NIH Collaboratory

Health Care Systems Research Collaboratory

The NIH Collaboratory Distributed Research Network: A Privacy Protecting Method for Sharing Research Data Sets

Jeffrey Brown, Lesley Curtis, and Rich Platt

June 13, 2014

NIH Collaboratory

Health Care Systems Research Collaboratory

Previously

The NIH Collaboratory:

Data Sharing Principles- An Initial Discussion

Robert M Califf and Catherine Meyers

What is Reproducible Research?

- **Data**: Analytic dataset is available
- **Methods**: Computer code underlying figures, tables, and other principal results is available
- **Documentation**: Adequate documentation of the code, software environment, and data is available
- **Distribution**: Standard methods of distribution are employed for others to access materials

From Dr. Califf's Grand Rounds, May 30, 2014

"Are you just pissing and moaning, or can you verify what you're saying with data?"

What is PCORI's data-sharing policy?

We require that a complete, cleaned, de-identified copy of the final data set used in conducting the final analyses be made available within nine months of the end of the final year of funding.

pcornet

NIH data sharing policies

- The privacy of participants should be safeguarded
- Data should be made as widely and freely available as possible
- Data should be shared no later than the acceptance for publication of the main study findings
- Initial investigators may benefit from first and continuing use of data, but not from prolonged exclusive use

Policy is consistent with clinical research that has monitored data capture under informed consent

http://grants.nih.gov/grants/policy/data_sharing/data_sharing_guidance.htm#time2

Data sharing within health system research

- Routinely collected health system data come from a wide range of sources linked for analysis
  - Ambulatory facilities, hospitals, pharmacies, health insurers, public registries
- Data are rarely collected under informed consent for research
- Sharing of clinical data used for research requires special consideration
  - Patient privacy issues
  - Health care system proprietary and confidentiality issues
- Multi-site studies without a central data warehouse raise additional complications

NIH Collaboratory draft data sharing policy

- REQUIRED: All Collaboratory trials are expected to share one or more public use datasets through an **unsupervised data archive**.

- OPTIONAL: Collaboratory trials may also choose to make more **detailed data available through a more restricted data access mechanism** (eg, data enclave). This is appropriate when sharing would increase risk of re-identification or other misuse.

Paraphrased from Greg Simon's February presentation to NIH HCS Collaboratory Steering Committee:
https://www.nihcollaboratory.org/news/Pages/February2014_Steering-Committee_meeting.aspx

De-identified data may not be very useful

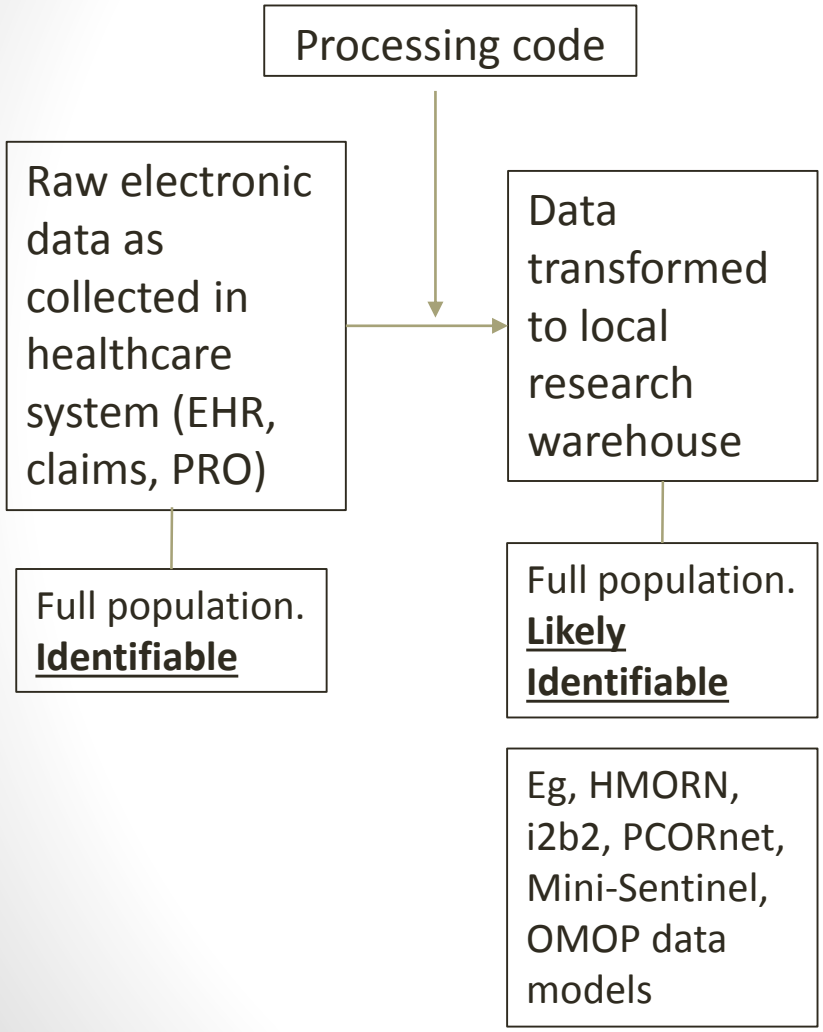
- Most studies need HIPAA identifiers like exact dates; some need zip code
- Data obfuscation (eg, date shifting) can be difficult to verify and can cause loss of value
  - No single obfuscation approach works in all situations
  - Seasonality and calendar year may be important confounders
  - Utilization patterns and procedures codes can reveal calendar time and age
- De-identification in the context of a multi-site study introduces a potential complicating factor if not done identically

What data are potentially shareable?

