NIH Collaboratory

Health Care Systems Research Collaboratory

Distributed Networking

Millions of people. Strong collaborations. Privacy first.

Jeffrey Brown, Lesley Curtis, Richard Platt
Harvard Pilgrim Health Care Institute and Harvard Medical School
Duke Medical School

March 15, 2013

The goal

- Facilitate multi-site research collaborations between investigators and data stewards by creating secure networking capabilities and analysis tools

Not the goal

We will **not** create a new stand-alone network with its own research agenda or content experts

Investigators will **not** have access to data without data stewards' active engagement

Reminder: Mini-Sentinel's foundation

- Strong collaborations between investigators and data partners
  - Creation of a community of trust with shared goals, backed by clear governance policies
  - Data partners' participation as collaborators
  - Data partners' voluntary participation on a case-by-case basis

The NEW ENGLAND JOURNAL of MEDICINE

February 10, 2011. Volume 364: 498-9

Perspective

Developing the Sentinel System — A National Resource for Evidence Development

Rachel E. Behrman, M.D., M.P.H., Joshua S. Benner, Pharm.D., Sc.D., Jeffrey S. Brown, Ph.D., Mark McClellan, M.D., Ph.D., Janet Woodcock, M.D., and Richard Platt, M.D.

The Food and Drug Administration (FDA) now has the capacity to "query" the electronic health information of more than 60 million people, posing specific questions in order to monitor the safety of approved medical products. This information to answer additional convening an ongoing series of discussions among stakeholders to address the near- and long-term challenges inherent in implementing the Sentinel System.[3] In 2009, the FDA gave the Harvard Pilgrim Health Care Institute the lead role

Use case: Assess disease burden/outcomes

- An NIDDK program officer wants to characterize the use and outcomes of insulin pumps for diabetes
- The Collaboratory networking center uses pre-existing ("canned") programs to query electronic data from millions of people to assess:
  - Frequency of use
  - Characteristics of the users (age, sex, prior treatment history)
  - Frequency of selected outcomes before and after initiation of use

Use case: Pragmatic clinical trial design

- Investigators planning a multi-center pragmatic trial of stroke prevention regimens want to assess the feasibility of embedding a clinical trial in care settings
- The Collaboratory networking center queries electronic health data to :
  - Assess baseline hospitalization rate with a stroke diagnosis
  - Identify organizations with enough potential study participants
  - Identify potential study participants – all identifiable information stays with the host organization

Use case: Pragmatic clinical trial follow up

- Investigators conducting a multi-center pragmatic trial of stroke prevention regimens want to simplify follow up
- The Collaboratory networking center supports clinical organizations' periodic scans of their electronic data covering study participants to identify
  - Dispensing of prescription medications, including dates, names, and amounts dispensed
  - All inpatient and ambulatory medical encounters, with dates and diagnoses and procedures

Use case: Reuse of research data

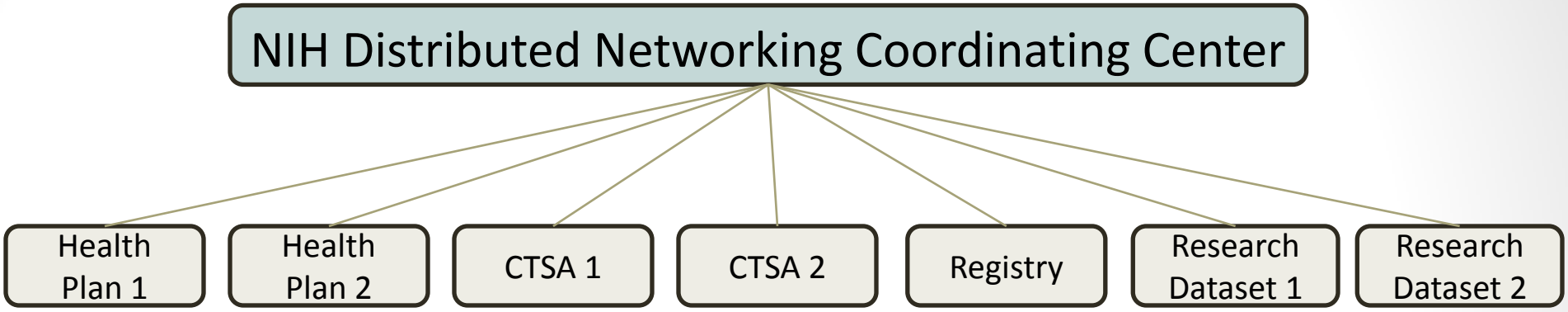
- A clinically rich research dataset of patients with incident hypertension contains longitudinal records of all blood pressure measurements, BMI, medical utilization, diagnoses, treatments, and laboratory test results
- The data steward uses the Collaboratory's networking capability to allow an investigator at another organization to submit analytic programs
- The output does not contain direct identifiers

