determinations for each Use Case to determine a potential data user's ability to link datasets and the limitations and controls imposed on the linked dataset by the contributing datasets. The outcome of the dataset linkage and use determination for linking all datasets for each Use Case is shown in the following lists. The project team also performed pairwise linkage and use determinations for all dataset within each use case. Each pairwise dataset linkage and use determination for Use Cases 1, 2, and 3, is presented in Appendix E.

## Use Case 1: Effects of the COVID-19 pandemic on mental health of children. Are related outcomes more severe for children in foster care?

### ALL DATASETS: NHANES (Dataset 1), NSDUH (Dataset 2), MTF (Dataset 3), and AFCARS (Dataset 4)

**Question 1: Can the datasets be linked?**

Yes, NHANES, NSDUH, MTF, and AFCARS can be linked provided:

**NCHS RDC/ICPSR-NAHDAP/NDACAN staff:**

1. Shares NHANES, MTF and AFCARS data de-identified of all direct identifiers [Controls 1a, 3a, 4a]
   a. For NHANES, certain indirect identifiers (geography) may be included
   b. For MTF, state and ZIP code may be included
2. Performs disclosure review prior to sharing of (a) NHANES data by NCHS Disclosure Review Board/NCHS Confidentiality Officer, (b) MTF data, and (c) AFCARS data (removing county FIPS code with >1,000 records, recode date of birth to the 15th and adjust all other dates accordingly) [Controls 1b, 3b, 4b]
3. Reaches agreement with NHANES and MTF data sources on an approach to perform disclosure review that meets each data sources' requirements [Controls 1c, 3c]

**The researcher/user:**

1. Obtains authorizations for linkage and sharing for NSDUH and obtains authorization for linking for MTF and AFCARS [Authorization gaps 2a, 3a, 4a] - ***Assumption***
2. Ensures that NSDUH, MTF and AFCARS data used for linking to NHANES data must be either vital statistics, health, nutrition, or other related records [Limitation 1a]
3. Uses NHANES, NSDUH, MTF (after obtaining permission from ICPSR/NAHDAP staff) and AFCARS data within NCHS RDC [Limitations 1b, 2a, Controls 1h, 2g, 3e] - ***Assumption for MTF***
4. Uses linked NHANES, NSDUH, MTF, and AFCARS data for statistical purposes only [Limitation 1c, 2b, 3a]

5. Submit RDC application for accessing NSDUH [Control 2a]
6. Obtains approval from NCHS Confidentiality Officer, NCHS RDC, and DHANES/NHANES, SAMHSA staff, NAHDAP staff, and NDACAN staff on the proposed linkage [Controls 1d, 2b, 3g, 4d]
7. Signs and completes the NHANES Data Use/Access Agreement, NHANES Non-Disclosure CIPSEA Agent Form, NSDUH Designated Agent Form, NSDUH Data Access Agreement (DAA), NAHDAP VDE RDUA, NDACAN Terms of Use Agreement [Controls 1e, 1f, 2c, 2d, 2f, 3d, 4c]
8. Completes confidentiality training for NHANES and NSDUH data access [Controls 1g, 2e]
9. Obtains IRB approval or exemption from their institution for accessing MTF [Control 3f]

**Question 2: What limitations do the linked datasets inherit?**

**For using NHANES, researchers/users:**

1a. Can only link NHANES data to vital statistics, health, nutrition, and other related records
1b. Must use NHANES data within the NCHS RDC (on-site enclave)
1c. Must use NHANES data only for health statistical reporting and analysis

**For using NSDUH, researchers/users:**

2a. Must use NSDUH data within the NCHS RDC
2b. Must use NSDUH data only for health statistical reporting and analysis

**For using MTF, researchers/users:**

3a. Must use MTF data for broad research or statistical purposes

**For using AFCARS, researchers/users:**

4a. Must use AFCARS data in accordance with their approved research described in Section I.1 of the NDACAN Terms of Use Agreement

**Question 3: What controls do the linked datasets require?**

**For sharing NHANES, NCHS RDC staff must**:

1a. Share NHANES data de-identified of all direct identifiers; certain indirect identifiers (geography) may be included
1b. Perform disclosure review prior to sharing the NHANES restricted-use data through the RDC (NCHS Disclosure Review Board/NCHS Confidentiality Officer)
1c. Perform disclosure review of the output before releasing it (RDC and DHANES)

**For sharing MTF, ICPSR/NAHDAP staff must:**

3a. Share fully de-identified data (for MTF restricted-use data, this does not include state and ZIP code)
3b. Perform disclosure review prior to sharing
3c. Perform disclosure review of analysis outputs prior to removing output data from the VDE

**For sharing AFCARS, NDACAN staff must:**

4a. Share AFCARS data de-identified of all 18 HIPAA identifiers
4b. Perform disclosure review of data prior to sharing (removing county FIPS code with >1,000 records, recode DoB to the 15th and adjust all other dates accordingly)

**For accessing NHANES, researchers/users must:**

1d. Obtain approvals from NCHS Confidentiality Officer, NCHS RDC, and DHANES/NHANES on the proposed research
1e. Execute Data Use/Access Agreement (Rules of Behavior)
1f. Sign Designated Agent Agreement
1g. Complete confidentiality training
1h. Access data within NCHS RDC (on-site enclave)

**For accessing NSDUH, researchers/users must:**

2a. Submit RDC application
2b. Obtain approval from SAMHSA staff on the research proposed in the RDC application
2c. Sign Designated Agent Form (DAF)
2d. Sign Data Access Agreement (DAA)
2e. Complete confidentiality training
2f. Sign the SAMHSA RDC Student Data User Acknowledgement form and obtain advisor's signature, if researcher/user is a student
2g. Access data from the NCHS RDC

**For accessing MTF, researchers/users must:**

3d. Execute NAHDAP VDE RDUA between ICPSR (U-Mich) and the researcher's institution
3e. Access data only through the ICPSR VDE (virtual enclave)
3f. Obtain IRB approval or exemption from the researcher's institution
3g. Obtain review and approval from NAHDAP on the proposed research

**For accessing AFCARS, researchers/users must:**

4c. Execute of the NDACAN Terms of Use Agreement
4d. Obtain review and approval from NDACAN staff on the proposed research

**Question 4: What authorization gaps exist?**

**For NSDUH:**

2a. Information on authorizations for linkage and sharing is not available/found

**For MTF:**

3a. Information on authorizations for linkage is not available/found

**For AFCARS:**

4a. Information on authorizations for linkage is not available/found

## Use Case 2: What is the impact of COVID-19 infection on pediatric cancer survivors? Or what is the impact of COVID-19 infection on future pediatric cancer outcomes?

### ALL DATASETS: NCCR (Dataset 1), CDC COVID (Dataset 2), and TMSIS (Dataset 3)

**Question 1: Can the datasets be linked?**

Yes, NCCR, CDC COVID, and T-MSIS can be linked provided:

**A. SEER/CDC COVID/CCW VRDS staff:**

1. Share NCCR, CDC COVID, and T-MSIS data fully de-identified of all direct identifiers [Controls 1a, 2a, 3a]

2. Perform disclosure review [Controls 2b, 3d]

   a. For CDC COVID data, to suppress data fields with low frequency (<5) prior to sharing
   b. For T-MSIS data, to review analysis outputs prior to sharing output data from the VRDC

**B. Federal entities who create linked T-MSIS data and share the linked data:**

3. Take the data into their SORN under the Privacy Act when performing data linkage using non-standard TAFs containing PII [Control 3b]

4. Treat secondarily shared data as if they are a HIPAA Covered Entity and follow a process similar to CMS for releasing data including entering into a DUA with the researcher [Control 3c]

5. Link NCCR data only based on applicable state laws for participating NCCR registries, in accordance with protocol for linkage approved by the NCCR data providers, and the RIF Application approved by ResDAC team and CMS Privacy Board and in according with the Information Exchange Agreement (IEA) between

CMS the Participating Agency - all of which specify the scope of linkage [Limitations 1a, 3a]

**C. The researcher/user:**

1. Obtains authorizations for data linkage and sharing for CDC COVID [Authorization gap 1] – ***Assumption***

2. Uses the linked NCCR, CDC COVID and T-MSIS data for broad research use (must be of public health significance) that justifies the initial disclosure and solely for the study described in the T-MSIS RIF Request Application, and ensures findings are publicly available [Limitations 1b, 2a, 3b]

3. Submits NCCR Data Analysis Plan, NCCR DUA, and CDC COVID Restricted Access Data Use Agreement (RIDURA), T-MSIS RIF Data Use Agreement, Attachment A: RIF Application, RIF Application Key Personnel Supplement, RIF Specifications Worksheet, and Data Management Plan Self-Attestation Questionnaire (DMP SAQ) [Controls 1b, 1c, 2b, 3e]

4. Obtains approval from NCI Office of Data Sharing and the Surveillance Research Program's Data Release group and ResDAC team on the proposed linkage [Controls 1d, 3f]

5. Obtains IRB LOD from the researcher's institution for NCCR, CDC COVID and T-MSIS data (or NCI BRANY as needed for NCCR) and IRB approval from the NCCR state registry [Controls 1e, 1f, 2c, 3h]

6. Uses an institutional account (known as eRA Commons) and obtains verification of Signing Official by the NCI Office of Data Sharing and the Surveillance Research Program's Data Release group for NCCR [Control 1g]

7. Works with T-MSIS, NCCR, and CDC COVID staff to establish federal agency authorization for T-MSIS linkage with NCCR and CDC COVID and to upload CDC COVID data obtained from CDC GitHub private repository and T-MSIS data obtained via encrypted shipped disks to SEER*Stat (client server software), where

NCCR data can be accessed [Controls 1h, 2e, 3i] – ***Assumption for NCCR, CDC COVID and T-MSIS***

**Question 2: What limitations do the linked datasets inherit?**

**For using NCCR, researchers/users:**

1a. Must only link NCCR data based on applicable state laws which specify the scope of linkage (e.g., Idaho) and in accordance with protocol for linkage approved by the data providers (i.e., state registry)

1b. Must use or disclose NCCR data only for the purposes for approved research

**For using CDC COVID, researchers/users:**

2a. Must use CDC COVID data for broad research (must be of public health significance)

**For using TMSIS, researchers/users:**

3b. Must use T-MSIS data for research use that justifies the initial disclosure and solely for the study described in detail in the RIF Request Application, and must ensure findings are publicly available