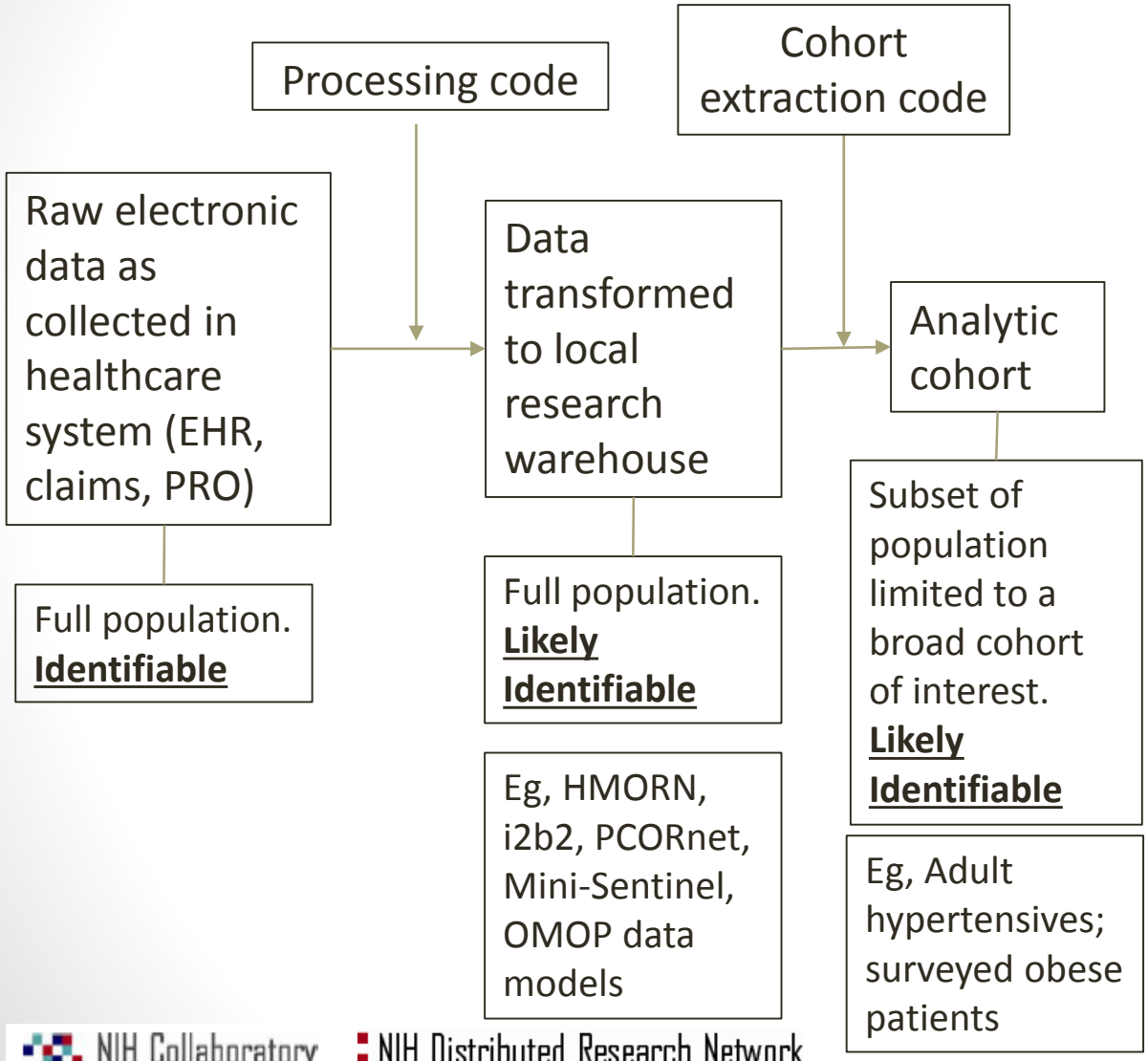
Raw electronic data as collected in healthcare system (EHR, claims, PRO)

Full population. **Identifiable**

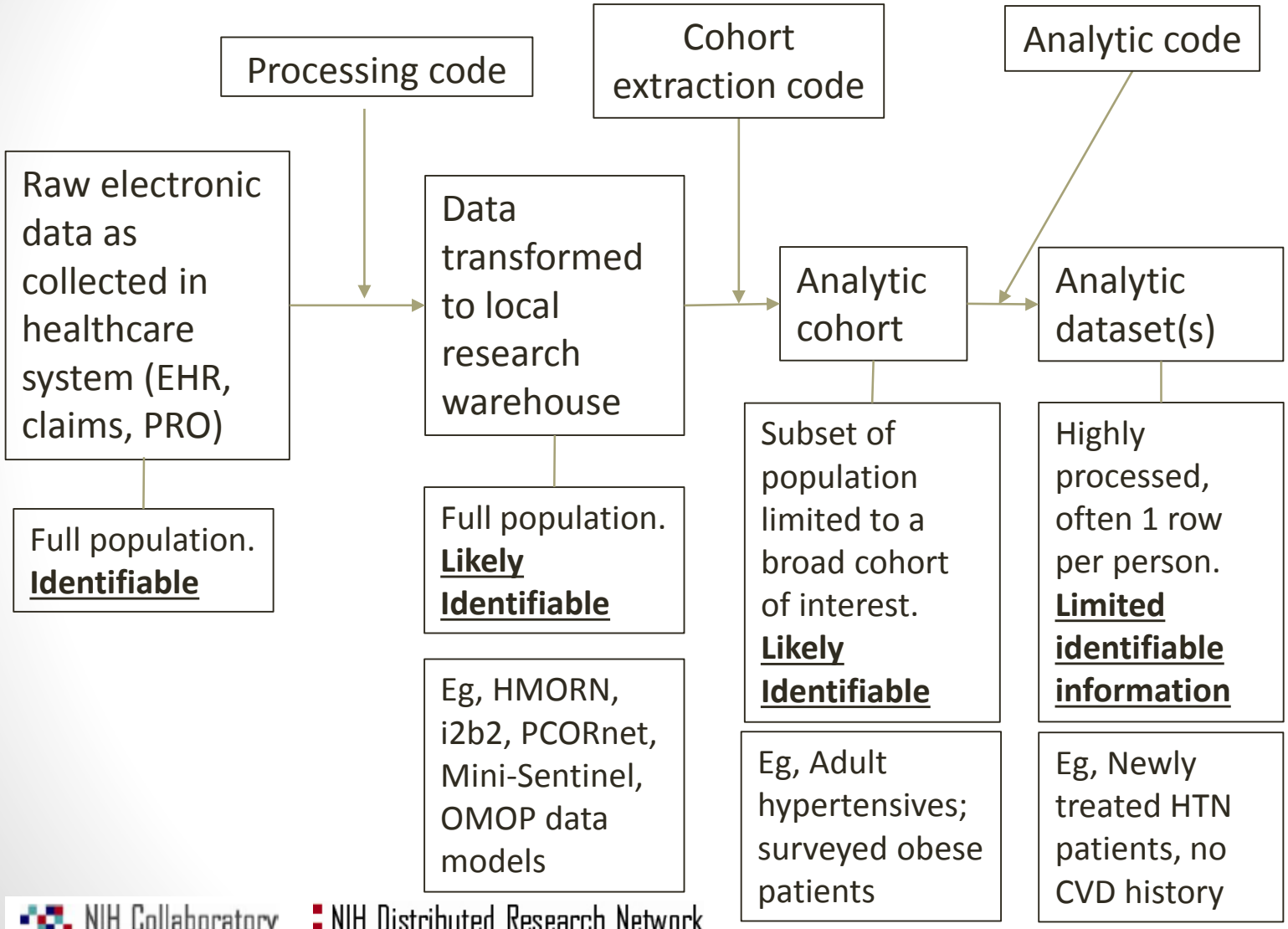
What data are potentially shareable?