Use case: Single study private network

- A multi-center pragmatic trial team wants to create a pooled final analysis data file
- The Collaboratory networking center establishes a private distributed network
  - To distribute programs that create separate analysis files at each site
  - To securely transfer the analysis files to the analyst

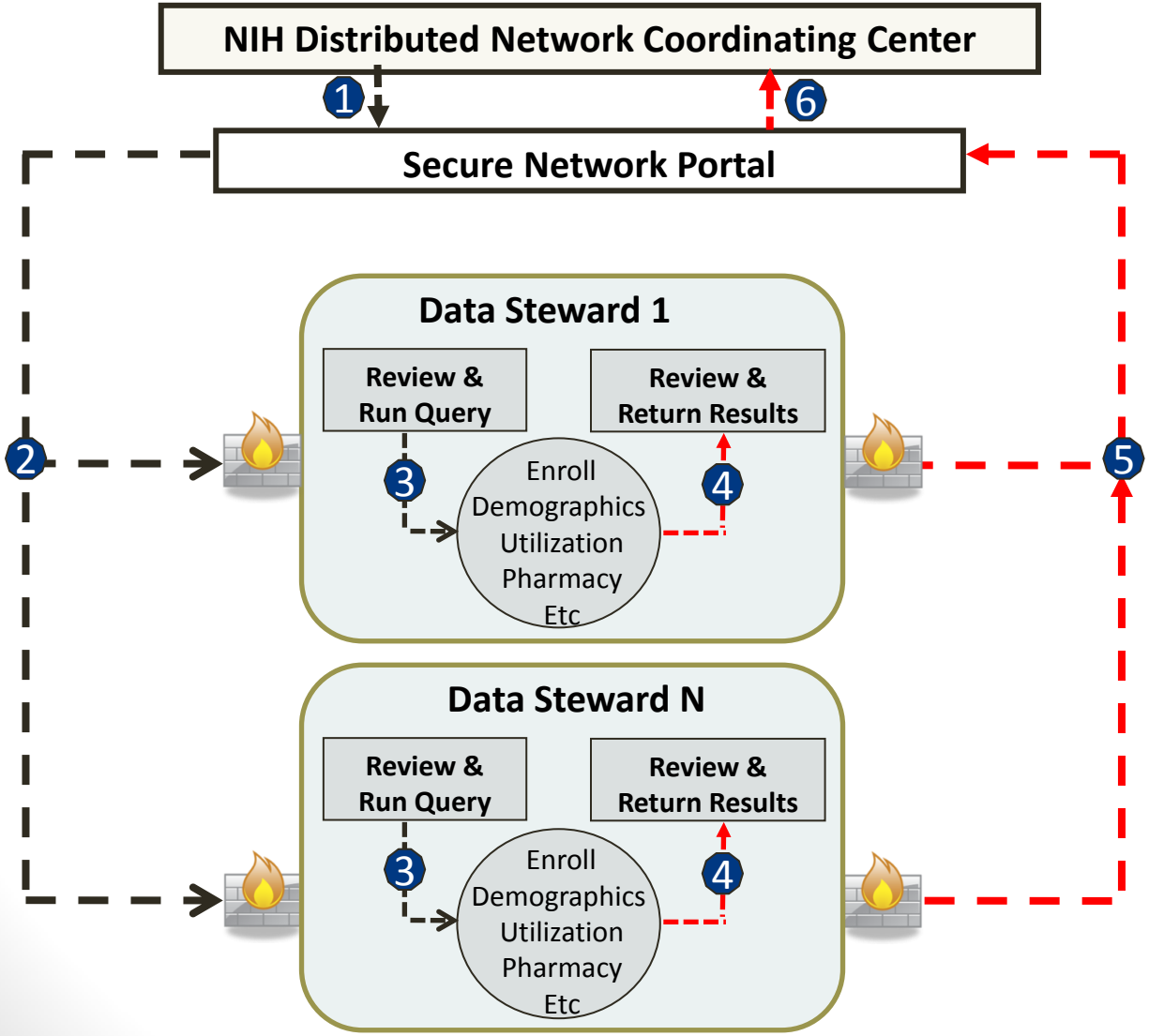
Benefits

- Assessing disease burden
  - New capability, speed, low cost, privacy protection
- Trial design / follow-up
  - New capability, speed, low cost, privacy protection
- Reuse of data
  - HIPAA compliance
    - Avoids need to create limited or de-identified datasets
    - In some cases, full datasets are more useful
  - Data sharing
