National Institute of Diabetes and Digestive and Kidney Diseases

Central Repository

NIDDK-CR Resources for Research

Data Science and Meet the Expert Webinar Series

REPOSITORY
TECHNOLOGY ADVANCEMENT
BIOSTATISTICS
CLINICAL STUDIES
REGULATORY
Office of Clinical Research Support
NIH National Institute of Diabetes and Digestive and Kidney Diseases

July 31, 2025

NIH National Institute of Diabetes and Digestive and Kidney Diseases

Central Repository

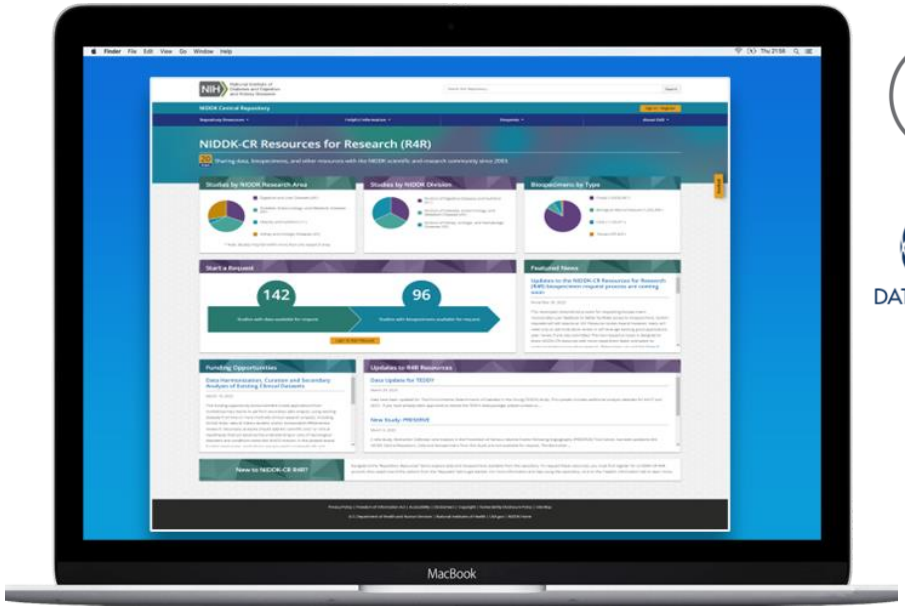
NIDDK Central Repository Overview

Mission

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

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

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


NIDDK-CR Resources for Research (R4R)
Studies by NIDDK Research Area
Studies by NIDDK Division
Biospecimens by Type
Start a Request
142
96
Featured News
Funding Opportunities
Data Harmonization, Curation and Secondary Analysis of Existing Clinical Datasets
Updates to R4R Resources
Data Update for TEDDY
New Study: PRESERVE
New to NIDDK-CR R4R?
MacBook

CORE TRUST SEAL

WORLD DATA SYSTEM

Imaging Data Files
15.8 M

Clinical Datasets
>8,400
from 189 clinical studies

Biospecimens
>16 M

Registered Users
6,976

Weekly Users
>5,000

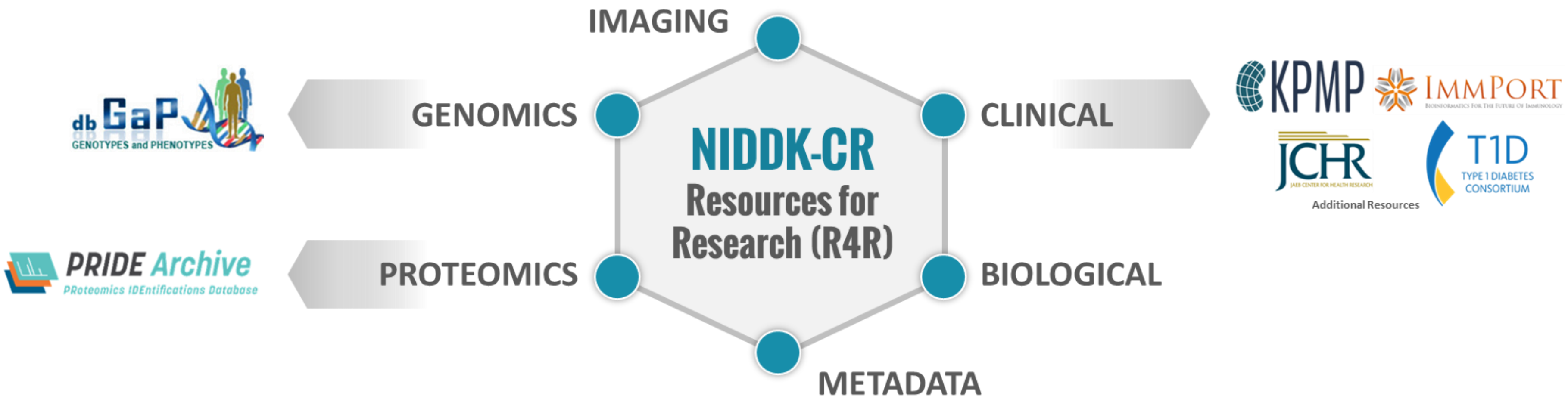
Public Releases
>875

NIH National Institute of Diabetes and Digestive and Kidney Diseases

Central Repository

NIDDK Data Sharing Ecosystem

The NIDDK-CR is a part of the broader NIH-funded biomedical data ecosystem and plays a key role in NIH's FAIRness and TRUSTworthiness goals. The NIDDK-CR houses a broad range of data types for secondary research, provides access to biospecimens, and direct links to other repositories with additional resources such as genomics data.


IMAGING
dbGaP GENOTYPES and PHENOTYPES
GENOMICS
CLINICAL
KPMP
IMMPORT
JCHR
T1D TYPE 1 DIABETES CONSORTIUM
Additional Resources
NIDDK-CR
Resources for Research (R4R)
PRIDE Archive
PRoteomics IDEntifications Database
PROTEOMICS
BIOLOGICAL
METADATA

FAIRsharing.org standards, databases, policies

DataCite FIND, ACCESS, AND REUSE DATA

re3data.org REGISTRY OF RESEARCH DATA REPOSITORIES

DK net

Google Dataset Search / Schema.org

NIH U.S. National Library of Medicine ClinicalTrials.gov

Vivli CENTER FOR GLOBAL CLINICAL RESEARCH DATA

PLANNING PHASE

figshare

NIH HEAL INITIATIVE

NIH National Institute of Diabetes and Digestive and Kidney Diseases

Central Repository

Future Functionality: Analytics Workbench

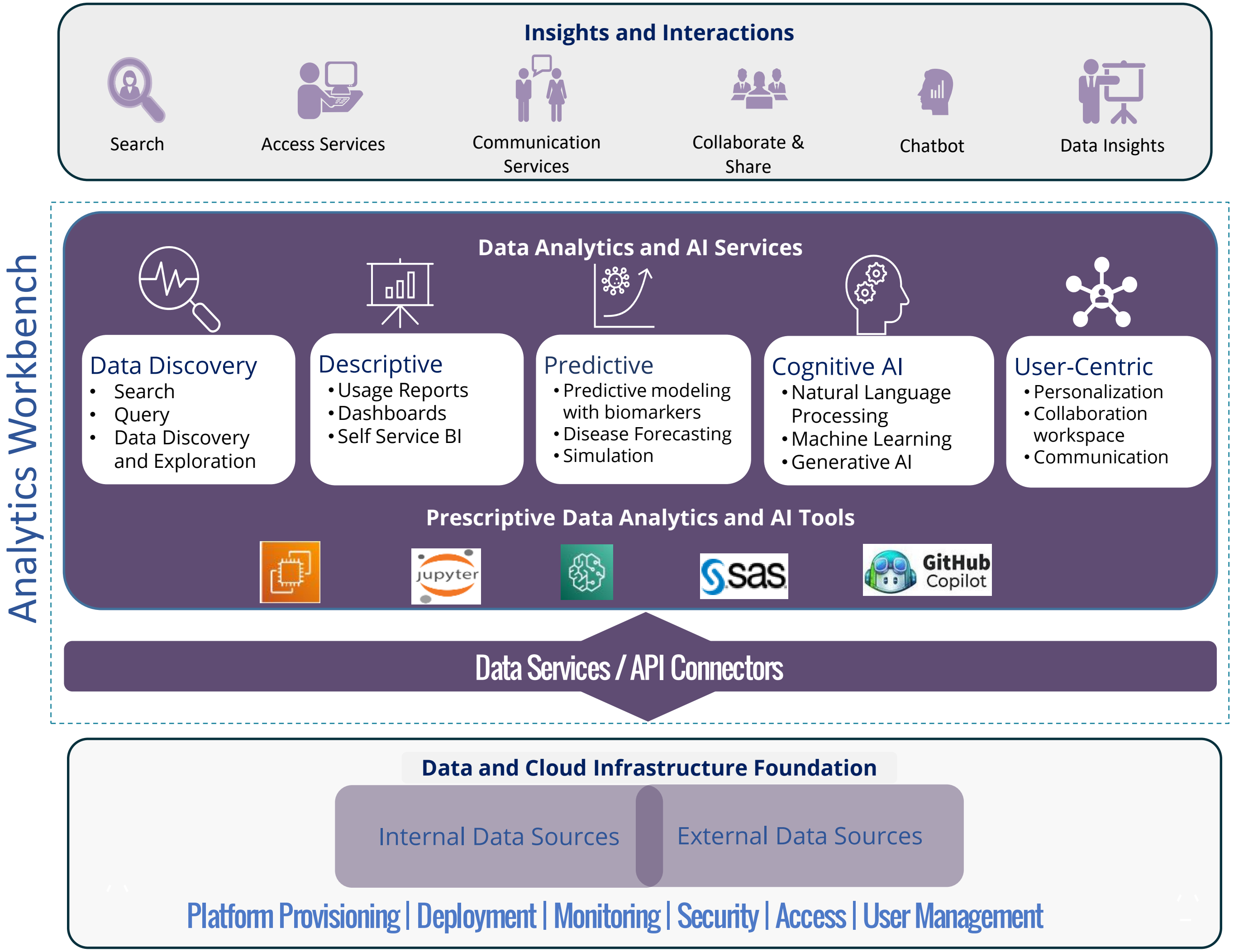
***Streamlining** end-to-end data science lifecycle and discovery of data-driven biomedical insights.*