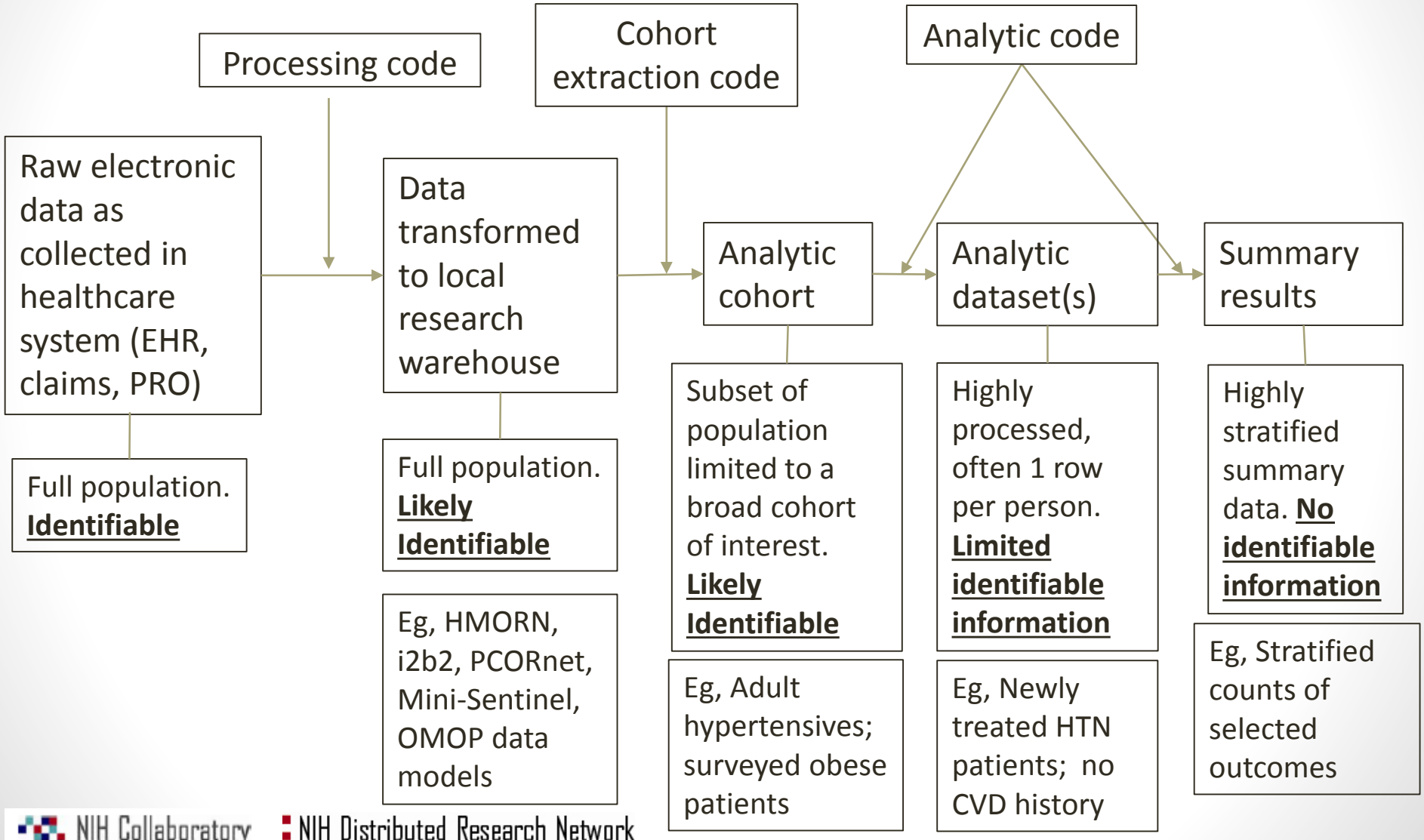
What data are potentially shareable?



What data are potentially shareable?



What data are potentially shareable?