    - Avoids need for some data use or business associate agreements
    - Preserves clinical organizations' sharing restrictions
- Private network
  - Secure access, auditable procedures



- Leverages existing networks' data and analysis tools
  - Can use many data types, e.g., EHR, claims, registries
  - Can use many data models, e.g., Mini-Sentinel, i2b2, OMOP
  - Can use existing querying tools, e.g., Mini-Sentinel modular programs
- Every use requires the agreement of the data steward

What is a distributed research network?



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

2- Data stewards retrieve query

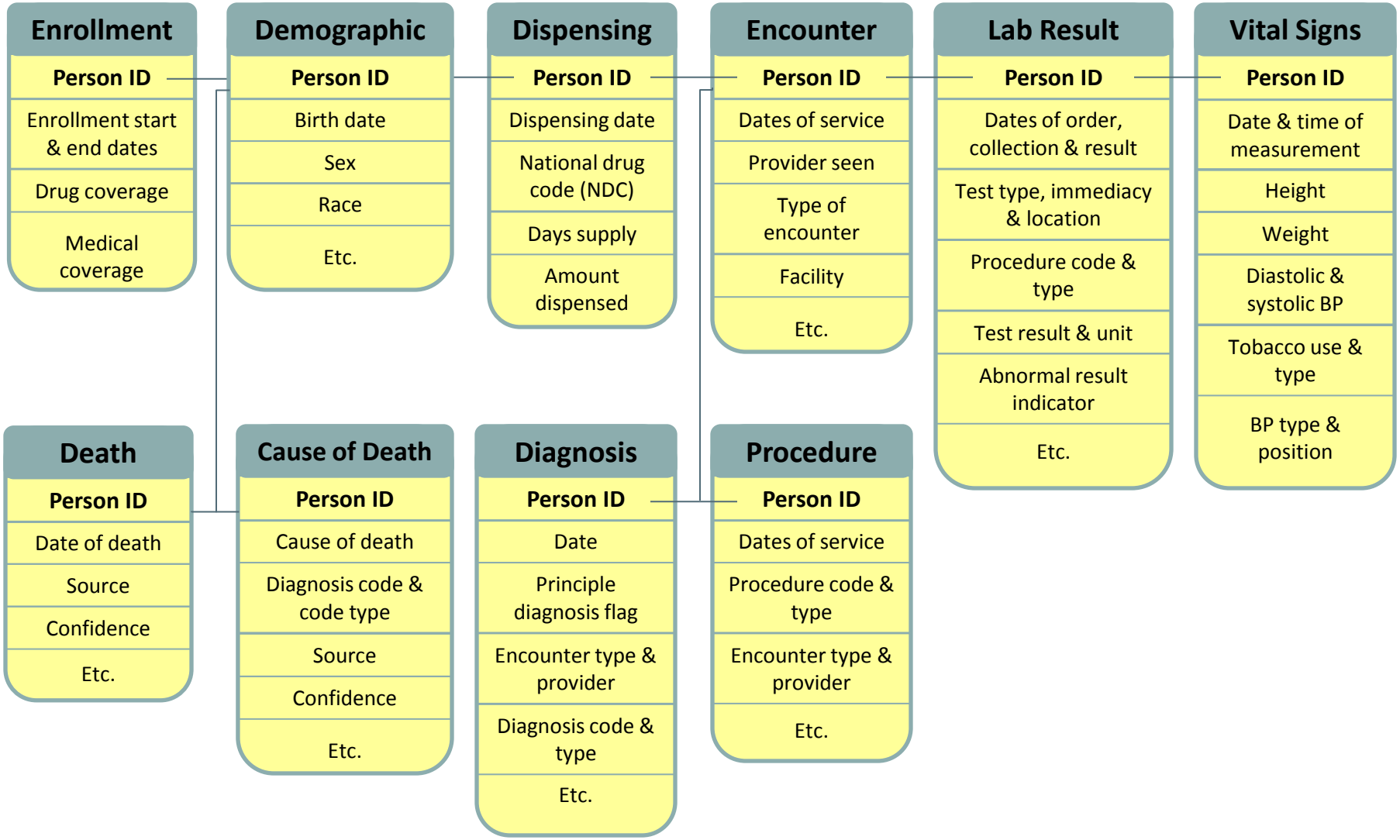
3- Data stewards review and run query against their local data