Innovation and ease of use

A cloud-based analytics environment where researchers and data scientists can access a suite of integrated analytics tools and cloud computing resources to participate in data challenges and AI innovation.

Expected Benefits of Analytics Workbench:

- Promote Collaboration
- Support AI Innovation
- Minimize Data Movement
- Improve User Experience
- Discover Data Insights
- Advance NIDDK Research Mission



Analytics Workbench

Insights and Interactions

Search | Access Services | Communication Services | Collaborate & Share | Chatbot | Data Insights

Data Analytics and AI Services

Data Discovery
- Search
- Query
- Data Discovery and Exploration

Descriptive
- Usage Reports
- Dashboards
- Self Service BI

Predictive
- Predictive modeling with biomarkers
- Disease Forecasting
- Simulation

Cognitive AI
- Natural Language Processing
- Machine Learning
- Generative AI

User-Centric
- Personalization
- Collaboration workspace
- Communication

Prescriptive Data Analytics and AI Tools

jupyter | sas | GitHub Copilot

Data Services / API Connectors

Data and Cloud Infrastructure Foundation

Internal Data Sources | External Data Sources

Platform Provisioning | Deployment | Monitoring | Security | Access | User Management

NIH National Institute of Diabetes and Digestive and Kidney Diseases

Central Repository

NIDDK-CR Data Science Centric Challenge Series

Goals of NIDDK-CR Data-science centric challenge series

- Develop tools, approaches, models and/or methods to increase data interoperability and usability for artificial intelligence (AI) and machine learning (ML) applications
- Augment and enhance existing data for future secondary research, including data-driven discovery by AI/ML researchers
- Discover innovative approaches to enhance the utility of datasets for AI/ML applications

Visit our website for more information on our data-centric movement and to learn more about our past data-challenges

National Institute of Diabetes and Digestive and Kidney Diseases
Central Repository

Secondary Data Science and Meet the Expert Webinar Series

About the Series

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

Upcoming Webinars

- Today – Challenges, opportunities, and considerations for secondary researchers using electronic health records and real-world data sources
- August 28 – Impact and Innovations from use of NIDDK-CR Resources

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

National Institute of Diabetes and Digestive and Kidney Diseases

Central Repository

Meet the Experts

Jasmin Phua is Head of Government Solutions at Datavant, leading data strategy, enterprise architecture, and data governance for public sector clients. Jasmin previously co-founded Health Data Link, a privacy-preserving record linkage technology company, now part of Datavant. Prior to that Jasmin was the Executive Director for a public-private partnership with the State of Illinois Department of Health, serving as a health data hub for the region connecting data for hospitals, labs, acute care facilities, and skilled nursing facilities throughout the state.

Hythem Sidky, PhD, is the Technical Lead for the National Clinical Cohort Collaborative (N3C) at the National Center for Advancing Translational Sciences (NCATS). Dr. Sidky's experience spans the public and private sectors, where he has directed the development of cutting-edge AI solutions for real-world evidence, biomedical imaging, and clinical data interoperability. He has developed and implemented novel machine learning models for causal inference, patient phenotyping, the early detection of sepsis, and graph deep learning solutions to analyze healthcare provider networks. Dr. Sidky holds a Ph.D. in Biomolecular Engineering and an M.S. in Applied Mathematics from the University of Notre Dame.

National Institute of Diabetes and Digestive and Kidney Diseases

Central Repository

Challenges, Opportunities, and Considerations for Researchers using Electronic Health Records and Real-World Data Sources

July 31, 2025

Jasmin Phua
Head of Government Solutions
Datavant

Hythem Sidky, PhD
Technical Lead, National Clinical Cohort Collaborative (N3C), NCATS

Session Goals

- Types of real-world data (RWD)
- Real-world data use cases and considerations
- Ingredients for a privacy-preserving RWD research infrastructure
  - National Clinical Cohort Collaborative (N3C)
  - Connecting electronic health record data with Medicare and Medicaid administrative claims, and mortality data
  - Design & Analysis Considerations

What is Real-World Data (RWD)?

Information collected through normal healthcare activities or day-to-day life

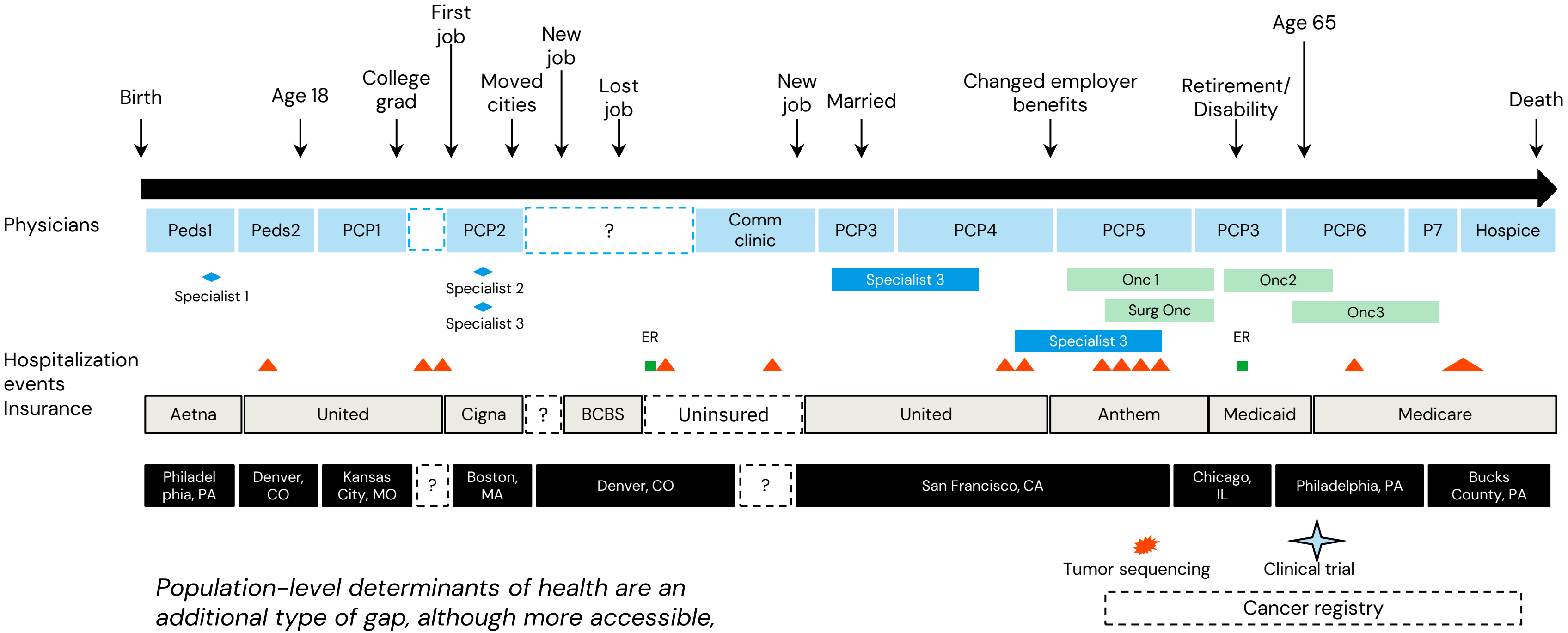
Outside the confines of a clinical trial

Diverse, multimodal RWD types are coming online every year

- Open and Closed Claims
- Patient Advocacy Group Registries
- Mortality
- Electronic Health Records (EHR)
- Social Determinants of Health (SDoH)
- Labs
- Wearable Technology
- Apps
- Grocery and lifestyle purchases
- Imaging Data
- Device Data (i.e., EKG, glucose monitors)
- Specialty Pharmacy Data
- Medical Records and Clinical Notes
- Government registries
- Hub Support Program data
- Patient Reported Outcomes (PROs)
- Digital Engagement (online behaviors)
- Genetic Testing and Whole Genome Sequencing
- Sensors
- Sentiment and Health Engagement
- Weather Data
- Financial Data (debt, income)

What causes healthcare data fragmentation and gaps?