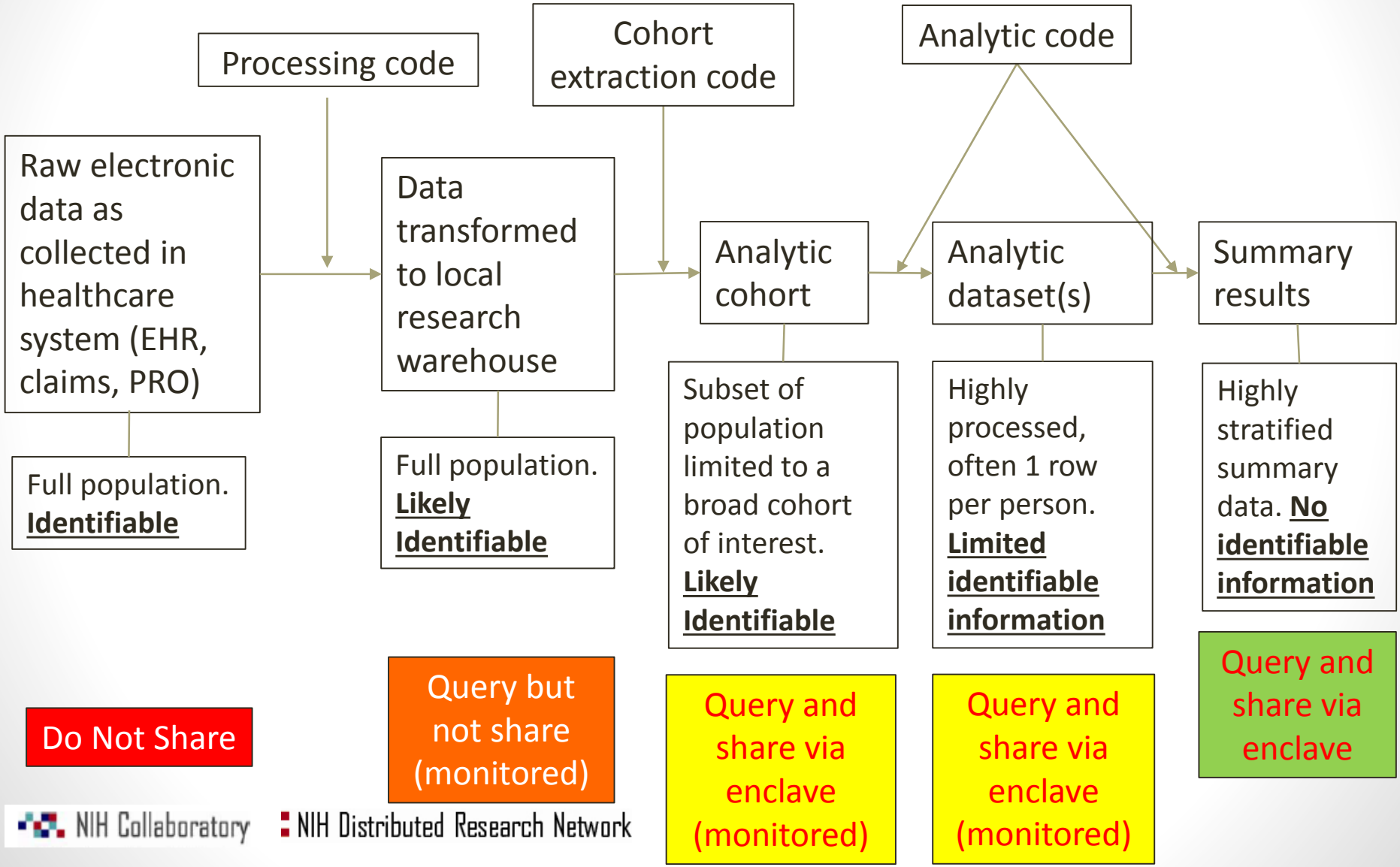
MHRN Mental Health Research Network

Technical options for data sharing (in ascending order of **data generator** control):

- Unsupervised data archive: Release appropriately de-identified data to any potential users
  Control of dataset contents only
- Unsupervised public data enclave: Allow any user to send any question to the data
  Control of dataset contents, query logic and return of results
- Unsupervised private data enclave: Allow specific users to send any question to the data
  Control of dataset contents, query logic, return of results, and user qualifications
- Supervised data archive: Release specific datasets to specific users
  Control of dataset contents, user qualifications and specific authorized use (e.g. DUA)
- Supervised private data enclave: Specific users may ask to send specific questions to data
  Control of dataset contents, user qualifications, query logic, return of results and topic

More control = more expense for infrastructure and governance.
(e.g. supervised means live people are involved)

What data are potentially shareable?



NIH Collaboratory

Health Care Systems Research Collaboratory

The NIH Distributed Research Network
New Functionality and Future Potential

Millions of people. Strong collaborations. Privacy first.

Jeffrey Brown, PhD for the NIH Health Care Systems Collaboratory EHR Core

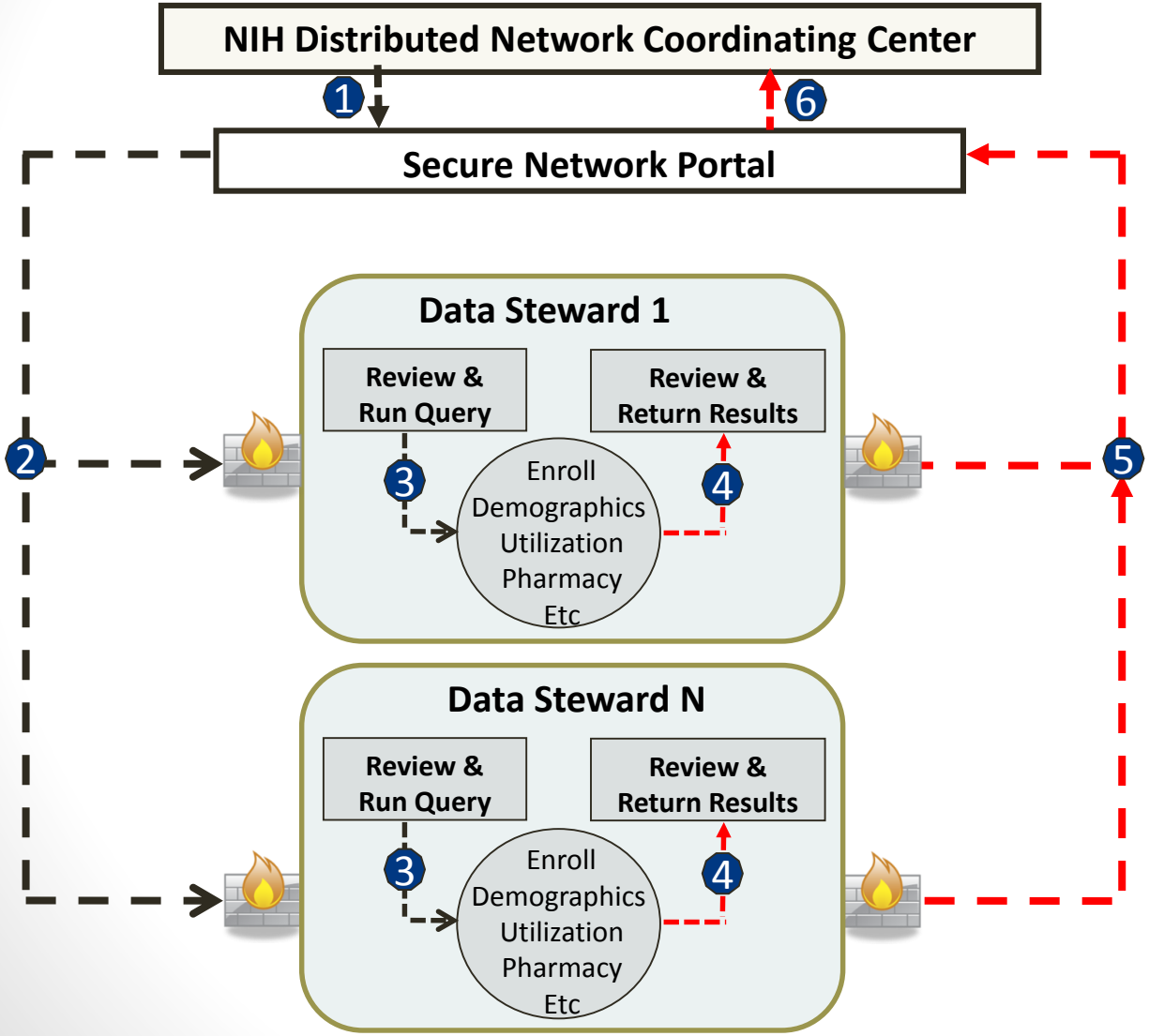
Harvard Pilgrim Health Care Institute and Harvard Medical School

September 13, 2013

Use cases

- Assess disease burden/outcomes
- Pragmatic clinical trial design
- Single study private network
- Pragmatic clinical trial follow up
- **Reuse of research data**

What is a distributed research network?