4- Data stewards review results

5- Data stewards return results via secure network

6 Results are aggregated

Mini-Sentinel's Common Data Model



Mini-Sentinel's distributed dataset data checks

- ~400 data checks per refresh
- 100+ tables per data partner per refresh

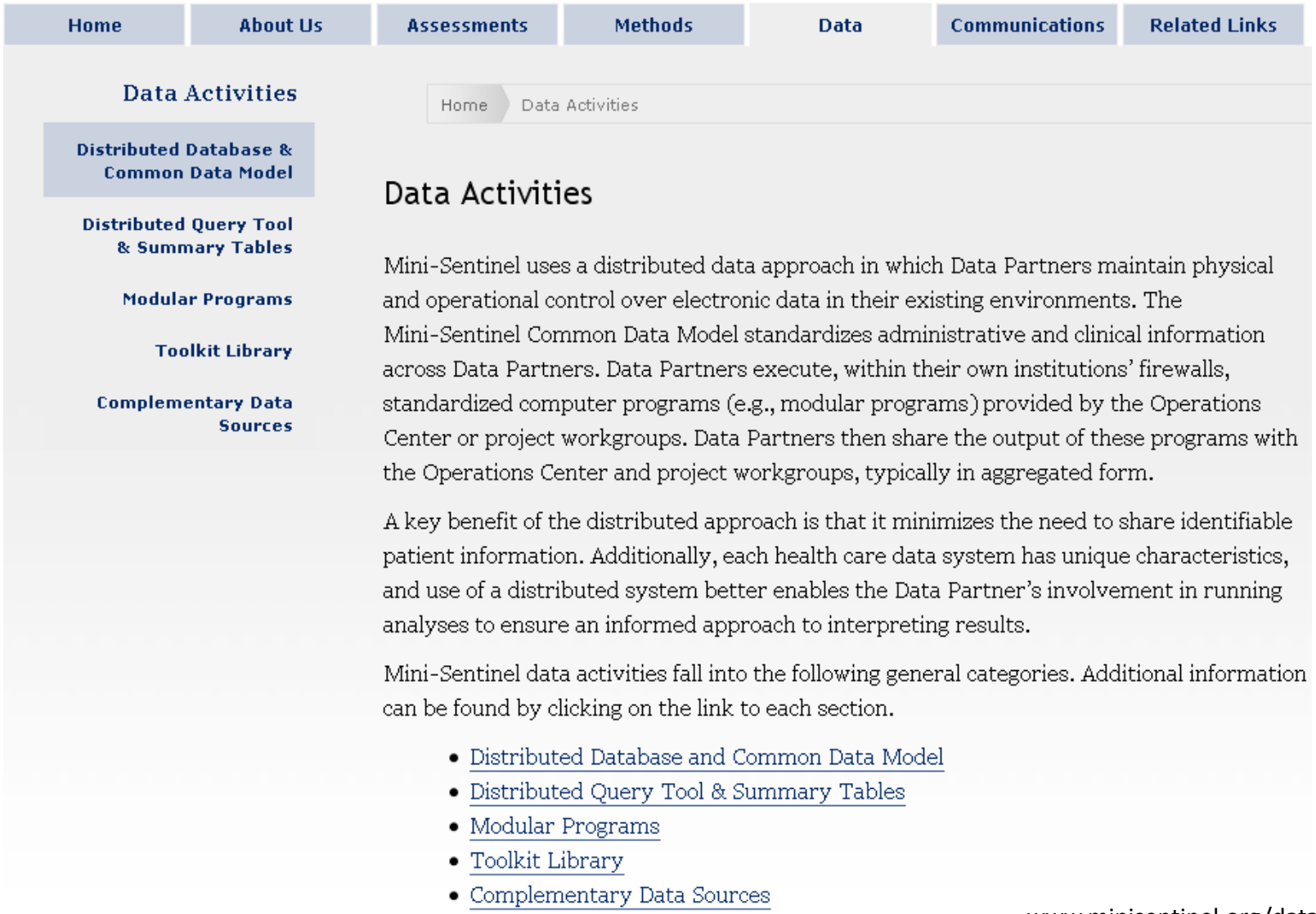
Obs	ENCTYPE	ADATE	COUNT	PERCENT
1	AV	2000	7030952	5.1370
2	AV	2001	7454699	5.4466
3	AV	2002	8014346	5.8555
4	AV	2003	8261199	6.0358
5	AV	2004	8251011	6.0284
6	AV	2005	8857635	6.4716
7	AV	2006	9576674	6.9969
8	AV	2007	10240959	7.4823
9	AV	2008	11831682	8.6445
10	AV	2009	13785025	10.0716
11	AV	2010	14499322	10.5935
12	AV	2011	14988289	10.9508
13	ED	2000	193108	0.1411
14	ED	2001	213180	0.1558
15	ED	2002	231296	0.1690
16	ED	2003	232122	0.1696
17	ED	2004	230756	0.1686
18	ED	2005	266406	0.1946
19	ED	2006	291381	0.2129
20	ED	2007	314060	0.2295
21	ED	2008	343936	0.2513
22	ED	2009	400500	0.2926
23	ED	2010	414312	0.3027
24	ED	2011	451881	0.3
25	IP	2000	432504	0.3
26	IP	2001	477466	0.3
27	IP	2002	517710	0.3
28	IP	2003	543660	0.3
29	IP	2004	543692	0.3
30	IP	2005	587863	0.4