Illustrative patient example



Birth
Age 18
College grad
First job
Moved cities
New job
Lost job
New job
Married
Changed employer benefits
Retirement/ Disability
Age 65
Death

Physicians
Peds1 | Peds2 | PCP1 | | PCP2 | ? | Comm clinic | PCP3 | PCP4 | PCP5 | PCP3 | PCP6 | P7 | Hospice

Specialist 1
Specialist 2
Specialist 3
Specialist 3
Onc 1
Onc2
Surg Onc
Onc3
Specialist 3

Hospitalization events
ER
ER

Insurance
Aetna | United | Cigna | ? | BCBS | Uninsured | United | Anthem | Medicaid | Medicare

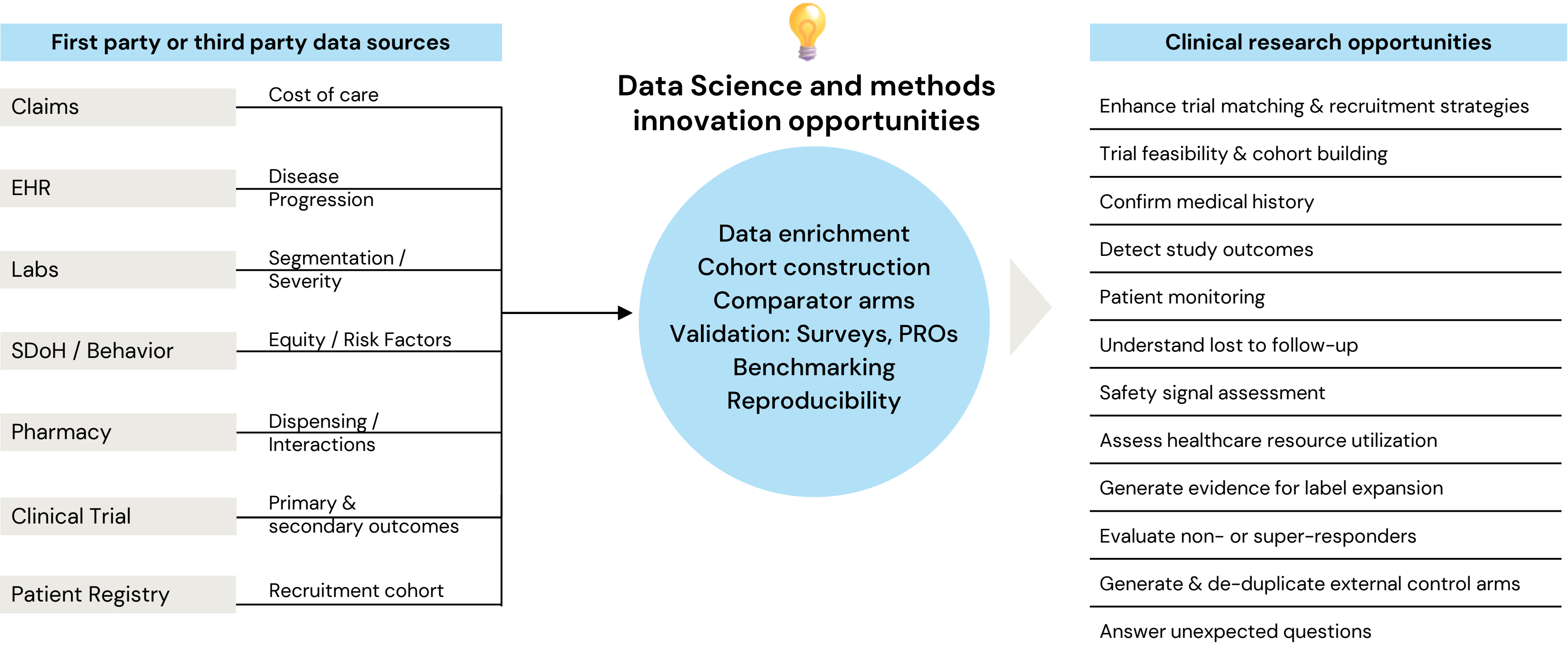
Philadelphia, PA | Denver, CO | Kansas City, MO | ? | Boston, MA | Denver, CO | ? | San Francisco, CA | Chicago, IL | Philadelphia, PA | Bucks County, PA

Tumor sequencing
Clinical trial
Cancer registry

Population-level determinants of health are an additional type of gap, although more accessible, e.g. environmental exposures

Rethink: Data integration & reuse as part of your data strategy

Rethink study design in terms of primary data collection vs observational data, and hybrid approaches



First party or third party data sources

Claims	Cost of care
EHR	Disease Progression
Labs	Segmentation / Severity
SDoH / Behavior	Equity / Risk Factors
Pharmacy	Dispensing / Interactions
Clinical Trial	Primary & secondary outcomes
Patient Registry	Recruitment cohort

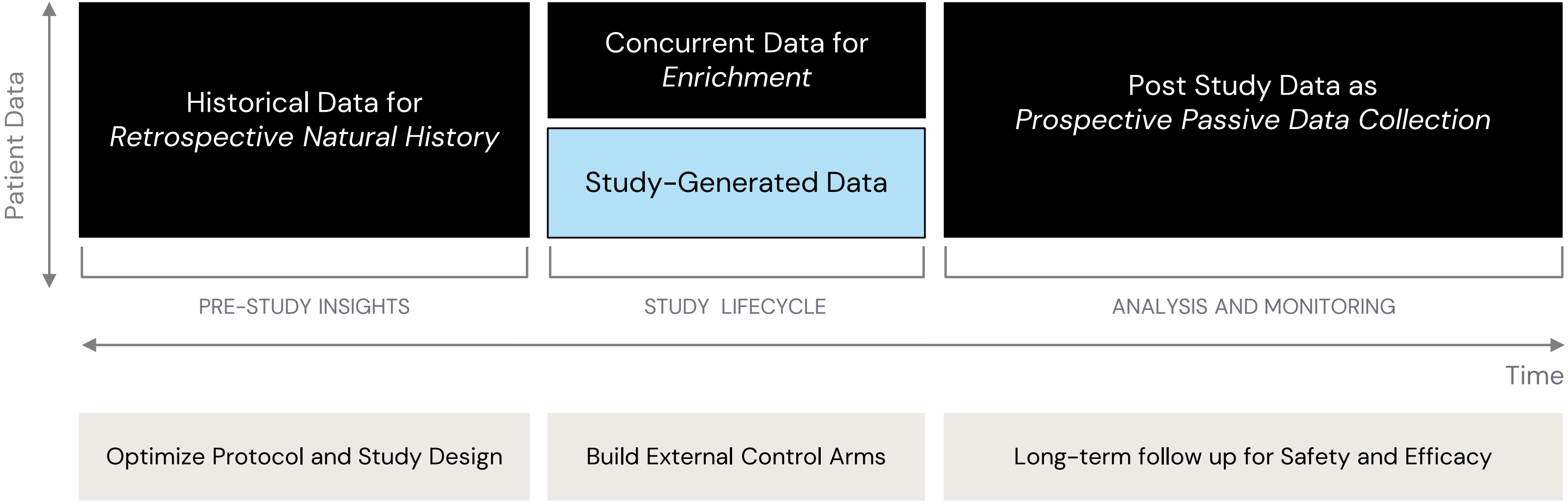
Data Science and methods innovation opportunities

Data enrichment
Cohort construction
Comparator arms
Validation: Surveys, PROs
Benchmarking
Reproducibility

Clinical research opportunities

- Enhance trial matching & recruitment strategies
- Trial feasibility & cohort building
- Confirm medical history
- Detect study outcomes
- Patient monitoring
- Understand lost to follow-up
- Safety signal assessment
- Assess healthcare resource utilization
- Generate evidence for label expansion
- Evaluate non- or super-responders
- Generate & de-duplicate external control arms
- Answer unexpected questions