1- User creates and submits query (a computer program)

2- Data stewards retrieve query

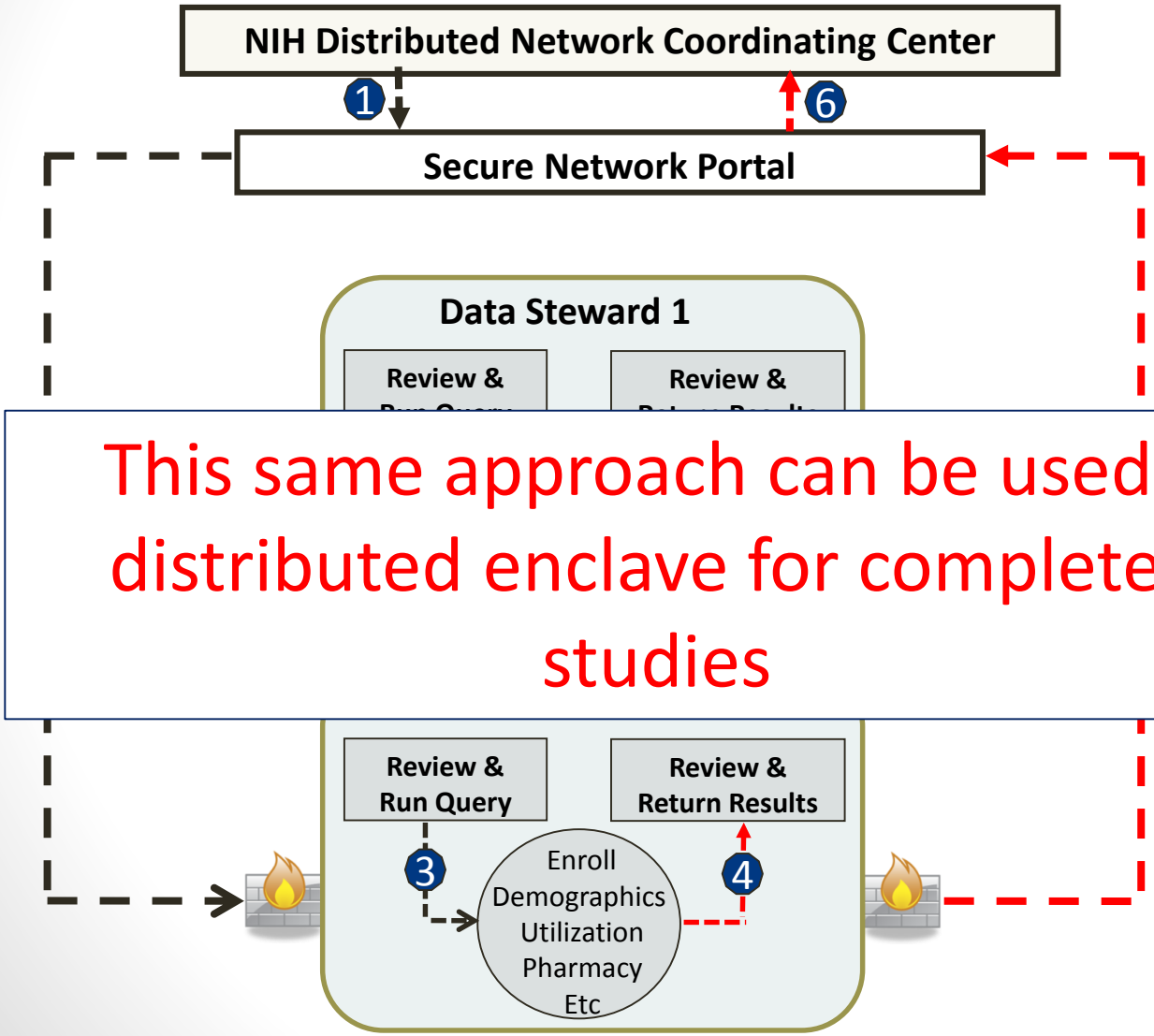
3- Data stewards review and run query against their local data

4- Data stewards review results

5- Data stewards return results via secure network

6 Results are aggregated

What is a distributed research network?



This same approach can be used a distributed enclave for completed studies

1- User creates and submits query (a computer program)