- **Federal entities who create linked T-MSIS data:**

3a. Must only link T-MSIS data in accordance with the RIF Application approved by the CMS Privacy Board which specifies the scope of linkage and in accordance with the Information Exchange Agreement (IEA) between CMS and Participating Agency which specifies the scope of linkage for federal entities performing linkage with non-standard TAFs containing PII

**Question 3: What controls do the linked datasets require?**

**For sharing NCCR, SEER staff must:**

1a. Share NCCR data fully de-identified of all direct identifiers (including exact ages and dates and geographic information, except quintiles)

**For sharing CDC COVID, staff must:**

2a. De-identify data of all direct identifiers

2b. Perform disclosure review to suppress data fields with low frequency (<5) prior to sharing CDC COVID data

**For sharing TMSIS:**

**A. Federal entities who create linked T-MSIS data and share the linked data must:**

3a. De-identify T-MSIS data of all 18 HIPAA identifiers as per HIPAA

3b. Take the data into their SORN under the Privacy Act when performing data linkage using non-standard TAFs containing PII

3c. Treat secondarily shared data as if they are a HIPAA Covered Entity and follow a process similar to CMS for releasing data including entering into a DUA with the researcher

**B. CCW VRDC staff must:**

3d. Perform disclosure review of analysis outputs prior to sharing output data from the VRDC

**For accessing NCCR, researchers/users must:**

1b. Submit Data Analysis Plan

1c. Execute the NCCR DUA

1d. Obtain review by and approval from the NCI Office of Data Sharing and the Surveillance Research Program's Data Release group on the proposed research

1e. Obtain IRB LOD from the researcher's institution, or from NCI central IRB (BRANY) if the researchers institution does not have an IRB (as needed/applicable)

1f. Obtain IRB approval from the state registry

1g. Use an institutional account (known as eRA Commons) and obtain verification of Signing Official by the NCI Office of Data Sharing and the Surveillance Research Program's Data Release group

1h. Access the data from SEER*Stat (client server software)

**For accessing CDC COVID, researchers/users must:**

2b. Execute the Restricted Access Data Use Agreement (RIDURA)

2c. Obtain IRB LOD from researcher's institution (as needed/applicable)

2d. Access data from CDC GitHub private repository

**For accessing TMSIS, researchers/users must:**

3e. Submit RIF Data Use Agreement, Attachment A: RIF Application, RIF Application Key Personnel Supplement, RIF Specifications Worksheet, and Data Management Plan Self-Attestation Questionnaire (DMP SAQ)

3f. Obtain review and approval from ResDAC team on the proposed research

3g. Obtain review and approval from CMS' Data Privacy Safeguard Program (DPSP) on the Data Management Plan Self-Attestation Questionnaire (DMP SAQ)

3h. Obtain IRB LOD from the requesting institution

3i. Access data from the VRDC or through encrypted shipped disks

**Question 4: What authorization gaps exist?**

**For CDC COVID:**

2a. Information on authorizations for linkage and sharing is not available/found

## Use Case 3: SARS-CoV-2 Vaccination and Asthma-Related School Absence – Does SARS-CoV-2 vaccination result in reduced asthma-related school absences at 3/6/12+ months post-vaccination?

### ALL DATASETS: N3C (Dataset 1), PEDSnet (Dataset 2), RADx-UP (Dataset 3), and EPA AQS (Dataset 4)

**Question 1: Can the datasets be linked?**

Yes, N3C, PEDSnet, RADx-UP, and EPA can be linked provided:

**A. N3C/PEDSnet/RADx Data Hub/EPA staff:**

1. Share N3C, PEDSnet, and RADx data de-identified of all direct identifers
    a. for N3C, limited datasets or synthetic datasets can also be shared; ZIP codes entirely for all geographic units containing 20,000 or fewer people should be removed; and full five-digit ZIP codes of predominantly AI/AN communities should be replaced with partial ZIP codes
    b. for PEDSnet, HIV-related data and reproductive and mental health care data for minors should be removed
    c. for EPA, full geographic identifiers including site address, ZIP code, CBSA, county, and state are shared

    [Controls 1a, 1b, 2a, 2b, 3b, and 4a]
2. Has waiver of consent from NIH IRB for sharing data through the NCATS N3C Platform [Control 1c]

3. Performs risk review prior to sharing of PEDSnet data (data transformations, such as date shifts, replacement labels for free text fields and geographic information, and removing HIV/pregnancy/mental health data) [Control 2c]

4. Ensures the RADx studies are registered in dbGaP [Control 3a]

**B. N3C Data Providers:**

5. Execute a Data Transfer Agreement (DTA) with NCATS [Control 1d]

6. Obtain institutional or external IRB approval [Control 1e]

**C. The researcher/user:**

1. Uses the linked N3C and PEDSnet data for general COVID-19 research purposes specified and approved by the PEDSnet participating sites and PEDSnet Steering Committee [Limitations 1b, 2b, and 3a]

2. Does not use the linked data to make assumptions about Tribal affiliation [Limitation 1c]

3. Complies with the N3C Community Guiding Principles and the Attribution and Publication Principles [Limitation 1e]

4. Has an eRA commons or Login.gov account [Control 3c]

5. Executes the Institutional Data Use Agreements (DUA) with NCATS and PEDSnet and Responsible Use of Data Agreement (RUD) with PEDSnet [Controls 1f and 2f]

6. Submits Data Use Request (DUR) for approval by N3C Data Access Committee, request form for approval by the PEDSnet Research Committee, and Data Access Request (DAR), which includes the Data Use Certification (DUC) Agreement, the Genomic Data User Code of Conduct, and the RADx SM Data User Code of Conduct [Controls 1h, 2d, and 3d]

7. Ensures the Signing Official from the investigator's institution reviews, approves, and co-signs the request [Control 3e]

8. Completes NIH IT training, attests to the N3C Data User Code of Conduct, and completes Human Subjects Research Protection training to access N3C data [Control 1i]

9. Provides IRB letter of determination for N3C data access and if determined to be Human Subjects Research, provide IRB approval with IRB reliance for site providing data (NPRA Master Reliance Agreement (MRA) or SMART IRB MRA) for PEDSnet [Controls 1k and 2e]

10. Obtains approvals from PEDSnet prospective site PI approval and PEDSnet Executive Committee, the AHARO Center/Comprehensive Health Center IRB and RADx Data Hub Data Access Committee on the proposed linkage [Controls 2g, 2h, 3h, and 3f]

11. Works with N3C staff to obtain Class 2 or Class 0 designation for PEDSnet and RADx-UP datasets so that all three datasets can be linked using PPRL [Limitations 1d and 2c] - ***Assumption***

12. If non-N3C data are designated as Class 2, uses/accesses N3C data within the N3C Enclave and obtains approval from PEDSnet staff to export PEDSnet data into the N3C Enclave and RADx Data Hub staff to export RADx-UP data into the N3C Enclave [Limitations 1a and 2a; Controls 1j, 2i, and 3g] - ***Assumption for PEDSnet and RADx-UP***

13. Ensures they have an existing institutional N3C Data Use Agreement, dual authentication and authorization, signed institutional linkage honest broker agreement for multiple datasets, an approved data use request (DUR) by the data access committee (DAC), and local institutions IRB letter of determination for N3C Class 2 or Class 0 designation for PEDSnet and RADx-UP linkage. If non-N3C data are designated as Class 0, ensures they also have an Interconnect agreement. [Control 1k]

14. Has approval from participating sites for linkage with the external dataset and the External Dataset Committee in the Tools and Resource subgroup and NCATS for EPA linkage [Limitations 1l and 1m]

Note: Controls 4a and 4c are not required for this linkage as EPA Air Quality Data has already been brought into the N3C Enclave and is already available for linkage to N3C.

**Question 2: What limitations do the linked datasets inherit?**

**For using N3C, researchers/users must:**

1a. Use N3C data within the N3C Enclave [N3C]

1b. Use N3C data for COVID-19 general research purposes [N3C]

1c. Not use AI/AN data and ZIP code information to make assumptions about Tribal affiliation [N3C]

1d. Work with N3C staff to link Class 2 or Class 0data using PPRL [N3C]

1e. Comply with the N3C Community Guiding Principles and the Attribution and Publication Principles [N3C]

**For using PEDSnet, researchers/users must:**

2a. Use the data in a workspace within the PEDSnet cloud enclave--OR--at their own institution if approved to have the data transferred to their institution by all PEDSnet institutions providing data for the request

2b. Use the data for purposes specified and approved by participating sites and the Steering Committee, namely using data from real-world clinical settings for research, qualitiy measurement, and improvement/advancement of child health, particularly studies that inform or directly address clinical decision making, including retrospective observational studies.

2c. Work with PEDSnet staff to link data conducted under a waiver of consent using PPRL

**For using RADx-UP, researchers/users must:**

3a. Use RADx-UP data for general research purposes

**Question 3: What controls do the linked datasets require?**

**For sharing:**

**A. N3C staff must:**

1a. Share N3C limited datasets (LDS), de-identified datasets, or synthetic datasets [N3C]

1b. Remove ZIP codes entirely for all geographic units containing 20,000 or fewer people and replace full five-digit ZIP codes of predominantly AI/AN communities with partial ZIP codes [N3C]

1c. Have waiver of consent from NIH IRB for sharing data through the NCATS N3C Platform [N3C]

**B. Data providers must:**

1d. Execute a Data Transfer Agreement (DTA) with NCATS [N3C]

1e. Obtain institutional or external IRB approval [N3C]

**For sharing, PEDSnet staff must:**

2a. Remove HIV-related data and reproductive and mental health care data for minors

2b. De-identify individual level data using the Safe Harbor method of de-identification of PHI

2c. Perform a risk review on the requested datasets as well as data transformations, such as date shifts, replacement labels for free text fields and geographic information, and removing HIV/pregnancy/ mental health data

**For sharing, RADx Data Hub stuff must:**

3a. Ensure the studies are registered in dbGaP

3b. Ensure that the data is de-identified by working with study teams to de-identify ZIP codes, shift dates, and adjust ages into categories for specific ages

**For sharing, EPA staff must:**

4a. Host ambient air data, which contains full geographic identifiers including site address, ZIP code, CBSA, county, and state, through EPA's Air Quality System (AQS)

**For accessing N3C, researchers/users must:**

1f. Execute Institutional Data Use Agreement (DUA) with NCATS

1g. Submit Data Use Request (DUR) for approval by N3C Data Access Committee

1h. Complete NIH IT training, attest to the N3C Data User Code of Conduct, and complete Human Subjects Research Protection training