Re-envision: RWD as integral study design component throughout person journey



Patient Data

Historical Data for *Retrospective Natural History*

Concurrent Data for *Enrichment*

Study-Generated Data

Post Study Data as *Prospective Passive Data Collection*

PRE-STUDY INSIGHTS

STUDY LIFECYCLE

ANALYSIS AND MONITORING

Time

Optimize Protocol and Study Design

Build External Control Arms

Long-term follow up for Safety and Efficacy

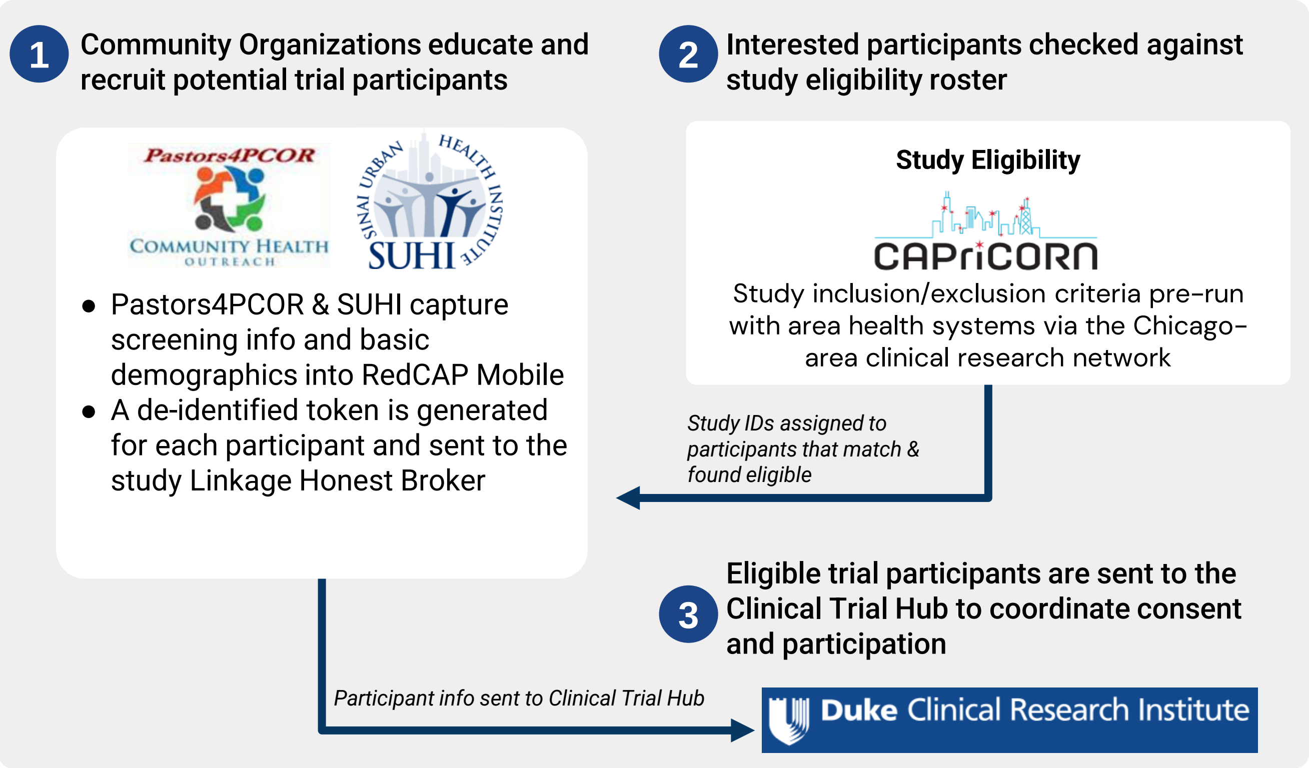
Re-envision: Enable new study recruitment and engagement models

Goal

Recruitment focused on at-risk and hard to reach patients

How

1. Recruit via **non-traditional** community organizations
2. Participant **eligibility pre-run** on area population based on electronic health records.
3. Checking for eligibility **lowers burden on participants to share information** related to study inclusion/exclusion criteria.



1 Community Organizations educate and recruit potential trial participants

- Pastors4PCOR & SUHI capture screening info and basic demographics into RedCAP Mobile
- A de-identified token is generated for each participant and sent to the study Linkage Honest Broker

2 Interested participants checked against study eligibility roster

Study Eligibility

Study inclusion/exclusion criteria pre-run with area health systems via the Chicago-area clinical research network

Study IDs assigned to participants that match & found eligible

3 Eligible trial participants are sent to the Clinical Trial Hub to coordinate consent and participation

Participant info sent to Clinical Trial Hub

Zimmerman LP, Goel S, Sathar S, Gladfelter CE, Onate A, Kane LL, Sital S, Phua J, Davis P, Margellos-Anast H, Meltzer DO, Polonsky TS, Shah RC, Trick WE, Ahmad FS, Kho AN. A Novel Patient Recruitment Strategy: Patient Selection Directly from the Community through Linkage to Clinical Data. Appl Clin Inform. 2018 Jan;9(1):114-121. doi: 10.1055/s-0038-1625964. Epub 2018 Feb 14. PMID: 29444537; PMCID: PMC5843765.

NCATS

COLLABORATE. INNOVATE. ACCELERATE.

National Clinical Cohort Collaborative

Past, present, and future

Hythem Sidky, PhD

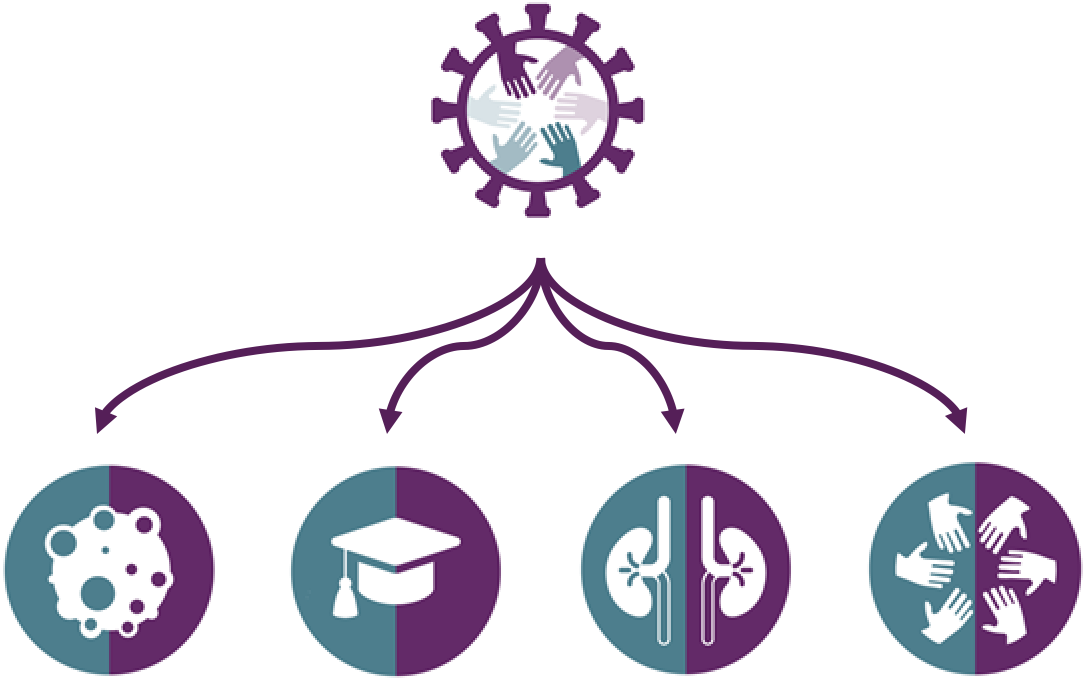
National Center for Advancing Translational Sciences

Why N3C?

- Urgent need for observational data at scale.
- In the US, there is no centralized healthcare, and therefore no centralized healthcare data.
- Data from a single person is spread across multiple providers across time and geography.

Why Expand?

- COVID-19 served a powerful proof-of-concept, demonstrating the value of large-scale clinical data harmonization and collaboration.
- Significant investments and teamwork have created a robust platform and governance model that can be expanded for broader use.
- Beyond COVID-19, chronic diseases and complex conditions remain critical public health priorities.
- Broader utilization of N3C infrastructure provides valuable opportunities for training researchers, clinicians, and students in real-world data analysis, informatics, and collaborative science.



NIH National Center for Advancing Translational Sciences

N3C Past and Present

N3C has grown from a COVID-19 response into a national translational research infrastructure, combining harmonized EHR data, scalable governance, and team science to accelerate discovery across diseases and institutions.

2020

N3C is Launched!

In response to the COVID-19 pandemic, the National COVID Cohort Collaborative was formed to create the largest publicly available, harmonized EHR dataset in U.S. history.

Today

N3C's impact

5000+ citations, H-index 33, 1589 authors. N3C enabled transformative research and care guidelines, disease definitions, and predictive models for outcomes across comorbidities.

2022-2023

Phase 1 Clinical Pilot

N3C successfully expanded beyond COVID-19, piloting clinical tenants for Alzheimer's, COPD, and Renal disease across 12 institutions.

2024-2027

Phase 2 Clinical Pilot

Building on Phase 1, Phase 2 scales with enhanced PPRL, data integration (e.g., CMS, SEER), and supports new tenants like cancer and renal.