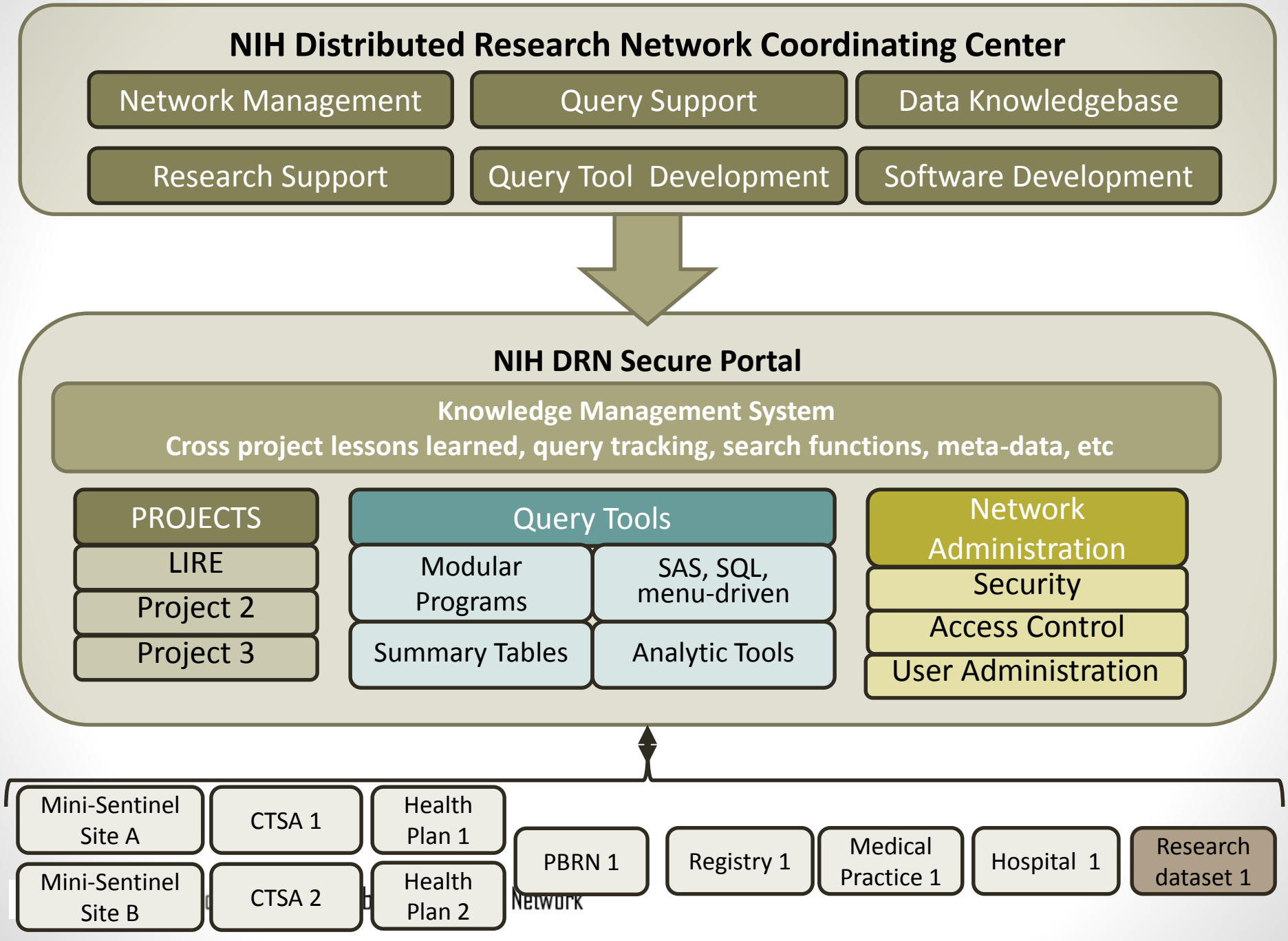
2- Data stewards retrieve query

3- Data stewards review run query against r local data

ata stewards review lts

5- Data stewards return results via secure network

6 Results are aggregated

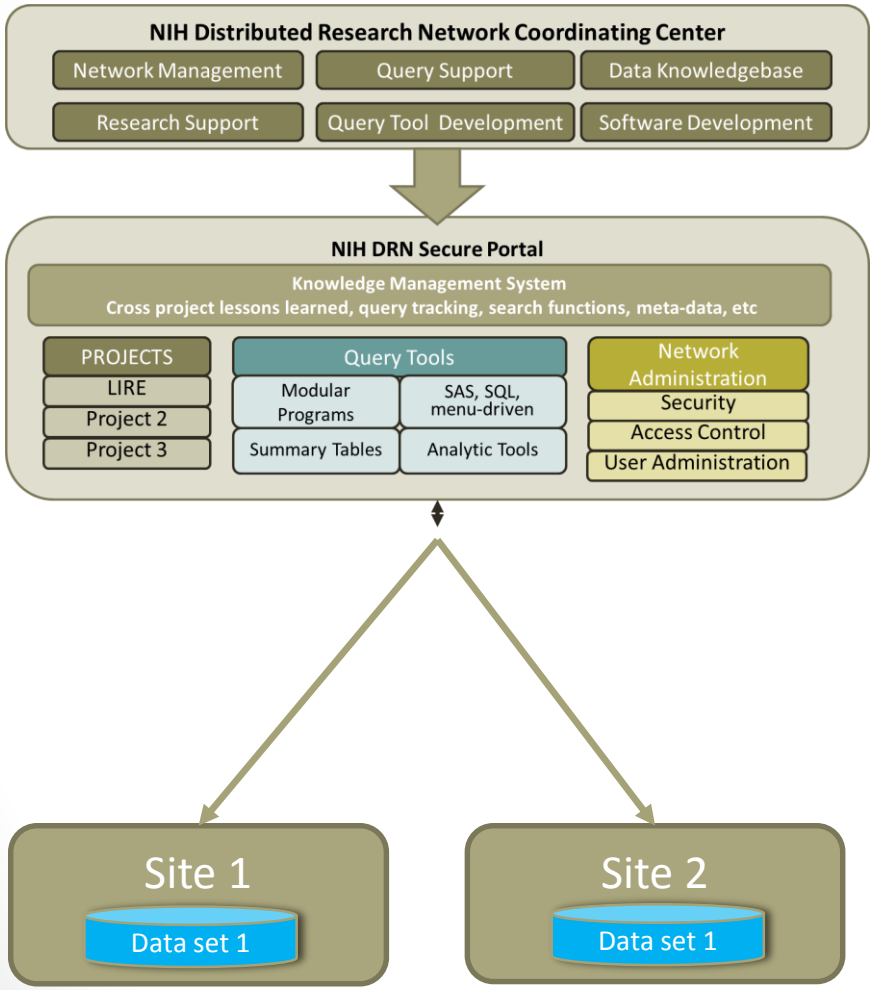

NIH Distributed Research Network Coordinating Center
Network Management
Query Support
Data Knowledgebase
Research Support
Query Tool Development
Software Development
NIH DRN Secure Portal
Knowledge Management System
Cross project lessons learned, query tracking, search functions, meta-data, etc
PROJECTS
LIRE
Project 2
Project 3
Query Tools
Modular Programs
SAS, SQL, menu-driven
Summary Tables
Analytic Tools
Network Administration
Security
Access Control
User Administration
Mini-Sentinel Site A
Mini-Sentinel Site B
CTSA 1
CTSA 2
Health Plan 1
Health Plan 2
PBRN 1
Registry 1
Medical Practice 1
Hospital 1
Research dataset 1

Storage and access

- The research dataset could live in the originating institution OR in a different secure location
- Control over access to the research dataset could live with a trusted 3rd party OR with the originating institution
- No need for multi-site studies to create a single analytic dataset for sharing