1i. Provide IRB letter of determination for data access

1j. Access the data within the N3C Enclave

**For accessing PEDSnet, researchers/users must:**

2d. Submit request form for approval by the Research Committee

2e. Undergo IRB review/determination (Human Subjects Review)

i. If IRB determines the proposed study is NHSR, then no further review/MRA required

ii. If IRB determines the proposed study HSR, the requester must provide IRB approval with IRB reliance for site providing data (NPRA MRA or SMART IRB MRA)

2f. Sign DUA (Data Use Agreement) and RUD (Responsible Use of Data) (Legal Review)

2g. Receive prospective site PI approval (Institutional Participation Approval)

2h. Receive PEDSnet Executive Committee approval (Network Participation Approval)

2i. Access the data through a workspace within the PEDSnet cloud enclave--OR--have the data transferred to their institution, the PEDSnet Study Approval request should specify, pending approval from all PEDSnet institutions providing data for the request

**For accessing RADx-UP data, researchers/users must:**

3c. Have an eRA commons or Login.gov account

3d. Submit a Data Access Request (DAR), which includes the Data Use Certification (DUC) Agreement, the Genomic Data User Code of Conduct, and the RADx SM Data User Code of Conduct

3e. Ensure the Signing Official from the investigator's institution reviews, approves, and co-signs the request

3f. Receive approval from the Data Access Committee

3g. Access the data through RADx Data Hub Jupyter Notebooks

**For accessing EPA data, researchers/users must:**

4c. Obtain data from AQS, an open access repository

**For linking N3C data, researchers/users must:**

1k. Work with N3C staff to verify and complete the following requirements for N3C Class 0 or Class 2 linkages [N3C]:

i. Existing institutional N3C Data Use Agreement

ii. Dual authentication and authorization

iii. Signed institutional linkage honest broker agreement for multiple datasets

iv. Approved data use request (DUR) by the federally staffed Data Access Committee (DAC)

v. Local institution's IRB letter of determination

vi. Interconnect agreement (for Class 0 only)

1l. Have agreement from participating sites for linkage with the external dataset [N3C]

1m. Have approval from the External Dataset Committee in the Tools and Resource subgroup and NCATS for linkage [N3C]

**For linking RADx-UP data, researchers/users must:**

3h. Work with AHARO Center/Comprehensive Health Center IRB to obtain approval for individual level data linkages [RADx-UP]

**Question 4: What authorization gaps exist?**

- No authorization gaps exist

## Objective 3: Generate considerations for a governance metadata schema based on governance analysis

### *3-1: Summarize findings from governance analysis*

**Findings from Governance Information Collection and Structuring**

For the first step of the governance analysis, the project team collected and assessed publicly available governance information for the eleven selected datasets. To adequately document the full governance information landscape required for dataset linkage and use determination, the project team was required to dig deeper, reviewing other data source documentation and conferring with data source stakeholders in writing and by interview (methods for data gathering for each data source are described in Table 2). Findings from this process are summarized in the following sections.

*Collecting information from a variety of sources was required to fully understand a dataset's governance*

The project team began the collection of governance information with the data source and/or dataset public website. The team found that web pages or documents clearly listing all governance information for the data lifecycle did not exist, nor was the governance information consistently included on the websites or in publicly available materials. One data source, the National Patient-Centered Clinical Network (PCORNet), publishes two web pages dedicated to explaining the governance of PEDSnet data, making much, but not all, information directly available to researchers and users. This level of readily accessible governance information was found to be unusual.

For two data sources, the EPA Air Quality Systems (AQS) and the CDC COVID-19 Case Surveillance Data, the online information served as the only available resource for governance information, and in both cases the web pages provided incomplete

governance information. For example, despite extensive research, the project team was able to find only limited governance information for the EPA Air Quality Systems (AQS) data from the EPA website. The project team determined that the Clean Air Act, the United States' primary federal air quality law intended to reduce and control air pollution nationwide, specifically authorizes both data sharing and data collection by state, local, tribal, and/or other federal air pollution control agencies for reporting to the EPA. However, no information was available to determine the origin of authorization for linking, accessing, and using AQS data. The project team confirmed that, because the AQS database contains readings from thousands of monitoring instruments throughout the country, and not data related to individuals, the collection, linkage, sharing, access, and use of AQS data does not require authorization by consent, assent, IRB and/or Institutional Certification. Ultimately, the team found that the ambient monitoring AQS data are in the public domain, ensuring open access and use of the AQS data.

After exhausting the governance information available from data source websites and the documents publicly available through resource websites, the project team pursued other resources to collect governance information. In total, nine of the eleven data sources in this assessment required collection of governance data from two or more sources, including, importantly, stakeholder contact. For NHANES, NSDUH, MTF, AFCARS, NCCR, T-MSIS, N3C, PEDSnet, and RADx-UP, the project team was only able to fully understand the governance by gathering information from the data source website, various documents (e.g., consent forms, study protocols, and IRB letters) provided by dataset stakeholders, and both written communication and interviews with stakeholders who provided governance information and confirmed the completeness and accuracy of information that was collected from other sources. The engagement of multiple dataset stakeholders, including Principal Investigators, project managers, program officials, technical leads, and legal staff, was often required to obtain the full governance information. For example, when collecting RADx data governance information, the project team gathered information directly from the two project co-PIs through interviews and the RADx Data Hub Program Director via email. Finally, the predecessor PPRL report served as a source of governance information for data sources used in both the prior project and this assessment.

The resources used by the project team to gather governance information for all datasets are summarized in Table 2 and described in detail for each dataset in Appendix D.

*Many of the sources of governance information are not publicly available or easy to access.*

Outside of data source websites and public information on laws impacting data governance and controls, the project team collected governance information from

materials and information that were accessed through personal contact with data source stakeholders. For example, the team collected governance information for NHANES from the NHANES/CDC website, but critically important to completing the assessment were discussions with two senior stakeholders at the National Center for Health Statistics (NCHS) and the Division of Health and Nutrition Examination Surveys (DHANES). In response to specific questions from the team, the stakeholders provided the project team with an NHANES linkage information document that is not publicly available.

The stakeholders also provided key clarifying information. For example, while the NHANES consent states that data may only be used for statistical purposes, stakeholders confirmed that this broad statement infers that the data may be used for general research purposes. The project team would not have been able to gather complete governance information or clearly and accurately interpret the information without access to data source stakeholders.

In some cases, the data source stakeholders were unable to share primary documentation with the project team. Thus, information required to determine the governance specifications, such as IRB approval language for data collection, linking, sharing, accessing, and use, and had to be collected by stakeholder interview.

Additionally, the project team found that certain governance information, as well as the process for establishing authorizations, for instance internal approvals required for data linking or sharing, are often not formally documented by the data source organization or institution. This information was then not available to the team for use in this assessment.

*Contact information for an individual who could answer questions about the dataset governance is not often publicly available*

The project team reviewed public websites and publicly available materials to determine the appropriate contact for the data source and typically found that personal contact information was not published. For data sources with help desk emails, the team submitted forms (e.g., CDC) and drafted emails (e.g., AFCARS) to initiate contact with the data source. The team then corresponded via email with these teams to identify the appropriate data source point of contact.

In a few cases, the project team was able to identify some data source points of contact based on contact information collected in the predecessor report (e.g., N3C, PEDSnet).

Finally, when all avenues for locating the data source contact from public resources were exhausted, the project team reached out to the members of the Project Governance Team (PGT) for leads or relied on ODSS contacts within HHS. The PGT and ODSS

provided the critical initial contact with the agency or program for nine of the eleven data sources, i.e., NHANES, NSDUH, MTF, AFCARS, NCCR, T-MSIS, N3C, PEDSnet, and RADx-UP. The project team then corresponded with the agency or program contacts to obtain the appropriate point of contact for the data source.

*Many datasets do not have explicitly stated authorization for linkage in the dataset documentation*

The project team found that governance documentation, when available, often lacked a definite statement of governance information. For example, informed consent forms, when applicable to the selected dataset, tended to contain vague language about "combining data" which does not explicitly allow or disallow data individual-level linkage.

For example, the consent form for RADx contains language about "combining data" that does not explicitly authorize individual-level data linkage: "The project is funded by the National Institutes of Health (NIH). It is part of the NIH RADx-UP, a health research program to learn more about Covid-19 disease. If you join RADx-UP, we will gather some data (information) about your child. We will combine the data from all who join, to understand how to help more people at risk for or with COVID-19." Although the RADx-UP investigators interpret this language as potentially allowing future data linkage, the project team considers this language to be too ambiguous (possibly, for example, meaning combining data from multiple participants) and for the purposes of this assessment concluded that the consent does not authorize or specify data linkage.

The consent for RADx participants residing in Hawaii contains language informing participants that the data will be used in the overall cohort for RADx-UP, including language that leaves the option for the data to be used for record linkage: "Your data can be used for other research studies." Through stakeholder interviews, the project team learned that the AHARO Health Centers can authorize individual-level data linkage. Linkage is possible but any linkage at an individual level must be approved by the Comprehensive Health Center IRB (AHARO Health Center). Anything beyond general research purposes, such as linking data and working with identifiers, must be approved by the community IRB. This information, however, is not available from any of the public websites or study documents made available to the project team.

*Data access and use authorizations are typically conveyed in data use agreements (DUAs) and other repository agreements with end-users and are rarely included in other data governance documentation*

Authorizations for data access are often formalized through data use agreements (DUAs) with secondary users such as researchers. For only two of the eleven data sources, AFCARS and the EPA AQS datasets, the project team was able to find explicit language

authorizing data access and use outside of DUAs. The other nine data sources rely on the execution of DUAs and the language executed in those agreements to authorize data access.

AFCARS data is managed by the National Data Archive on Child Abuse and Neglect (NDACAN). In discussion with the NDACAN team, the project team learned that NDACAN is authorized by the Children's Bureau to provide access to AFCARS data. The Children's Bureau authorizes NDACAN to share the data with researchers under contractual Terms of Use. Additionally, NDACAN provides researchers with authorization to use the AFCARS datasets, as formatted and provided by NDACAN, upon NDACAN approval of a Terms of Use Agreement. The Terms of Use agreement contains a description of the investigator's research purpose and affirmation to appropriately safeguard the data and limit the use of the data to research. The Agreement further prohibits re-distribution of the data, any attempt to identify individuals, and other activities defined as misuse in the Agreement and by U.S. laws. The Terms of Use agreement is publicly available on the NDACAN website: https://www.ndacan.acf.hhs.gov/datasets/request-dataset.cfm.

The EPA allows both access and use of Air Quality System (AQS) data. The AQS policy publicly states that the ambient monitoring data in EPA's AQS are public domain, and researchers are welcomed to download the data and use it freely. This information is available on the EPA Air Quality Data website: https://www.epa.gov/outdoor-air-quality-data/do-i-need-request-permission-use-monitoring-data-and-graphics-airdata.