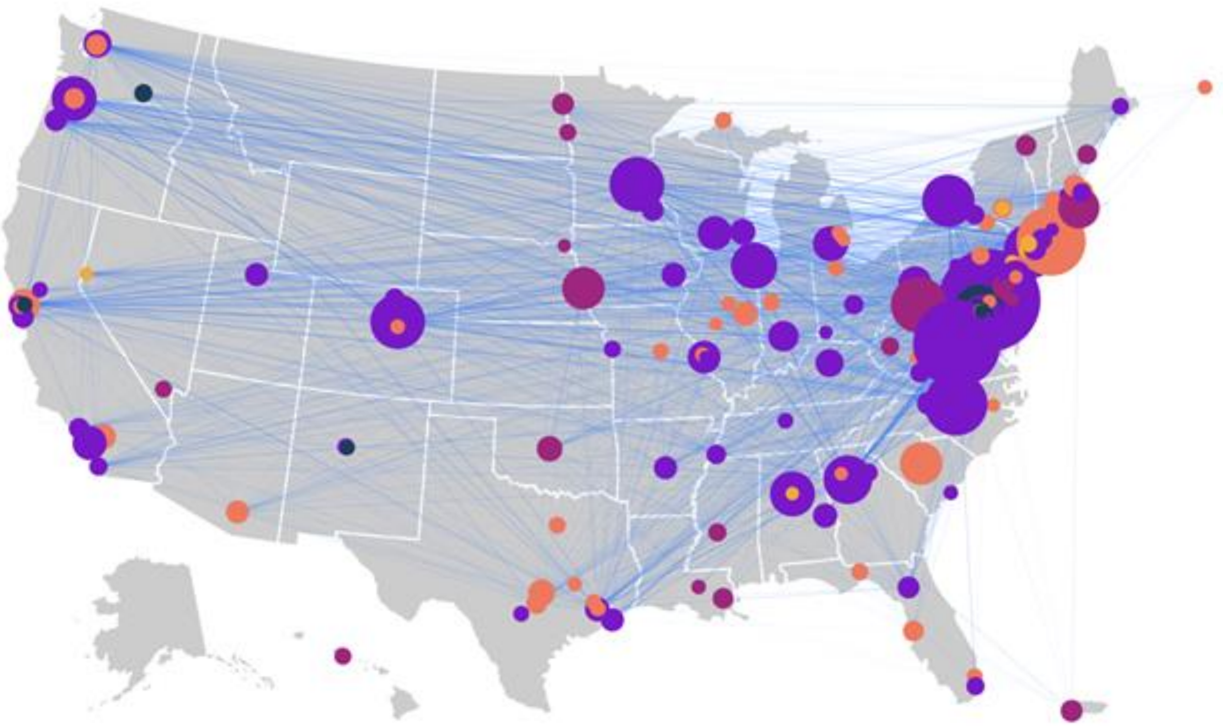
Continued Community Engagement

96 Data Contributors signed the original COVID Data Transfer Agreement

76 Data Contributors signed the COVID Data Transfer Agreement Extensions

12 institutions participated in the Phase I Clinical Pilot

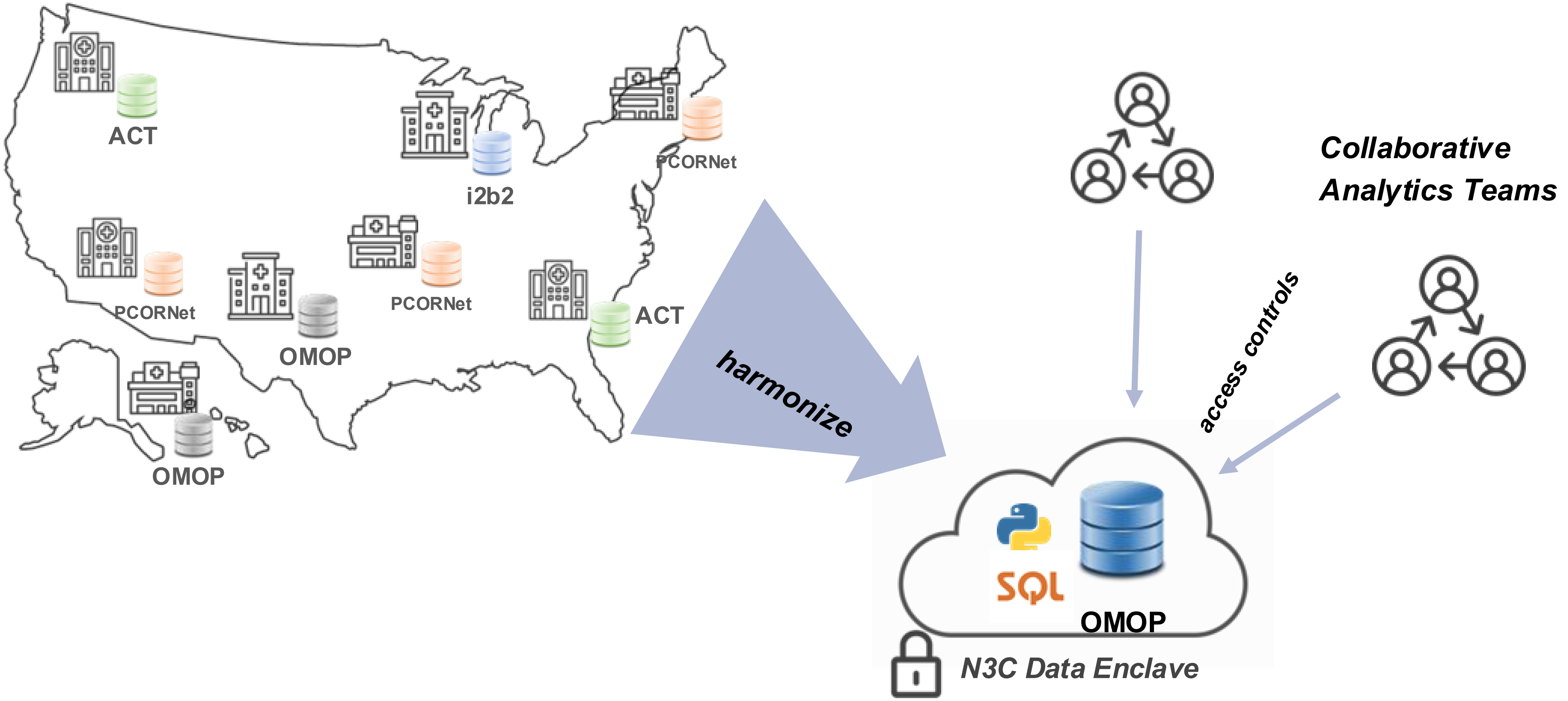
18 institutions are participating in the Phase II Clinical Pilot