Distributed data sharing options

Keep data behind institutional firewalls, **distributed querying**



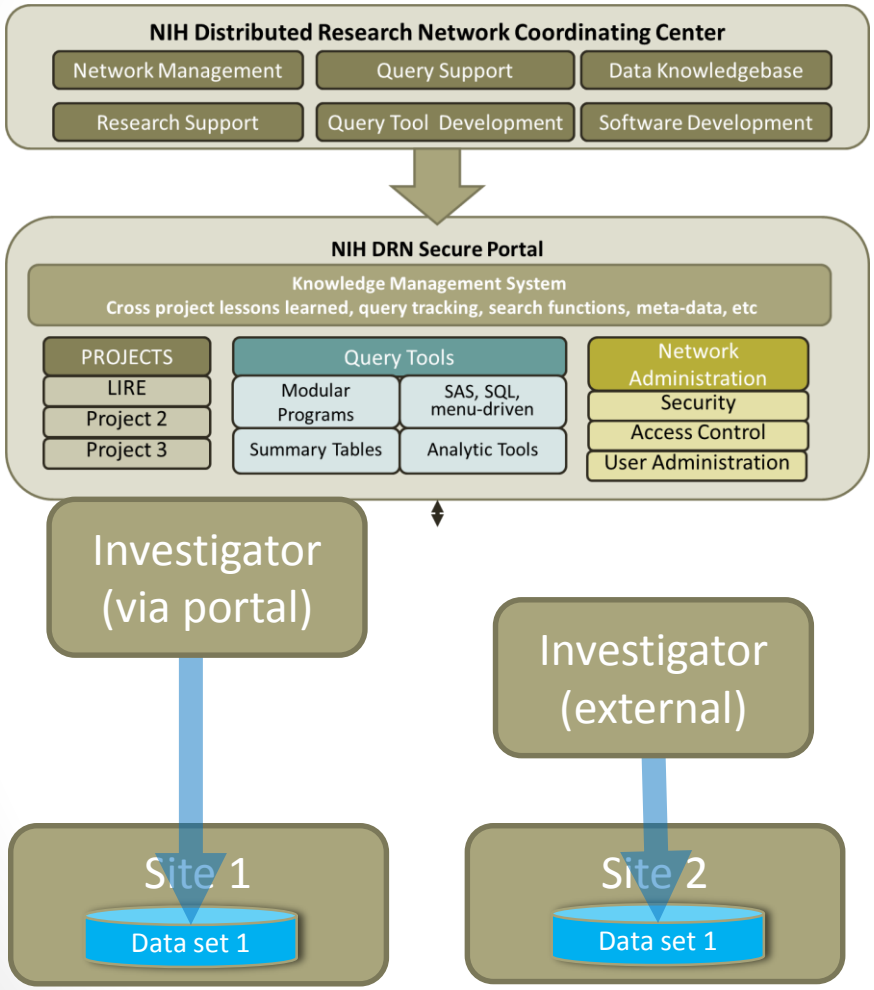
Governance, access controls, infrastructure, etc

Queries are sent, executed locally, and results returned

Sites directly control use, local staff execute queries and apply governance policies

Distributed data sharing options

Keep data behind institutional firewalls, **direct access**



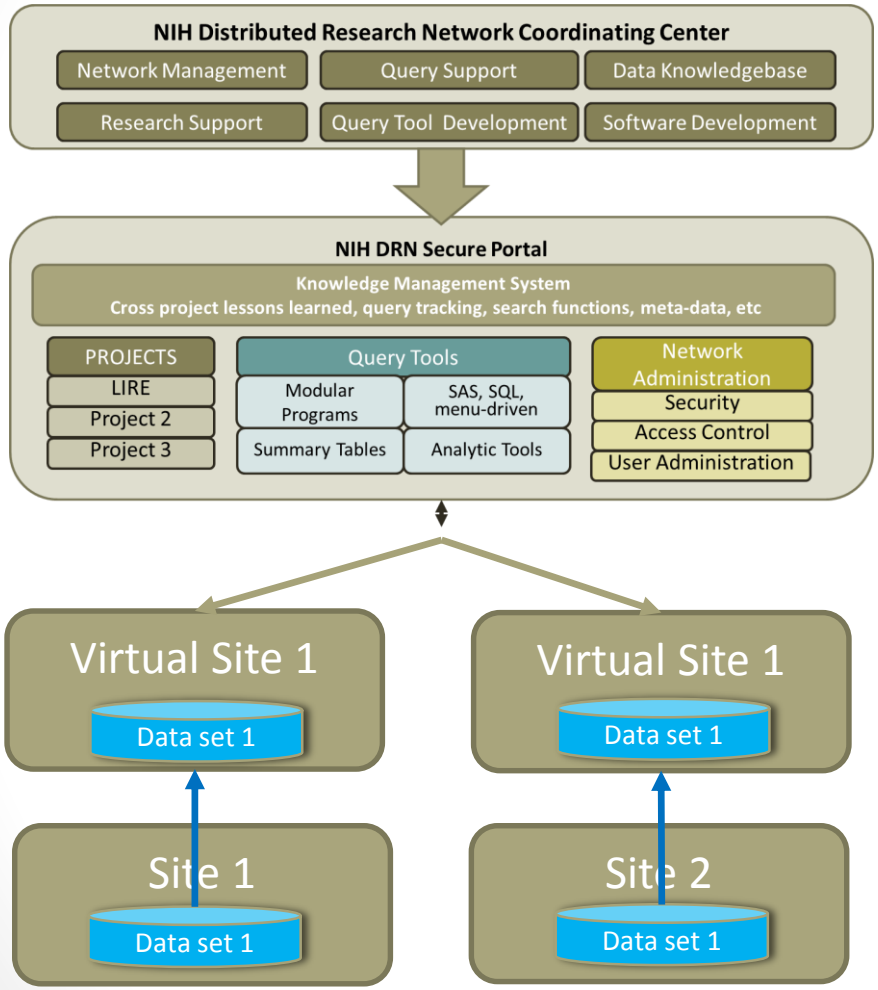
Governance, access controls, infrastructure, etc

Sites give direct access (eg, VPN) to their data source; no need to go through secure portal

Sites control VPN, local staff apply governance policies as part of access agreement