Conversely, for many datasets, data use or access agreements serve as the primary place where data access authorizations are documented. Examples include NHANES, NSDUH, MTF, NCCR, CDC COVID, and T-MSIS. Together the terms and conditions of such agreements and other data repository policies document the rules on how the data must be accessed and used even if some of these rules are derived from upstream regulatory or other requirements (e.g., Federal Privacy Act, HIPAA, and Common Rule in the case of T-MSIS).

*State-level laws impact which data are collected for cancer registries, which impacts downstream agreements*

The National Childhood Cancer Registry (NCCR) relies on the agreements and policies unique to each of the twenty-three participating states. State cancer registries collect information about cancer diagnosis and treatment according to state laws. In some cases, the state health department directly manages the activities of the central cancer registries but in other cases this is contracted to another entity, usually a university (as is the case in New Jersey, Florida, Georgia, Kentucky, and others). Regulations vary by state. For example, Michigan registries are permitted to hold incidence but not longitudinal

data. Idaho Code Title 57, Chapter 17, 57-1703 (6) defines the scope of data collection. It becomes important to read state statutes and regulations to understand the provenance of data governance of cancer registry data so the associated requirements are incorporated in downstream agreements. This principle is important to consider for other laws as well.

*Permitted scope of linkage is not always explicitly documented and is sometimes determined on a case-by case basis*

When a dataset is authorized for linkage, it is further necessary to identify the scope of linkage, i.e., with what datasets the data may (or may not) be linked. To fully understand the scope of linkage for the datasets reviewed for this assessment, the project team often needed to interview the data source stakeholders. In most cases, the scope of linkage was not discoverable in documentation. In some cases, the scope of linkage relies, at least in part, on case-by-case approval by a governing body or committee.

The scope of data linkage for the eleven datasets is summarized in Tables 5, 6, and 7. For only one of the Use Case 1 datasets, NHANES, is the scope of linkage explicitly documented. The information is found in the Consent Template for 'Home Interview Consent' that states: "Health research using NHANES can be enhanced by combining your survey records with other data sources. The data gathered are used to link your answers to vital statistics, health, nutrition, and other related records." The consent form was found on the CDC website, and all information in Table 5 was confirmed by NCHS and DHANES staff. The NCHS ERB and DHANES/HANES approvals for linkage were uncovered through interviews with NCHS points of contact. For the other datasets in Use Case 1, no specific language regarding data linkage is in the documentation (noted as "Does not authorize/specify").

The results look much the same for Use Cases 2 and 3, as for Use Case 1, with scope of linkage details fully or partially unspecified, with some notable exceptions. For example, the scope of linkage for the Use Case 2 NCCR dataset was gathered from both the public website and in discussion with the NCCR technical lead. For NCCR, data originator agreements specify that the data can be linked according to the protocol for linkage approved by the data provider. For instance, the laws of some states contributing data to NCCR specify the scope of linkage. Idaho Code Title 57, Chapter 17, 57-1703 (6) specifies that linkages to gather treatment and other information are specifically allowed in 57-1805 (5), which states, "Nothing in this chapter shall prevent the department or authorized contractor from identifying and reporting cases using data linkages with death records, statewide cancer registries, and other potential sources."

The project team found that the permitted linkage scope for the Use Case 2 T-MSIS dataset was not available to the public on either the CMS (T-MSIS data source) or the Research Data Assistance Center (ReSDAC) websites. The team interviewed staff from the Center for Medicaid and CHIP Services (CMCS), Data and Systems Group (DSG) and the Office of Enterprise Data and Analytics (OEDA), Data Development and Services group to gather information on the scope of linkage. The team was able to determine that a researcher may request data linkages as part of a research protocol. CMS policy and Data Use Agreements specify that T-MSIS data may only be linked to external data in accordance with a CMS Privacy Board-approved Research Identifiable Files (RIF) Application that includes the protocol for the linkage of specific files. Data linkage with non-standard T-MSIS Analytic Files (TAFs) that contain additional variables is available only to federal entities, and in this case the federal entity must secure approval from the CMS Privacy Board and submit a letter of justification to the CMS Chief Data Officer. The letter of justification must include the name of other data files or sources of information to be included in the data linkage, the purpose for using the TAF in the analysis, and the process to be used for data linkage. Thus, the scope of linkage for this dataset is determined on a project-by-project basis during the application process.

The project team investigated the Use Case 3 N3C scope of linkage using the National Center for Advancing Translational Science (NCATS) website, interviews, and written exchanges with NCATS informatics team leadership, and information from the predecessor PPRL report. Bringing together information from these resources, the team was able to determine that N3C data may be linked with external datasets when the external datasets are classified as Class 2 or Class 0 by the External Dataset Committee in the Tools and Resource subgroup and NCATS, both of which approve the datasets for import and linkage with N3C data. N3C electronic health record (EHR) datasets will be part of the linkage with the external datasets if the EHR data contributor has signed a formal Linkage Honest Broker Agreement (LHBA) with Regenstrief Institute and agreed to link with the “external” datasets that are accessible through the enclave for Class 2 datasets or outside the enclave for Class 0 datasets (6). Class 2 linkage has already been implemented for the Mortality and CMS data that reside within N3C. Class 0, or linkage with datasets outside the enclave, has not yet been implemented.

Although not specified in the following tables, data governance that impacts the scope of linkage are most often derived from data collection and data linkage governance origins (review Appendix E).

**Table 5: Summary of governance that specifies whether data can be linked and the scope of data linkage based on all authorizations and applicable regulations/policies for Use Case 1.**

| Governance for Scope of Data Linkage | NHANES | NSDUH | MTF | AFCARS |
|---|---|---|---|---|
| Whether the data can be linked | Assent/Consent, NCHS ERB approval, and DHANES/HANES approval specify that the data can be linked. | Does not authorize/specify | Does not authorize/specify | Does not authorize/specify |
| With what other data can it be linked or can it not be linked (scope of linkage) | Assent/Consent specifies that the data can be linked to vital statistics, health, nutrition, and other related records. | Does not authorize/specify | Does not authorize/specify | Does not authorize/specify |

**Table 6: Summary of governance that specifies whether data can be linked and the scope of data linkage based on all authorizations and applicable regulations/policies for Use Case 2.**

| Governance for Scope of Data Linkage | NCCR | CDC Covid | T-MSIS |
|---|---|---|---|
| Whether the data can be linked | Certain state laws/regulations specify that the data can be linked (e.g., Lousiana).<br><br>Data originator agreements specify that the data can be linked. | Does not authorize/specify | CMS Privacy Board authorizes data linkage for researchers for research purposes via an approved RIF Application that specifies the scope of linkage.<br><br>CMS Privacy Board, Chief Data Officer approval of a letter of justification for linkage, and Information Exchange Agreement (IEA) authorizes data linkage for federal entities performing linkage with non-standard TAFs containing PII. |
| With what other data can it be linked or can it not be linked (scope of linkage) | Certain state laws/regulations specify the scope of linkage (e.g., Idaho).<br><br>Data originator agreements specify that the data can be linked according to the | Does not authorize/specify | CMS policy and Data Use Agreement specify that data can only be linked to external data in accordance with the RIF Application approved by the CMS Privacy Board which specifies the scope of linkage. |

| Governance for Scope of Data Linkage | NCCR | CDC Covid | T-MSIS |
|---|---|---|---|
| | protocol for linkage approved by the data provider. | | Information Exchange Agreement (IEA) between CMS and Participating Agency specifies the scope of linkage for federal entities performing linkage with non-standard TAFs containing PII. |

**Table 7: Summary of governance that specifies whether data can be linked and the scope of data linkage based on all authorizations and applicable regulations/policies for Use Case 3.**

| Governance for Scope of Data Linkage | N3C | PEDSnet | RADx-UP | EPA |
|---|---|---|---|---|
| Whether the data can be linked | LHBA specifies that the data can be linked.<br>Participating PPRL sites specify that data can be linked with particular external datasets.<br><br>The External Dataset Committee in the Tools and Resource subgroup and NCATS approval specifies that data can be linked. | Consent (when obtained) specifies that data can be linked.<br><br>PEDSnet Steering Committee approval specifies that data can be linked according to the approved research plan.<br><br>Individual PEDSnet sites, through a study participation vote, specify that the sites can participation in data linkage on a study-by-study basis.<br>IRB specifies that PEDSnet data can be linked for research conducted under | Parental informed consent and assent do not specify linkage.<br>Raw language referring to "other research studies" is interpreted by the study PI as leaving the option open for data linkage.<br><br>AHARO Health Centers/Comprehensive Health Center IRB specifies that data can be linked at an individual level only if the IRB approves the linkage. | Does not authorize/specify |

| Governance for Scope of Data Linkage | N3C | PEDSnet | RADx-UP | EPA |
|---|---|---|---|---|
| | | a waiver of consent. | | |
| With what other data can it be linked or can it not be linked (scope of linkage) | N3C policy designation of external datasets for linking specifies that:<br>a. External datasets must be classified as Class 0, 2, 3, or 4 to be considered for N3C linkage. A dataset which is categorized as class 2 can be imported but will require hashing.<br>b. Class 1 linkages are not permitted.<br>Participating PPRL sites specify linkages with external datasets on a case-by-case basis.<br>External Dataset Committee in the Tools and Resource subgroup and NCATS determines the scope of linkage by approving external datasets for import and linkage within N3C. | IRB specifies that data can be linked using PPRL for research conducted under a waiver of consent.<br>Individual PEDSnets study sites specify the scope of data linkage on a study-by-study basis. | AHARO Health Centers/Comprehensive Health Center IRB specifies that any data linkages at an individual level or outside of general research purposes must be approved by the AHARO Health Centers IRB. | Does not authorize/specify |

## Findings from Dataset Linkage and Use Determination

*To link datasets, a researcher must discover and interpret the data access requirements of each of the to-be-linked datasets*

To fully understand the data access requirements of each dataset to be linked, a researcher or entity seeking the linkage must collect and evaluate the wide breadth of information that comprises data access governance, e.g., data access or use agreements,

repository policies, federal/state regulations. There may be multiple sets of complex data access requirements to navigate.