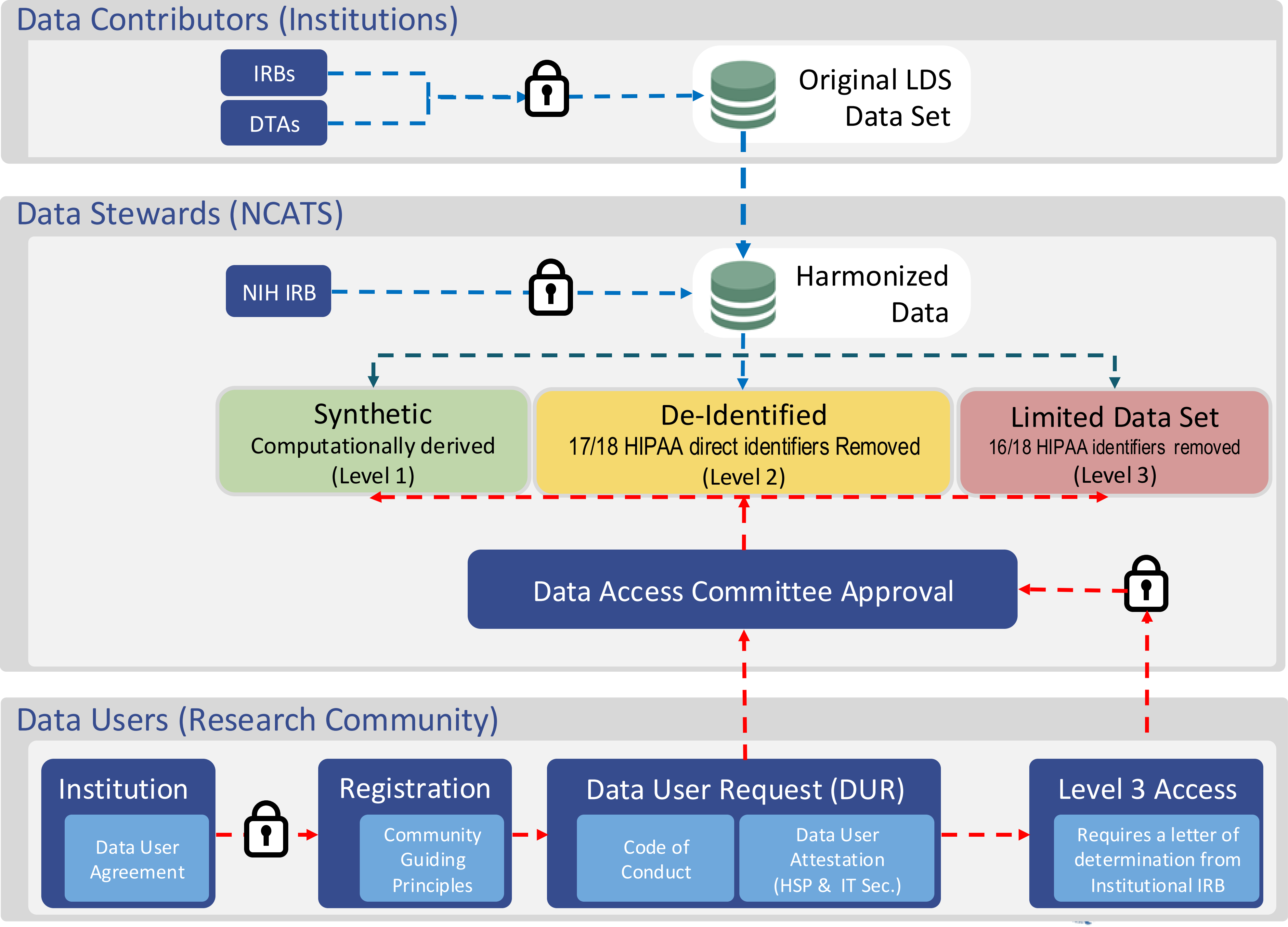
National
Clinical
Cohort
Collaborative

How it works

N3C: High Level



N3C: Data Governance and Access



Data Contributors (Institutions)

IRBs

DTAs

Original LDS Data Set

Data Stewards (NCATS)

NIH IRB

Harmonized Data

Synthetic
Computationally derived
(Level 1)

De-Identified
17/18 HIPAA direct identifiers Removed
(Level 2)

Limited Data Set
16/18 HIPAA identifiers removed
(Level 3)

Data Access Committee Approval

Data Users (Research Community)

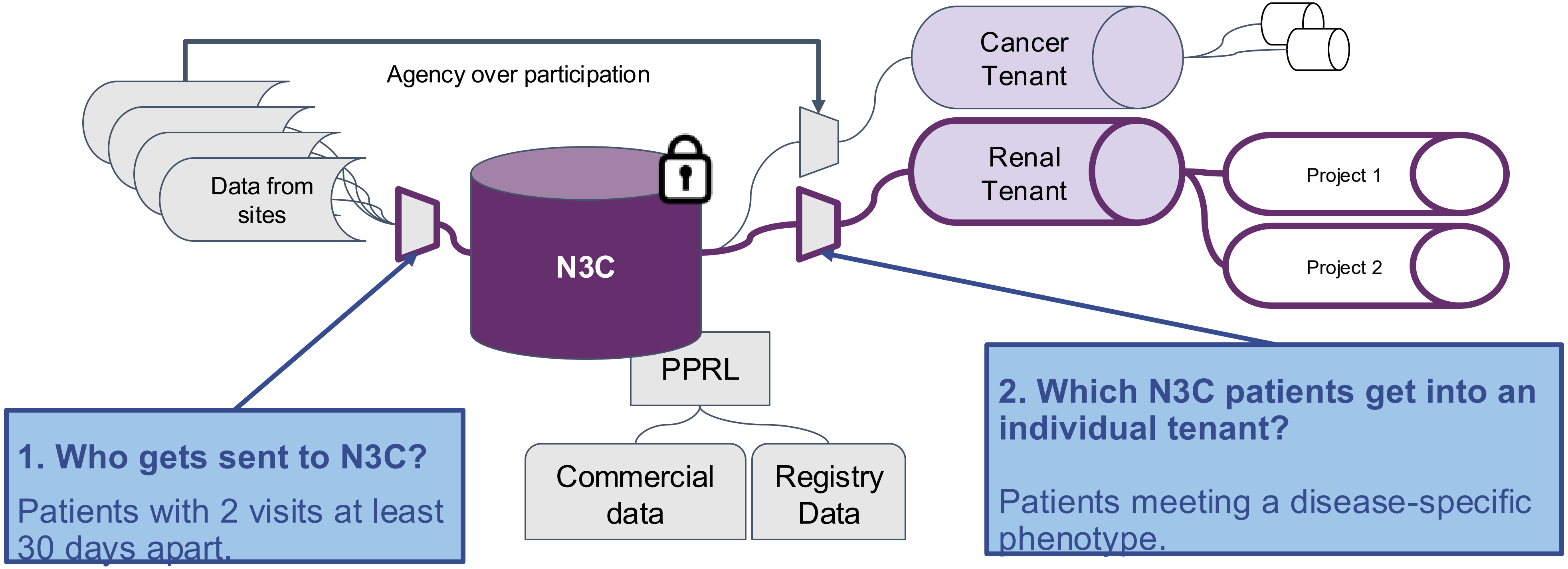
Institution
Data User Agreement

Registration
Community Guiding Principles

Data User Request (DUR)
Code of Conduct
Data User Attestation (HSP & IT Sec.)

Level 3 Access
Requires a letter of determination from Institutional IRB

N3C Phase 2 Clinical Pilot Model



1. Who gets sent to N3C?

Patients with 2 visits at least 30 days apart.

2. Which N3C patients get into an individual tenant?

Patients meeting a disease-specific phenotype.

Objectives of the N3C Clinical Pilots

- N3C Clinical pilots were meant to help NCATS more accurately understand the financial, infrastructure, and community resources needed to develop and maintain future tenants.
- Pilots will facilitate refining operations, governance, and technical architecture.
- Establish partnerships with CMS, HRSA, NCI, NIDDK, and other HHS agencies.
- Next-generation healthcare interoperability is being developed (HL7 FHIR US Core).
- New capabilities will expand the space of scientific questions that can be asked and answered.

National
Clinical
Cohort
Collaborative

EHRs are not Enough

The Need For Multiple Data Sources

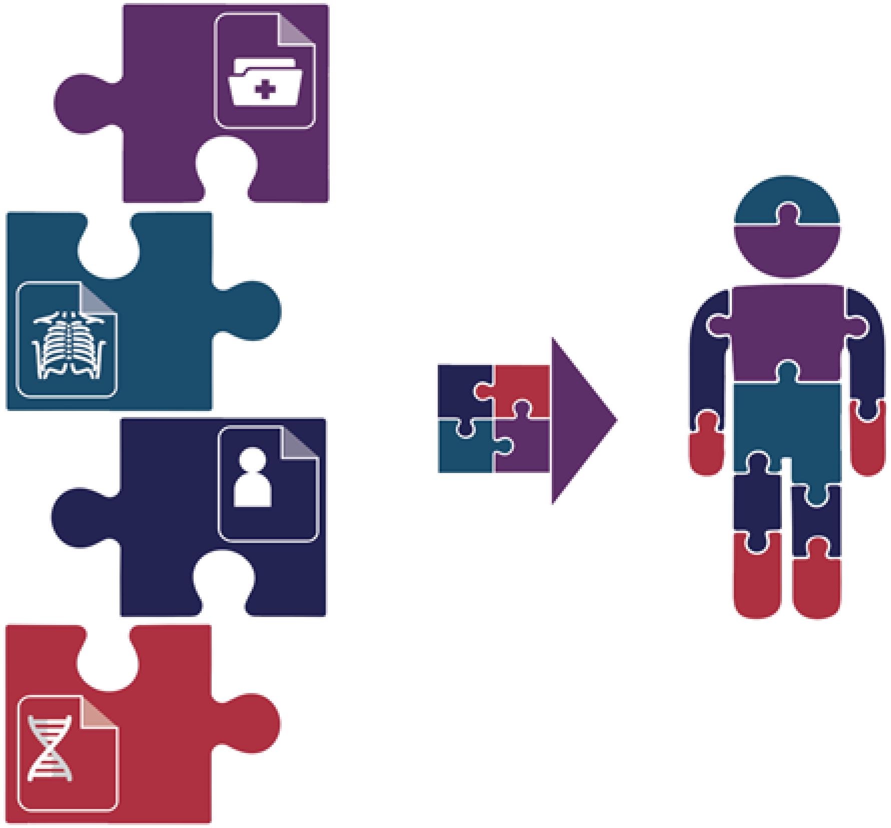
Why Link Multiple Data Sources?

- **Overcomes limitations of siloed datasets**
  EHR and other data sources capture different, complementary aspects of care.

- **Improves completeness and continuity**
  Linked data fill gaps in treatment, outcome, and longitudinal follow-up.

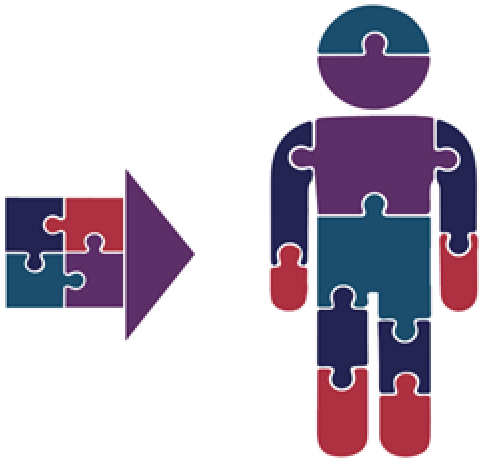
- **Reduces bias in observational studies**
  More accurate measurement of exposures, covariates, and endpoints.