Distributed data sharing options

Data stored externally, 3rd party storage



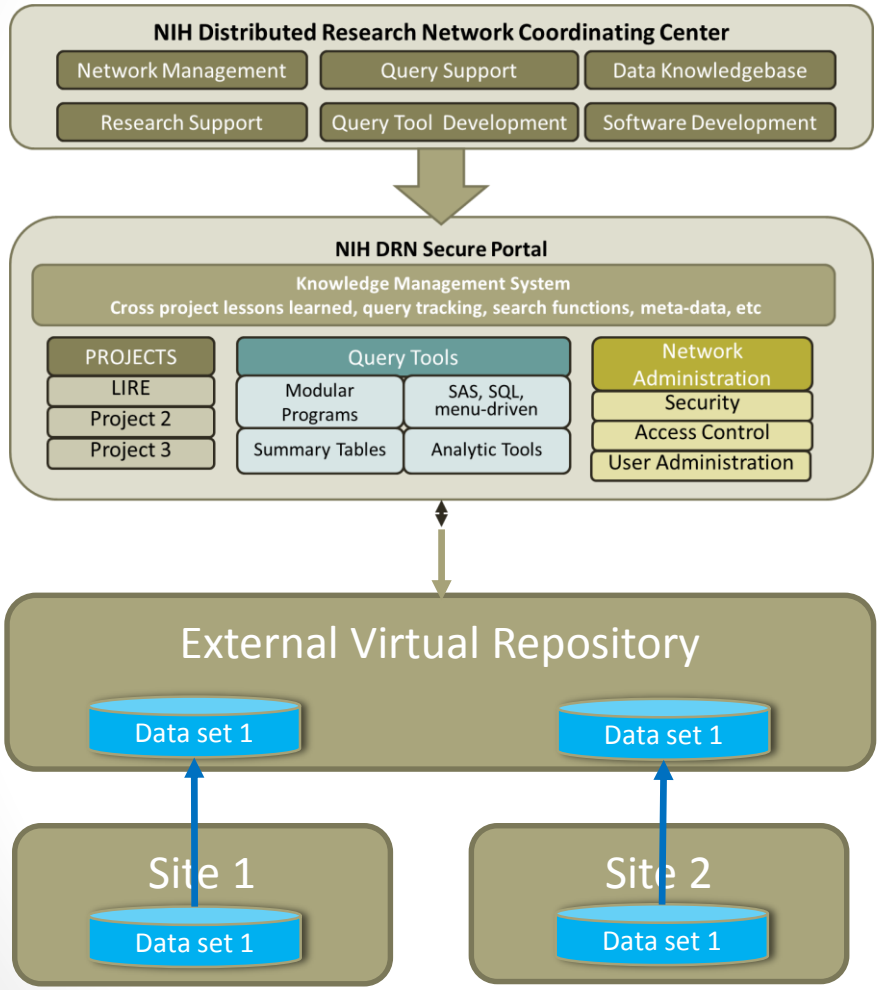
Governance, access controls, infrastructure, etc

Queries sent to virtual site, or sites give direct access (eg, VPN)

Sites send data to external location for storage, governance policies applied by site staff or a proxy

Distributed data sharing options

Data stored externally, 3rd party storage



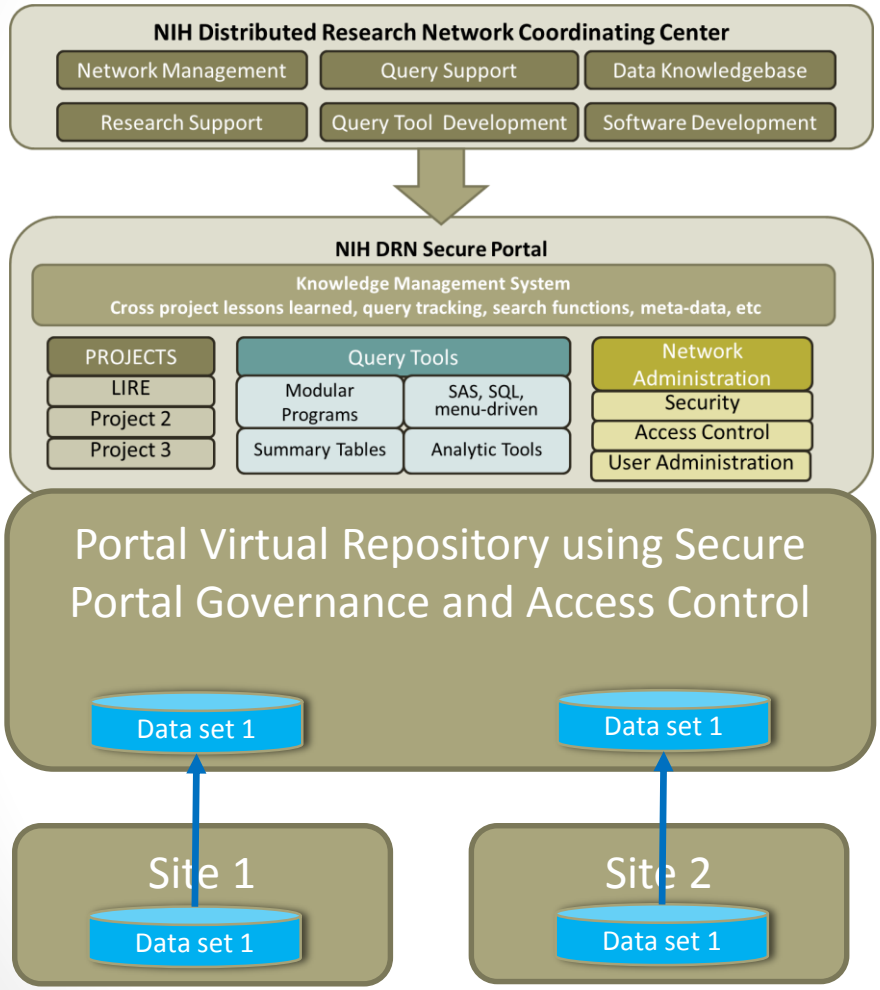
Governance, access controls, infrastructure, etc

Queries sent to virtual site, or sites give direct access (eg, VPN)

Sites send data to external location for storage, governance policies applied by site staff or a proxy

Distributed data sharing options

Data stored within NIH Collaboratory DRN secure portal



Governance, access controls, infrastructure, etc

Queries sent to virtual site within secure portal

Sites send data to secure portal for storage, governance policies applied via software and coordinating center as proxy

Key elements for a data enclave

- Discovery of available data resources and organizations
- Information about data use requirements
- Query distribution
- Secure and auditable
- Access controls and permissions
- Query interface
- Knowledge management
- Testing environment (eg, test database)
- Data storage and governance function
  - When investigators do not want to maintain local control of data source

Discovery: Data source metadata

- ClinicalTrials.gov ID#
- Access and use restrictions
- Data dictionaries, documentation, analytic code
- Publications based on dataset
- Tools available for querying or using the dataset
- Availability of TEST dataset
- Contact info for data steward
- Governance

Advantages of enclaves for data sharing

- Data sets that could not be shared externally due to privacy and proprietary concerns can be used for research
- Enables research community to confirm and extend analyses, and propose new uses of data that would otherwise not be available
- Infrastructure efficiencies
- Build a community of tools and researchers

NIH Collaboratory DRN can currently support these data sharing needs

- Platform enables re-use of research dataset with appropriate controls for patient privacy, access, governance, and proprietary concerns
- Distributed analyses limited to the software/hardware capabilities of the enclave
- Governance over usage must be established and implemented for each resource
  - Review committees? Policies?
- Oversight, maintenance, and development costs

NIH Collaboratory

Health Care Systems Research Collaboratory

The NIH Collaboratory

Discovery and Sharing of Data Resources Using Existing Tools and Infrastructure

Jeffrey Brown, Lesley Curtis, and Rich Platt

Special thanks to Greg Simon
and Rob Califf