The project team found a wide variety of sources of information pertaining to data access governance for this assessment. For a single dataset, NHANES, the team discovered a series of applicable policies, including Protected Data Policy, Section 308(d) of the Public Health Service Act, the Confidential Information Protection and Statistical Efficiency Act (CIPSEA), all of which ensure the confidentiality of collected data, and each of which had to be researched and interpreted. In addition to the policies, the process for receiving approval and the conditions for accessing NHANES data are not insignificant. To access NHANES data, a researcher or user must submit an application that specifies the proposed research questions, what data the research requires, and the desired output. The NCHS privacy officer is responsible for reviewing and approving the application. Before receiving the NHANES dataset, the researcher must complete confidentially training, execute a data use/access agreement outlining rules of behavior and data safeguards, and sign a non-disclosure CIPSEA agent form, which makes the researcher an agent under the statute and legally liable for proper use of the data. Additionally, before output can leave the RDC, RDC staff and data owners (DNAHES/NHANES) review the output to ensure it conforms to the approved application and cannot reidentify an individual.

Researching data access governance for the MTF dataset, the project team found that data access is authorized by the National Addiction & HIV Data Archive Program (NAHDAP) Restricted Data Use Agreement for Restricted Data in the Virtual Data Enclave (NAHDAP VDE RDUA) and the MTF PI at the University of Michigan, who authorizes data access through the Inter-university Consortium for Political and Social Research Virtual Data Enclave (ICPSR VDE). ICPSR and NAHDAP specify that for data access, a researcher must execute the NAHDAP VDE restricted Data Use Agreement between ICPSR (through the University of Michigan) and the researcher's institution, obtain IRB approval or exemption from the researcher’s institution, obtain review and approval from NAHDAP on the proposed research, and only access data through the ICPSR VDE.

For a pairwise linkage of NHANES and MTF data, a user must understand and comply with all of the data access requirements described above. For the full NHANES, MTF, NSDUH, and AFCARS linkage examined in Use Case 1, the governance becomes even more complicated. The researcher must thoroughly research the data access governance information for each dataset in the proposed linkage and synthesize the data access governance from each of the datasets to determine what access conditions and controls are passed to the resulting linked dataset. The synthesis may be nuanced and complex and rules and controls could possibly conflict, necessitating resolution (Figure 3).

*Authorizations for data linkage and sharing should be addressed prior to linking and using the data*

For this assessment, the project team collected and reviewed more than sixty authorizations, regulations and policies that allow for the collection, access, linkage, sharing, and use of the eleven selected datasets. All governance information passed down from a contributing dataset to the linked dataset was collected and interpreted to assess the ability to link datasets, and if linked, the rules that the linked dataset will inherit (from authorizations or other policies). Identifying authorizations for data linking and sharing can take time and possibly multi-modal research.

In some cases, the project team identified gaps in authorization for linking and sharing of datasets. In both the results of linkage determination for each of the three use cases, found in the "Outcomes from Linkage and Use Determination" section, and Appendix E, gaps in authorization and possible solutions for these gaps generated by the project team are noted as 'Assumptions' under Question 1, and authorization gaps are explicitly listed in the results of linkage determination under Question 4. Authorization gaps may require negotiation and resolution prior to linking the datasets and using the linked dataset. The project team's linkage determination identified authorization gaps in Use Cases 1 and 2. For three of the four Use Case 1 datasets - MTF, NSDUH, and AFCARS - information about authorizations for linkage is not available or was not found in the course of this assessment. Likewise, authorizations for linkage of Use Case 2 dataset, CDC COVID-19, were not available or not found. As described in the predecessor PPRL report, linkage authorizations are important for moving forward with record linkage. Where linkage authorization gaps exist, the onus is on the researcher or entity seeking the linkage to pursue authorization before datasets are linked and used.

Authorizations for sharing were also not found for Use Case 1 NSDUH and Use Case 2 CDC COVID-19; however, since sharing authorizations primarily refer to governance origins that allow for the exchange between a data originator and central system that makes the data accessible to other researchers, and there are already established mechanisms for making these data accessible to secondary users, it is possible these gaps are not barriers for moving forward with including these datasets in a record linkage implementation or use case.

*Linked dataset governance converges on the most constraining requirements of the contributing datasets*

A critical step in the linkage determination is evaluating how the governance from each contributing dataset impacts the governance of the resulting linked dataset, and what controls the linked dataset inherits from each data source. Governance is inherited from every step across the data lifecycle. The project team assessed the conditions of each of the fifteen pairwise linkages performed for this assessment and finally determined the conditions of the linkage of the collective datasets within each use case.

The project team found that requirements for location of data access are a source of particularly constraining inherited governance. In Use Case 3, a dataset resulting from the linkage of PEDSnet and RADx-UP data inherits complex data access constraints that require resolution. For accessing PEDSnet data there are two options: PEDSnet data may be accessed through a workspace within the PEDSnet cloud enclave, or the PEDSnet dataset may be transferred to a researcher's institution pending approval of the request from all PEDSnet institutions providing data for the proposed data linkage. RADx contributes the requirement that RADx-UP data must be accessed through RADx Data Hub Jupyter Notebooks. The project team determined that a researcher linking the PEDSnet and RADx-UP datasets must work with PEDSnet staff and RADx Data Hub staff to determine whether RADx-UP data downloaded through Jupyter Notebooks may be transferred to the PEDSnet cloud enclave workspace, or, alternatively, obtain approval to transfer PEDSnet data to the researcher's institution to integrate with the downloaded RADx-UP data. This process requires human action and approvals from multiple entities.

The project team similarly found that data use requirements often pass on constraining governance to linked data. The combined inheritance of the three datasets comprising Use Case 2 (NCCR, CDC COVID, and T-MSIS) imposes a set of data use requirements: the linked dataset must be used for broad research use; the research must be of public health significance; the research must justify the initial disclosure; the linked data may be used solely for the study described in the T-MSIS RIF Request Application; and the findings must be made publicly available. Thus, users of the linked data must adhere to all of the above data use requirements, including the most constraining requirement that the linked data must only be used for the study described in the T-MSIS RIF Request Application.

The project team found that constraints on the scope of linkage based on governance from one dataset limit which datasets can be linked for an entire use case. For example, all linked datasets that include NHANES data also inherit the NHANES requirement that NHANES data must only be linked to vital statistics, health, nutrition, and other related records, which impacts exactly which NSDUH, MTF and AFCARS data are linked for Use Case 1. However, the team found that more frequently, explicit restrictions on the scope of linkage were broad or were not stated at all and instead required review on a case-by-case basis from governing bodies, such as approved linkages for the T-MSIS, which are

based on the information provided in the T-MSIS RIF Request Application or N3C datasets, which are based on approvals from the PPRL participating site providing the data, the N3C External Dataset Committee, and NCATS.

*Apparent conflicts in governance inherited from contributing datasets introduce complexity in defining appropriate use of linked datasets*

In the course of this assessment, the project team found conflicts in data governance requirements of datasets to be linked within a Use Case. The governance policies of data repositories are designed to protect the privacy of participants represented in the data, and these policies came into play when linking datasets. Conflicts in data governance may require adjudication prior to data linkage and use, and a researcher or entity seeking the linkage may need to work with data source staff to resolve the conflicts. Possible solutions to conflicts in inherited data governance are indicated as "Assumptions" under Question 1 in the linkage determination in Table 6 and Appendix E.

With Use Case 1, the project team determined that access and use of NHANES data must take place within the NCHS RDC. The MTF data must be accessed through the ICPSR Virtual Data Enclave. To implement the NHANES and MTF pairwise dataset linkage or the four-dataset linkage (NHANES, NSDUH, AFCARS, and MTF), the researcher must find a way to resolve this conflict, for example by working with ICPSR/NAHDAP to obtain approval to export MTF data into the NCHS RDC. This is a major assumption as it requires ICPSR/NAHDAP to agree to let the user circumvent the access location requirement. All permissions must be secured before data movement, linkage, and use.

To execute the Use Case 2 all-dataset linkage, federal agency authorization is required to link the T-MSIS linkage with NCCR and CDC COVID datasets. The CDC COVID data obtained from the CDC GitHub private repository and T-MSIS data obtained via encrypted shipped disks could be uploaded to the to the client-server SEER*Stat application through which the NCCR data can be accessed.

As in Use Case 2, the location where the Use Case 3 datasets will be linked and used must be resolved before moving ahead with data linkage. For N3C, data must be accessed and used within the N3C Enclave, whereas PEDSnet data must be accessed within the PEDSnet cloud enclave or requesters may have the data transferred to their institution. Therefore, a researcher must obtain approval from PEDSnet staff and all PEDSnet data providers to treat N3C as the requester's institution and allow the PEDSnet data to be exported into the N3C Enclave. Additionally, a researcher must work with RADx Data Hub staff to obtain approval to export RADx-UP data into the N3C enclave. No permissions are required for the EPA AQS public-domain dataset to be linked with

PEDSnet, RADx-UP and N3C. The issue of location of linkage and use must be addressed before a linkage implementation moves forward.

*Linkage implementation must consider how the linked dataset is de-identified*

In the process of this assessment, the project team uncovered and noted different de-identification standards for sharing data applied to individual datasets. It is unclear, however, how the different standards will be applied to the linked dataset. For example, in Use Case 1, the AFCARS dataset follows the HIPAA Safe Harbor standard, meaning it does not contain any of the 18 HIPAA-defined personal identifiers, such as addresses or dates. When linked with the MTF dataset, locations smaller than state (ZIP code) are added, thereby creating a dataset that no longer follows the HIPAA Safe Harbor standard. Similarly, in Use Case 3, the PEDSnet policies specify that individual level data must be de-identified using the Safe Harbor method of de-identification before sharing data. Linking PEDSnet data with EPA AQS data, requires using geographic identifiers from both datasets to fulfill the linkage to the other EPA AQS variables. For both of these Use Cases, a decision would need to be made regarding whether the geographic information should be retained, removed, or transformed in the linked dataset prior to sharing. Additionally, the prior PPRL report noted that data linkage inherently increases re-identification risk, which varies widely depending on the content of the datasets to be linked and the pattern of overlap of the variables contained in the resulting linked dataset. Linkage implementations often apply additional controls to account for that added risk in the linked dataset, such as modifying variables that are shared or performing a disclosure review that may determine further de-identification is needed before sharing the linked dataset.

The prior report also observed that each record linkage implementation involves agreeing on one de-identification standard for all datasets linked. For each use case in this project, this decision would be needed prior to sharing linked data, to address discrepancies in de-identification standards for individual datasets and possibly mitigate any added risk of the linked dataset. Furthermore, for linked datasets that incorporate data based on geographical information (e.g., the EPA dataset) but require removing actual location data from the dataset (e.g., to meet HIPAA Safe Harbor), stakeholders will need to consider whether the geo-linked data enables users to deduce location even if location information is stripped from the linked dataset.