- **Facilitates cross-validation**
  Conflicting values across sources can be reconciled to enhance data integrity.

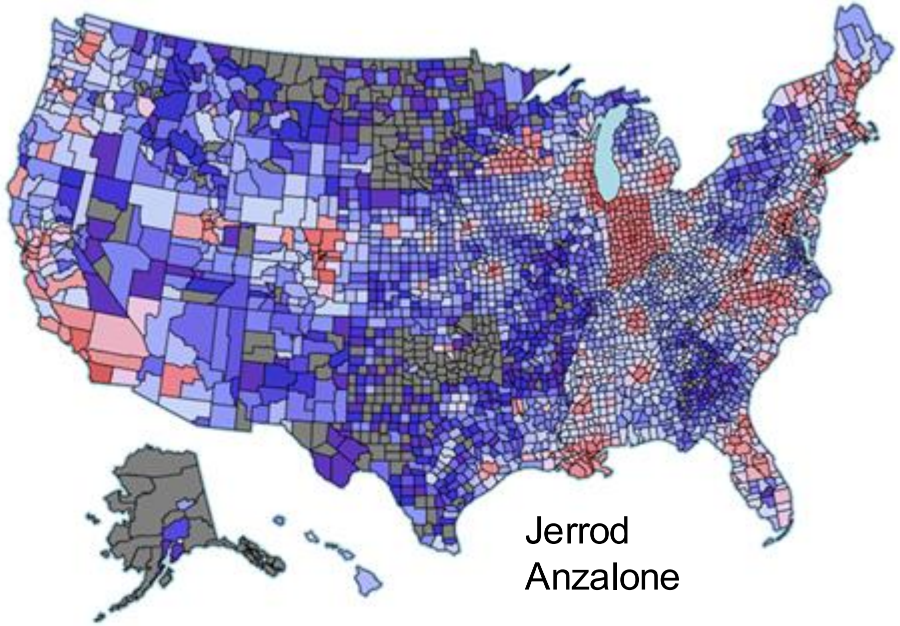


N3C – Multiple data sources

- CMS Data
- SDOH Data
- Viral Variant
- Mortality Data
- Clinical Data
- *#Vaccine data*
- **Imaging MIDRC*
- **SEER, SRTR, NAACCR**



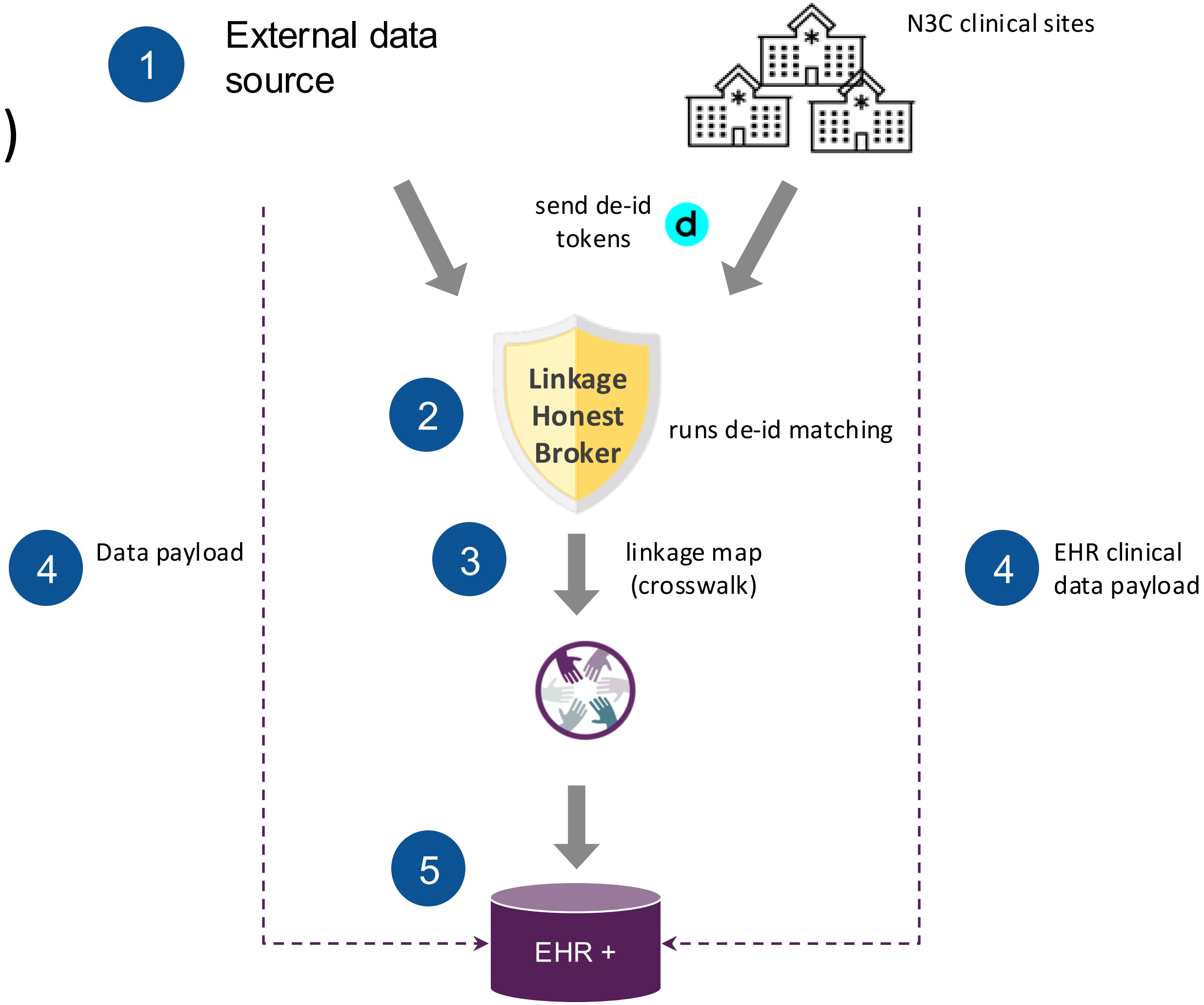
Privacy Preserving Record Linkage, (PPRL) a means of connecting records using secure, pseudonymization processes in a data set that refer to the same individual across different data sources while maintaining the individuals' privacy



Jerrod Anzalone

Linked Data Flow with N3C

1. Tokenization of Data (Datavant tokens)
2. Honest Broker De-Identified Matching
3. Matches Communicated
4. Clinical Payloads Sent to N3C
5. NCATS Links N3C and External Data



How Important is Data Linkage?

EHR alone:

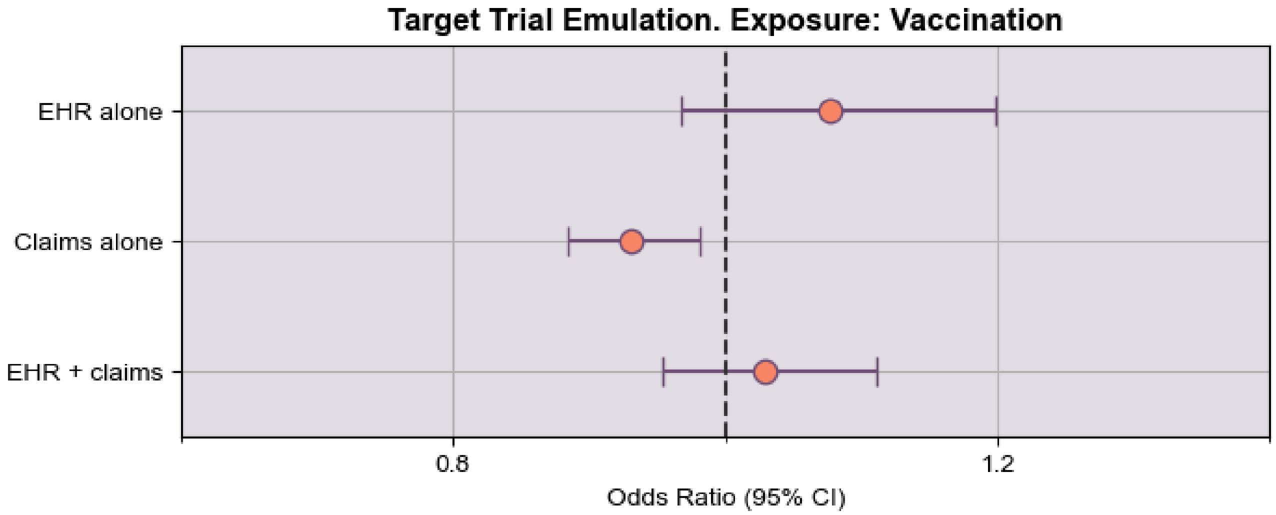
- Non-significant association between vaccination and *higher* odds of stroke OR = 1.19, 95% CI : (0.97, 1.20), p = 0.18