### *3-2: Present Considerations for Developing a Generalizable Data Governance Metadata Schema*

The wide-spread adoption of metadata is revolutionizing the way that data is used. Metadata provides context for data and makes data more findable and persistently meaningful for reuse. Metadata in the context of computer systems was first mentioned when, in 1967, Stuart McIntosh and David Griffel of MIT noted the need for digital "meta language" (7). Today, metadata is used to identify and describe data of all types, across many disciplines, from planetary science (NASA Planetary Data System (8)), to agriculture (Online Farm Trials Database (9)), and culture and folklore (e.g., metadata schema for folk takes in the Mekong River Basin (10)). This project aims to inform the creation of structured metadata of data governance information to inform appropriate linkage and use of HHS and other federally funded datasets. The findings from this governance assessment and data linkage determination confirm that developing a generalizable data governance metadata schema will support, for the benefit of the public, re-use of valuable data assets to answer important scientific questions that can only be addressed with the linkage of data from multiple resources. Some considerations to this end are included here.

*Public sharing of data governance information will facilitate the ability to create linked datasets*

Obtaining data governance information is critical to implementing record linkage but requires significant effort. The project team retrieved and evaluated more than sixty sources of governance information in the course of this data linkage determination. All eleven of the data sources provided some form of governance information on the resource website, however for nine of the eleven datasets the project team consulted at least three sources of governance information. To collect the necessary governance information on more than half of the datasets (6 of 11), the team had to rely on communication with data source stakeholders (often PIs, Project Directors and Technical Directors) to provide information or confirm the completeness and validity of information collected from other sources. The breadth of data sources the team consulted is considerable, including: data source websites; dataset webpages; consent documents; protocols; IRB, privacy board, and ethics review board approval documents; Data Use Agreements and Data Access Agreements; brochures; presentations; and in the case of NHANES, an internal document provided by the stakeholder that shed light on data linkage. The project team often followed a trail of web pages, for example, from the data source website to the website describing laws like Section 306 of the Public Health Service Act and the Food Quality Protection Act of 1996. Although in some cases the project team was able to find documents like consent forms and templates for Data Use Agreements online, the availability of these documents was not consistent.

The project team relied heavily on data source and dataset stakeholders to provide information for this assessment. The team was introduced to the stakeholders by the Project Governance Team and through ODSS contacts. The team was advantaged to have the support of the senior HHS leaders comprising the PGT and ODSS for connection to individuals who were responsive and willing to discuss the data source and governance of the datasets. Researchers attempting data linkage are unlikely to have contacts like these, leading to a significant gap in the governance information profile.

Despite the experienced project governance team's best effort, and after an exhaustive search for all discoverable sources of governance information, there remain gaps in governance information. For example, the scope of data linkage for CDC COVID-19 data could not be found and remains undetermined. Appendix D provides the full landscape of data governance information capture and interpretation for all data sources and datasets. These gaps as well as uncertainty around interpretation of disparate data governance information could be eliminated by a standardized, well-defined data governance metadata model and the mandate that the metadata be collected in a standard manner and made publicly and readily available in a predictable and easy-to-find location alongside shared datasets.

Even when governance information was publicly available, the project team were typically required to contact data source experts to confirm authorizations and controls, filling gaps when possible so the team could fully understand the data governance. This was the case for nine of the eleven datasets included in this linkage assessment. A primary point of contact for a data source agency or program would assist users with general questions (e.g., how the linked dataset can be accessed and used), clarifications, and resolution of potential conflicts among datasets to be linked, for instance regarding the scope of linkage, access, and use of the linked data. As the Findings show, clarifying publicly available information, assisting with interpretation of those information, and negotiating conflicts in governance for data linkage and use often require human intervention. Providing a point of contact to assist researchers may be logistically challenging, however the implementation of publicly available, standardized dataset governance and linkage metadata would reduce the need for staff to support researchers in this process.

*Linkage governance information should be explicit, complete, and easily discoverable in data source web pages or publicly accessible documentation*

To be complete, linkage governance information for individual datasets should ideally describe the following:

- Governance origins, including consent, assent, IRB and privacy board approval, data originator agreements, data sharing/use/access agreements, and applicable laws, regulations, and policies
- Authorization for linkage (I.e., whether a dataset can be linked with another dataset)
- Scope of linkage (I.e., with what other dataset(s) or kind of dataset(s) the data may or may not be linked)
- Limitations and controls (I.e., how the linked dataset that includes this dataset must be shared, accessed and used)

The data structure created by the project team for this report can serve as a foundation for explicitly and completely describing these governance elements in an accessible manner. The tables in Appendix D contain detailed governance information contributing to linkage determination for all datasets, as well as the provenance of the data governance origins. For each step of the data lifecycle (collection, linkage, sharing, access, and use), the information is broken down into authorizations (e.g., consents, data originator agreements), and applicable laws, regulations, and policies, and finally the governance based on the collective origin of the governance (synthesis of all authorizations and laws, policies, and regulations). In addition, information on PII availability and access and description of any prior data linkages are included to help clarify how data linkage may be implemented. Appendix D is in part designed to show the progression from collection to interpretation of the governance information. The interpretation of the governance information is the step where these governance elements begin to take shape as standardized metadata. For dataset linkage to be a readily usable tool for researchers, these metadata must be complete, as clear as possible through standardized definition, and publicly accessible.

The project team also noted that the various stages of the data lifecycle and associated governance requirements involve or impact multiple stakeholders. This was highlighted by a lack of consistency in how points of contact interviewed interpreted the definition of “sharing”. Because sharing data often involves multiple parties with different roles and responsibilities, the framework needed to accommodate data originators who contribute data, data repositories who accept the data and then make it accessible to secondary users, and the secondary users accessing and using the data – all of whom contribute to data sharing activities. Importantly, all of these parties would also play a role in implementing record linkage and in sharing, accessing, and using linked data. Additional parties may be engaged to facilitate record linkage, for example, honest brokers who facilitate participant-level matching. While information about the responsible party was

not structured in a standard manner in this assessment, further consideration should be given to whether a structured schema could benefit from capturing the roles of various parties involved in implementing data governance across the data lifecycle.

*When linkage governance is not explicitly specified in governance documentation, it may be inferred or determined by an appropriate authority. Once the linkage governance is identified or decided, it should be added to the publicly available linkage governance information for the dataset for users to access.*

To determine the governance of many of the datasets, the project team distilled and interpreted language from the data governance origin materials. For example, for NCCR, the project team researched the statutes of the individual states contributing to the registry, along with NAACR and SEER policies, to understand the scope of linkage. The project team was able to determine that the Louisiana Tumor Registry allows data to be linked with external databases to improve the accuracy and completeness of follow-up data or for research. Idaho Code Title 57, Chapter 17, 57-1703 (6) defines the scope of data collection to include all cancers and reportable benign tumors diagnosed and/or treated within the state of Idaho by hospitals or other facilities providing screening, diagnostic or therapeutic services to patients with respect to cancer, and from physicians, surgeons, and all other health care providers diagnosing or providing treatment for cancer patients. It further permits identification and report of cases using data linkages with death records, statewide cancer registries, and other potential sources. Given that NCCR includes data from Cancer in North America (CiNA) North American Association of Central Cancer Registries (NAACCR) and the NCI's Surveillance, Epidemiology and End Results (SEER) Registries, a user would have to examine the governance origins for the 23 contributing registries, which include: California, Connecticut, Florida, Georgia, Hawaii, Idaho, Illinois, Iowa, Kentucky, Louisiana, Massachusetts, New Jersey, New Mexico, New York, Ohio, Pennsylvania, Seattle (Puget Sound), Tennessee, Texas, Utah, Wisconsin. These 23 NCCR registries represent 66% of all U.S. children, adolescents, and young adults ages 0-39 based on 2018 U.S. population census.

Reviewing and documenting the scope of linkage for each of the NCCR participating states is burdensome for an individual researcher who may also lack experience with data governance assessment and linkage determination. Preferably, an agency, program or investigator offering data for linkage, sharing, and/or use would assemble the governance information and publish it for potential data users and other entities to reference. This process would be most efficient and accurate if an established

governance metadata schema were available to guide the development and presentation of this information.

Additionally, linkage often requires additional decisions to be made regarding the linked dataset beyond the rules derived from individual datasets. As described above, decisions may be needed regarding conflicts of location of access, discrepancies between de-identification standards, and any additional controls that are warranted to mitigate risk introduced by the linkage. These decisions may be specific to a given record linkage implementation, but it may be useful to share such decisions to communicate appropriate use of the linked data as well as inform future linkage implementations involving that individual dataset or similar datasets.

*Standardized and consistently available governance metadata would simplify the complex process of determining whether two or more datasets can be linked and promote adherence to governance requirements.*

The Findings in this report describe a data landscape in which some critical data governance information is publicly available, while much is locked in organizational documentation that is not readily accessible. In the context of three real-world use cases, the project team used many techniques and their extensive expertise to document the authorizations and policies that impact linkage of eleven data sets and the limitations and controls inherited by the linked dataset. These activities demonstrate the potentially time-consuming process an investigator or other stakeholder would undertake to gather the governance information required for linking datasets.

The team's linkage determination produced a complex series of pairwise evaluations leading to an all-dataset linkage outcome, with authorizations and applicable controls and limitations that apply to the linked datasets noted. The process as it exists today requires a researcher or other stakeholders to interpret the language used in each of a series of documents like consent forms, IRB determinations, federal legal statutes, and data use agreements and synthesize the information to establish a series of data linkage and use conditions. The interpretation of complex provisions from each document can be nuanced. The integration of these rules with those of a dataset to be linked requires meticulous attention to detail, ability to assimilate data governance concepts, and ability to recognize incomplete information and conflicts in governance that may require resolution prior to data linkage. In many cases the process is burdensome and time consuming, and the governance information is at risk of misinterpretation.

The process and outcome of this project makes evident the need for the development and operationalization of standardized data governance metadata and related decision-making procedures to facilitate research made possible or more powerful by dataset

linkage. In fact, the documentation of the assessment, reflecting the assessment process, breaks the governance information landscape down into key areas for metadata development, for instance annotation of consent, description of state laws, and extraction of governance information from Data Use Agreements. The relevant clauses and provisions of these documents are often convoluted, text-heavy, explained in heterogeneous ways, using inconsistent terminologies and definitions, and located in different documents or variable parts of a given document.

A clear, well-defined governance metadata standard would facilitate the process of governance metadata collection, sharing, and human interpretation (in addition to machine-readability), which in turn promotes adherence throughout the data lifecycle, including linkage and use of linked data. A metadata standard would also facilitate harmonization of governance information that may already exist across the data ecosystem. Making this information readily discoverable across data repositories and datasets will relieve the researcher or other stakeholders of the burden of searching for data governance information from many sources, remove or reduce uncertainty from the process of data linkage, and facilitate accurate interpretation of governance requirements. Furthermore, a standardized approach would encourage the use of shared data for innovation and application to complex research questions that existing data may address.

*Adoption of a generalizable, scalable, and machine-readable linkage governance metadata schema that can be broadly and effectively employed by future linkage decision makers and data users will require a concerted effort and coordination across federal and other health agencies that generate large volumes of clinical, administrative, survey and other types of data.*

Defining metadata and data standards is a time- and labor-intensive process. By way of process, typically, a common dictionary is formed, with the development of extensions to accommodate the annotation of discipline-specific or project-specific data. This effort is carried out by a team with representative experts from participating subject areas. For example, to develop imaging metadata for the National Cancer Institute's Human Tumor Atlas Network, imaging experts from each of the institutions generating images worked cooperatively through an iterative process to define and document standardized imaging metadata, including annotations for governance and provenance. Experts with each of the various imaging types, for instance pathology based diagnostic images and multiplexed immunoflourescence (MIF) imaging, were critical to the expedient development of relevant metadata. The generation and contribution of these metadata will be mandated, upon sharing data, allowing the implementation of the NCI Cancer Research Data Commons, which links datasets with tools and algorithms for dataset use

and reuse. These metadata support all computational search capabilities for the data commons areas. Likewise, the development of linkage governance metadata that can be broadly and effectively applied to biomedical datasets will require a cooperative and concerted effort across federal health agencies.

Developing generalizable, scalable, machine-readable governance metadata is a focus for not only OS-PCORTF and NIH, as described in the introduction of this report, but also for many scientific disciplines. For example, the NASA Data Strategy, published in 2021, includes objectives shared by HHS and other federal agencies, including: ensure data is Findable, Accessible, Interoperable, Reusable, Understandable, Secure, and Trustworthy; identify data stewards for all NASA data and systems; create a centralized data governance framework to manage core metadata on all systems; create reusable data assets. There is opportunity to create a national standard generalizable, scalable governance metadata schema that will benefit research across the federal data ecosystem. Adoption of governance metadata standards, once developed, will require education to create a data science-knowledgeable and capable community who will understand the need for and benefit of the approach. Furthermore, the agencies must develop and commit to a standardized approach for ensuring that these metadata and data standards are used across the data lifecycle consistently, starting with collection, through linking and sharing, to accessing and using the datasets.

# Conclusion

This assessment further solidifies the argument for a creating a metadata schema that specifies a standard for sharing data governance information alongside biomedical, administrative, or other datasets to inform future record linkage opportunities for patient centered outcomes research. Based on this dataset governance assessment, the report provides considerations for the development and implementation of a data governance metadata schema, including:

- Publicly sharing the data governance information specified by the schema in a predictable and easy-to-find location will facilitate the ability to create linked datasets
- Publicly shared data governance information, and the associated schema, should:
    - Explicitly describe whether linkage is permissible for a given dataset and, if so, include general guidance for what types of linkages are allowed or prohibited, and what rules and controls the linked data would inherit from the individual dataset
    - Incorporate the provenance of data governance origins including authorizations for data collection, linking, sharing, access, and use as well as applicable laws, regulations, and policies
    - Capture the roles and responsibilities of the multiple stakeholders involved in implementing data governance across the data lifecycle
    - Incorporate information regarding decisions made for previous and new linkages involving a given dataset to communicate appropriate linkage of the data and to inform future linkage involving the same dataset; this information may streamline decision making when linkage governance is not explicitly specified by any dataset governance source.
- The schema should describe data governance in a standard way to facilitate human interpretation and machine-readability, which in turn promotes adherence
- A concerted effort is required to encourage adoption of the schema across federal and other health agencies that generate datasets that could be linked and used by researchers

The framework created throughout this assessment serves as a critical foundation for developing the structure for a future data governance metadata schema. Additionally, in parallel to the development of a standardized schema, this framework could be used to facilitate data governance analyses for future record linkage implementations by helping stakeholders determine feasibility of linkage, surface rules and controls that must be respected for linked datasets, and reveal additional decisions that need to be made regarding how the linked data are shared (e.g., implementing additional controls to mitigate re-identification risk). This work will ultimately promote more thoughtful and appropriate record linkage efforts, build community trust, and yield more discoveries from patient centered outcomes research.

# References


1. **National Institute of Child Health and Human Development Office of Data Science and Sharing with Booz Allen Hamilton.** Privacy Preserving Record Linkage (PPRL) for Pediatric COVID-19 Studies. [Online] September 2022. https://www.nichd.nih.gov/sites/default/files/inline-files/NICHD_ODSS_PPRL_for_Pediatric_COVID-19_Studies_Public_Final_Report_508.pdf.

2. **U.S. Department of Health and Human Services.** [Online] https://www.hhs.gov/about/index.html.

3. **ASPE Office of the Secretary Patient-Centered Outcomes Research Trust Fund.** [Online] September 2022. https://aspe.hhs.gov/sites/default/files/documents/b363671a6256c6b7f26dec4990c2506a/aspe-os-pcortf-2020-2029-strategic-plan.pdf.

4. **National Institutes of Health Office of Data Science Strategy.** [Online] June 2018. https://datascience.nih.gov/sites/default/files/NIH_Strategic_Plan_for_Data_Science_Final_508.pdf.

5. *The FAIR Guiding Principles for scientific data management and stewardship.* **Wilkinson, Mark D, Dumontier, Michel and Aalbersberg, I Jsbrand Jan.** 160018, 2016, Scientific Data, Vol. 3.

6. **National COVID Cohort Collaborative (N3C).** N3C Privacy-Preserving Record Linkage. [Online] https://covid.cd2h.org/PPRL/.

7. **Griffel, David and McIntosh, Stuart.** ADMINS: A Progress Report. [Online] January 1967. http://dspace.mit.edu/bitstream/handle/1721.1/82974/09487802.pdf?sequence=1.

8. *PDS4: A model-driven planetary science data architecture for long-term preservation.* **Hughes, John S, et al.** Chicago : s.n., 2014. IEEE 30th International Conference on Data Engineering Workshops. pp. 134-141.

9. *Using agricultural metadata: a novel investigation of trends in sowing date in on-farm research trials using the Online Farm Trials database [version 2; peer review: 1 approved, 2 approved with reservations].* **Walters, Judi, Light, Kate and Robinson, Nathan.** 1305, November 2020, F1000 Research, Vol. 9.

10. *Metadata Schema for Folktales in the Mekong River Basin.* **Kwiecien, Kanyarat, et al.** 4, 2021, Informatics, Vol. 8.

# Appendix A: Project Governance Team

| Participant Name | Participant Title |
|---|---|
| Alison Cernich, Ph.D. | Deputy Director, *Eunice Kennedy Shriver* National Institute of Child Health and Development (NICHD) |
| Chun-Ju (Janey) Hsiao, Ph.D. | Deputy Data Officer, Office of Data Resources and Analytics (ODRA), National Institute on Aging (NIA) |
| Betsy Hsu, Ph.D. | Branch Chief, Surveillance Informatics Branch, Surveillance Research Program National Cancer Institute (NCI) |
| Vivian Ota-Wang, Ph.D. | Policy and Ethics Lead, Office of Data Science Strategy at Office of the Director NIH |
| Christine Fortunato, Ph.D. | Team Leader for Child Welfare Research, Office of Planning, Research and Evaluation, Administration for Children and Families (ACF) |
| Jennifer Haight, M.A. | Director, Division of Performance Measurement and Improvement Administration on Children Youth and Families (ACYF) |
| Elizabeth (Liz) Ginexi, Ph.D. | Program Director, Clinical Research in Complementary and Integrative Health Branch at the National Center for Complementary and Integrative Health (NCCIH) |
| Joshua Fessel, M.D., Ph.D. | Senior Clinical Advisor, Division of Clinical Innovation National Center for Advancing Translational Sciences (NCATS) |

# Appendix B: Glossary

| Term | Definition |
|---|---|
| Accessibility (data) | To be accessible, metadata and data should be readable by humans and machines, and it must reside in a trusted repository (NIH NLM) |
| Aggregate data | Summary statistics compiled from multiple sources of individual-level data. (NIH) |
| Authorization | Permission provided by a law/regulation/policy or an authority or an agreement to perform data lifecycle activities, including collecting, linking, sharing, accessing, or using the data |
| Common data model | A common data model (CDM) standardizes the definition, format and model content of data across participating data partners so that standardized applications, tools and methods can be applied. (PCORnet) |
| Controlled access | Application and eligibility requirements need to be met and approved (e.g., by a data access committee) to gain access. (NIH)<br><br>"Controlled access" and "access controls" refer to measures such as requiring data requesters to verify their identity and the appropriateness of their proposed research use to access protected data. (NIH) |
| Controls | Processes established to ensure compliance with governance for data sharing, access, and use (e.g., user must access data in a physical enclave, user must sign data use agreement, user must receive data access committee approval, etc.) |
| Data Access | Acquiring data from a data repository or other data sharing system |
| Data Collection | Obtaining data from participants for research, clinical, or administrative purposes |
| Database/Data repository | Virtual data storage that stores, organizes, and validates data, and makes the data accessible for use by others |
| Data linkage/record linkage | Combining information from a variety of data sources for the same individual. (AHRQ) In the context of this report, synonymous with record linkage. |
| Data originator/ contributor/submitter | Institutions/organizations/researchers that collect data from patients or study participants or that collect administrative |

| Term | Definition |
|---|---|
| | data; they may also be the party to submit the data to a repository for sharing |
| Data science | Interdisciplinary field of inquiry in which quantitative and analytical approaches, processes, and systems are developed and used to extract knowledge and insights from increasingly large and/or complex sets of data |
| Dataset | Collection of related sets of information composed of separate elements that can be manipulated computationally as a unit |
| Data Sharing | Making data available to the broader data user community, for example, by submitting the data to a data repository for dissemination |
| Data standards | Documented agreements on representation, format, definition, structuring, tagging, transmission, manipulation, use, and management of data |
| Data steward | A formal position or an assigned accountability with responsibility for the following areas: (HHS)<br><br>Adherence to an appropriately determined set of privacy and confidentiality principles and practices<br><br>Appropriate use of information from the standpoint of good statistical practices (such as by not implying cause and effect when the data only point to correlation)<br><br>Limits on use, disclosure and retention<br><br>Identification of the purpose for a specific use of the data<br><br>Application of “minimum necessary” principles<br><br>Verification of receipt by the correct recipient, wherever possible<br><br>Data de-identification (HIPAA-defined and beyond)<br><br>Data quality, including integrity, accuracy, timeliness, and completeness (NCVHS) |
| Data Use | Working with data for secondary research or other analytical purposes |