CMS alone:

- Significant association between vaccination and lower odds of stroke OR = 0.93, 95% CI : (0.88, 0.98), p = 0.008

Combined:

- Non-significant association between vaccination and stroke OR = 1.03, 95% CI: (0.95, 1.11), p = 0.45

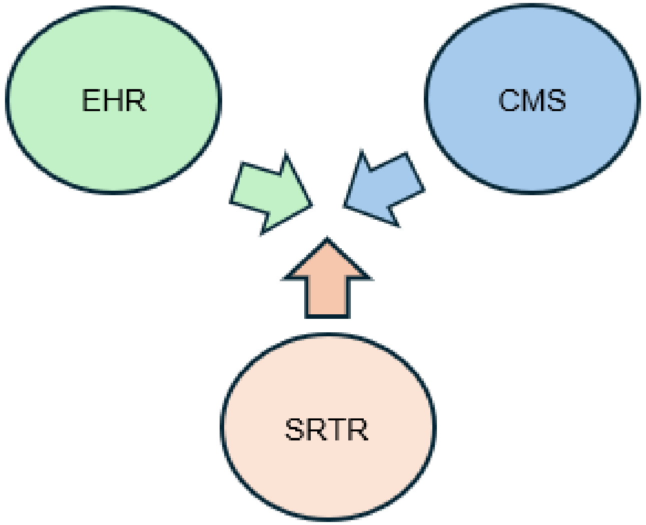


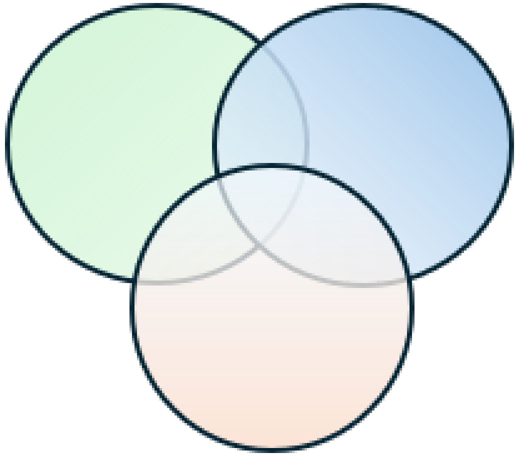
Goals of Renal Tenant

- Better understand and conceptualize the patient journey through CKD, dialysis, transplant, and beyond
    - Patient trajectories, dialysis “drop-ins”, organ allocation, organ donors, difference in death rates across data sets, transplant referral, understand bias and quantify issues in underserved populations, etc.
- Answer nuanced questions through the combination and linkage of EHR, billing, and transplant data
- Combining data is essential to arrive at the correct results/conclusions
- Create open science community, team science

Added Value of the Renal Tenant

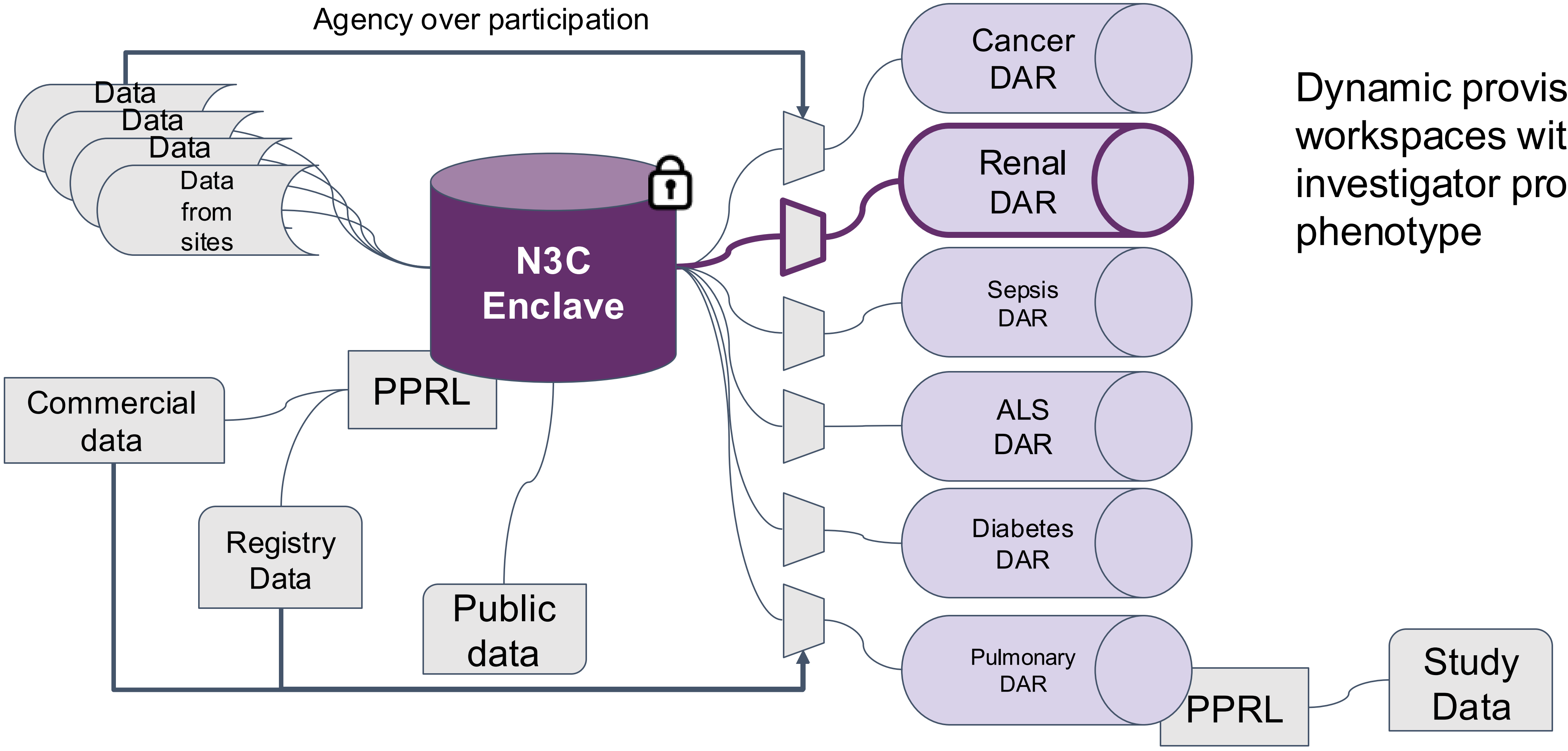
Identify the added value of **linking datasets** in the Renal Tenant on **priority topics for HHS**





Good Algorithmic Practice: Test Bed for Validation of AI

The future of N3C: Dynamic Workspaces



Dynamic provisioning of workspaces with investigator provided phenotype

Imagine if…

You want to study sepsis with 5 other institutions

Without N3C

Kickoff

Administrative Setup

Identify partners, Initiate data-sharing agreement negotiations, Start individual IRB processes

Month 4

Legal Negotiations

Finalize multiple institution-specific agreements, address privacy/security/legal compliance across multiple institutions.

Month 7

Technical Infrastructure Setup

Build/procure secure collaborative platform, establish compute environment (high cost).

Month 10

Data Harmonization

Independently clean and harmonize data at each institution. Link external datasets (high cost).

Month 13

Project Start

Data ready for analysis after significant delays.

With N3C

Kickoff

Setup

Submit a DAR. Single Master Data Transfer Agreement (already executed). Rapid institutional opt-out review process

Week 3

Data & Workspace Provisioning

Creation of secure dynamic workspace. Data already harmonized to OMOP. Immediate linkage to external data (CMS, mortality data, and more).

Week 5

Project Start

Begin collaborative analysis in secure, scalable environment.

National
Clinical
Cohort
Collaborative

NIH National Center for Advancing Translational Sciences

National
Clinical
Cohort
Collaborative

Questions?

Contacts:

- Jasmin Phua - jas@datavant.com
- Hythem Sidky, PhD - hythem.sidky@nih.gov

Upcoming Webinar: Impact and Innovations from use of NIDDK-CR Resources

- **Date:** August 28th from 2-4pm ET
- **Experts:**
  - **Dr. Adam Gaweda**, Assistant Professor in the University of Louisville Department of Medicine, on “AI-driven Personalized Predictive Modeling of Kidney Disease Progression”
  - **Dr. Prasanna Santhanam**, Associate Professor of Clinical Medicine and Oncology at Johns Hopkins University School of Medicine, and Co-founder of AI-Metab, LLC, on “AI, on Body Composition: Novel Methods to Improve Accuracy”
  - **Dr. Juliet Emamaullee**, Associate Professor of Surgery and Immunology (Clinical Scholar) at University of Southern California Keck School of Medicine, and an attending liver and kidney transplant surgeon at Keck Hospital and Children's Hospital-Los Angeles, on “Creation of the CHLA Acute Liver Failure Score to predict need for transplant in children with acute liver failure.”
- **Scan the QR code register**

Thank You!