| Term | Definition |
|---|---|
| Data use agreement (DUA) | A document which establishes who is permitted to use and receive data, and the permitted uses and disclosures of such information by the recipient. (HHS modified) |
| Data user (or secondary data user) | A person who accesses and uses data collected by another party for new research purposes |
| Deductive disclosure | Disclosure is revealing information that relates to the identity of a data subject, or some sensitive information about a data subject through the release of either tables or microdata. (HHS) |
| De-duplication | The process of removing redundant patient records from a database. (CDC) |
| De-identification | De-identified patient data is patient information that has had personally identifiable information (PII; e.g. a person's name, email address, or social security number), including protected health information (PHI; e.g. medical history, test results, and insurance information) removed. This is normally performed when sharing the data from a registry or clinical study to prevent a participant from being directly or indirectly identified. (NIH) |
| Electronic health records (EHRs) | EHRs are electronic versions of the paper charts in your doctor's or other health care provider's office. An EHR may include your medical history, notes, and other information about your health including your symptoms, diagnoses, medications, lab results, vital signs, immunizations, and reports from diagnostic tests such as x-rays. (HHS) |
| Enclave | A data enclave is a secure network through which confidential data, such as identifiable information from census data, can be stored and disseminated. In a virtual data enclave a researcher can access the data from their own computer but cannot download or remove it from the remote server. Higher security data can be accessed through a physical data enclave where a researcher is required to access the data from a monitored room where the data is stored on non-network computers. (NNLM)<br><br>[*see also data access model*] |
| Entity Resolution | Process of joining or matching records from one data source with another that describes the same entity. (Census) |

| Term | Definition |
|---|---|
| | In PPRL, hash codes/tokens are used to match individual records without using PII/PHI. (N3C) |
| FAIR | Findable, Accessible, Interoperable, Reusable |
| FAIR Guiding Principles | A set of guiding principles for scientific data management and stewardship that describe distinct considerations for contemporary data publishing environments with respect to supporting both manual and automated deposition, exploration, sharing and reuse. |
| Findable (data) | For data to be findable there must be sufficient metadata, a unique and persistent identifier, and the data must be registered and indexed in a searchable resource (NIH NLM) |
| Governance | Governance or data governance, as defined in this Report, comprises of the policies, limitations, processes, and controls that address ethics, privacy protections, compliance, risk management, or other requirements for a given record linkage implementation across the data lifecycle. |
| HIPAA Privacy Rule | The Standards for Privacy of Individually Identifiable Health Information are codified in 45 C.F.R. Parts 160 and 164 promulgated by the U.S. Department of Health and Human Services under the Health Insurance Portability and Accountability Act (HIPAA) of 1996. The HIPAA Privacy Rule establishes national standards to protect individuals' medical records and other individually identifiable health information (collectively defined as "protected health information") and applies to health plans, health care clearinghouses, and those health care providers that conduct certain health care transactions electronically. The Rule requires appropriate safeguards to protect the privacy of protected health information and sets limits and conditions on the uses and disclosures that may be made of such information without an individual's authorization. The Rule also gives individuals rights over their protected health information, including rights to examine and obtain a copy of their health records, to direct a covered entity to transmit to a third party an electronic copy of their protected health information in an electronic health record, and to request corrections. (HHS Health Information Privacy) |
| Honest broker | A party that holds de-identified tokens ("hashes") and operates a service that matches tokens generated across disparate |

| Term | Definition |
|---|---|
| | datasets to formulate a single Match ID for a specific use case. (N3C) |
| Institutional Review Board (IRB) | An institutional review board (IRB) is the institutional entity charged with providing ethical and regulatory oversight of research involving human subjects, typically at the site of the research study. (NIH)<br><br>An Institutional Review Board is an appropriately constituted group that has been formally designated to review and monitor biomedical research involving human subjects. An IRB has the authority to approve, require modifications in (to secure approval), or disapprove research. This group review serves an important role in the protection of the rights and welfare of human research subjects. (FDA) |
| Interoperability | According to section 4003 of the 21st Century Cures Act, the term 'interoperability,' with respect to health information technology, means such health information technology that—"(A) enables the secure exchange of electronic health information with, and use of electronic health information from, other health information technology without special effort on the part of the user; "(B) allows for complete access, exchange, and use of all electronically accessible health information for authorized use under applicable State or Federal law; and "(C) does not constitute information blocking as defined in section 3022(a)." (HIT) |
| Interoperability (data) in computer systems | The ability to exchange and make use of information from various sources and of different types (NIH ODSS)<br><br>The ability of data or tools from non-cooperating resources to integrate or work together with minimal effort (FAIR)<br><br>Data must share a common structure, and metadata must use recognized, formal terminologies for description (NIH NLM) |
| Letter of determination (LOD) | A letter of determination documents an IRB decision on the status of research. (HHS) |
| Limitations | Restrictions on data linkage and use (e.g., dataset must only be linked with other disease-relevant data, dataset must be used in a physical enclave, etc.) |

| Term | Definition |
|---|---|
| Machine learning | A field of computer science that gives computers the ability to learn without being explicitly programmed by humans |
| Metadata | Information describing the characteristics of data including, for example, structural metadata describing data structures (e.g., data format, syntax, and semantics) and descriptive metadata describing data contents (e.g., information security labels). (NIST) |
| Ontology | A set of terms or concepts defining the properties or identities of subjects (e.g., genes, proteins, conditions) and relationships between them; similar to a standardized vocabulary |
| Open Access | Data within this category presents minimal risk of participant identification. Access to these data does not require user certification, and researchers may explore data content without restriction. (NCI)<br><br>No access restrictions or registration required to access (NIH) [*see also data access model*] |
| Patient Identifier | Unique data used to represent a person's identity and associated attributes. (NIST) |
| Personally identifiable information (PII) | Any information that can be used to distinguish or trace an individual's identity, either alone or when combined with other information that is linked or linkable to a specific individual. (NIST and CODI) |
| Privacy preserving record linkage (PPRL) | A technique identifying and linking records that correspond to the same entity across several data sources held by different parties without revealing any sensitive information about these entities. (UK Office for National Statistics) |
| Protected Health Information (PHI) | Individually identifiable health information that is transmitted or maintained in any form or medium (electronic, oral, or paper) by a covered entity or its business associates, excluding certain educational and employment records. (NIH) |
| Provenance | The documented trail that accounts for the origin of a piece of data and where it has moved from to where it is presently (NIH NLM) |
| Reusable (data) | Data and collections must have clear usage licenses and clear provenance, and must meet relevant community standards for the domain (NIH NLM) |
| Software | Programs and other operating information used by a computer |

# Appendix C: Acronyms and Initialisms

| Abbreviation | Term |
|---|---|
| ACF | Administration for Children and Families |
| ACYF | Administration on Children Youth and Families |
| AFCARS | The Adoption and Foster Care Analysis and Reporting System |
| AQS | Air Quality System |
| CBSA | Core-Based Statistical Area |
| CCDI | NCI Childhood Cancer Data Initiative |
| CDC | Centers for Disease Control and Prevention |
| CDM | Common Data Model |
| CHIP | Children's Health Insurance Program |
| CiNA | Cancer in North America |
| CIPSEA | Confidential Information Protection and Statistical Efficiency Act |
| CMS | Centers for Medicare and Medicaid Services |
| CODI | Childhood Obesity Data Initiative |
| DAA | Data Access Agreement |
| DAC | Data Access Committee |
| DAF | Designated Agent Form |
| DHANES | Division of Health and Nutrition Examination Surveys |
| DUA | Data Use Agreement |
| EHR | Electronic Health Record |
| EPA | Environmental Protection Agency |
| FAIR | Findable, Accessible, Interoperable and Reusable |
| FISMA | Federal Information Security Management Act |
| HIPAA | Health Insurance Portability and Accountability Act |
| HHS | Health and Human Services |
| ICPSR | Inter-university Consortium for Political and Social Research |
| IRB | Institutional Review Board |
| LOD | Letter of Determination |

| Abbreviation | Term |
|---|---|
| MIS-C | Multisystem Inflammatory Syndrome in Children |
| MTF | Monitoring the Future |
| N3C | National COVID Cohort Collaborative |
| NAACR | North American Association of Central Cancer Registries |
| NAHDAP | National Addiction & HIV Data Archive Program |
| NCATS | National Center for Advancing Translational Sciences |
| NCCIH | National Center for Complementary and Integrative Health |
| NCCR | National Childhood Cancer Registry |
| NCHS | National Center for Health Statistics |
| NCI | National Cancer Institute |
| NCVSH | National Committee on Vital Health and Statistics |
| NDA | National Institute of Mental Health Data Archive |
| NDACAN | National Data Archive on Child Abuse and Neglect |
| NHANES | National Health and Nutrition Examination Survey |
| NIA | National Institute on Aging |
| NICHD | National Institute of Child Health and Human Development |
| NIH | National Institute of Health |
| NIST | National Institute of Standards and Technology |
| NLM | National Library of Medicine |
| NPCR | National Program of Cancer Registries (NPCR) |
| NSDUH | National Survey on Drug Use and Health |
| ODRA | Office of Data Resources and Analytics |
| ODSS | Office of Data Science and Sharing |
| OMB | Office and Management and Budget |
| OS-PCORTF | Office of Secretary Patient-Centered Outcomes Research Trust Fund |
| PCOR | Patient-Centered Outcomes Research |
| PCORI | Patient Centered Outcomes Research Institute |
| PCORnet | The National Patient-Centered Clinical Research Network |
| PGT | Project Governance Team |
| PHI | Personal Health Information |
| PII | Personally Identifiable Information |

| Abbreviation | Term |
|---|---|
| PI | Principal Investigator |
| POC | Point of Contact |
| PPRL | Privacy Preserving Record Linkage |
| RADx | Rapid Acceleration of Diagnostics |
| RDC | Research Data Center |
| RDUA | Restricted Data Use Agreement |
| SAMHSA | Substance Abuse and Mental Health Services Administration |
| SDOH | Social Determinants of Health |
| SEER | Surveillance, Epidemiology, and End Results |
| T-MSIS | Transformed Medicaid Statistical Information System |
| VDE | Virtual Data Enclave |