Obs	RXDATE	N
1	2000JAN	75816
2	2000FEB	68872
3	2000MAR	240058
4	2000APR	248527
5	2000MAY	261254
6	2000JUN	258289
7	2000JUL	241145
8	2000AUG	260316
9	2000SEP	252799
10	2000OCT	260813
11	2000NOV	254161
12	2000DEC	259611
13	2001JAN	275314
14	2001FEB	242270
15	2001MAR	278558
16	2001APR	260591
17	2001MAY	268647
18	2001JUN	267520
19	2001JUL	257699
20	2001AUG	279320

Obs	Age_group	COUNT	PERCENT
1	0.1 0-1 Yrs	602059	1.4996
2	02. 2-4 Yrs	1376997	3.4298
3	03. 5-9 Yrs	2553188	6.3595
4	04. 10-14 Yrs	2638462	6.5719
5	05. 15-18 Yrs	2135457	5.3190
6	06. 19-21 Yrs	1670742	4.1615
7	07. 22-44 Yrs	14770481	36.7906
8	08. 45-64 Yrs	11221814	27.9515
9	09. 65-74 Yrs	1854092	4.6182
10	10. 75+ Yrs	1324163	3.2982

Obs	px_codetype	enctype	COUNT	PERCENT
1	09	AV	3891384	0.2061
2	09	ED	940211	0.0498
3	09	IP	7716848	0.4088
4	09	IS	168596	0.0089
5	09	OA	510196	0.0270
6	C2	AV	4906255	0.2599
7	C2	ED	325738	0.0173
8	C2	IP	392155	0.0208
9	C2	IS	18219	0.0010
10	C2	OA	222605	0.0118
11	C3	AV	212648	0.0113
12	C3	ED	5276	0.0003
13	C3	IP	7755	0.0004
14	C3	IS	269	0.0000
15	C3	OA	2030	0.0001
16	C4	AV	1364119936	72.2580
17	C4	ED	95271865	5.0466
18	C4	IP	50242438	2.6614
19	C4	IS	3914519	0.2074
20	C4	OA	27959691	1.4810
21	HC	AV	252901204	13.3963
22	HC	ED	14811325	0.7846
23	HC	IP	8125355	0.4304
	HC	IS	1600478	0.0848
	HC	OA	31067795	1.6457
	ND	AV	16692216	0.8842
	ND	ED	639229	0.0339
	ND	IP	147970	0.0078
	ND	IS	12924	0.0007
	ND	OA	819916	0.0434
	OT	AV	194765	0.0103
	OT	ED	374	0.0000
	OT	IP	2607	0.0001
	OT	IS	1367	0.0001
	OT	OA	348	0.0000

Ready to use tools for common data model



www.minisentinel.org/data_activities

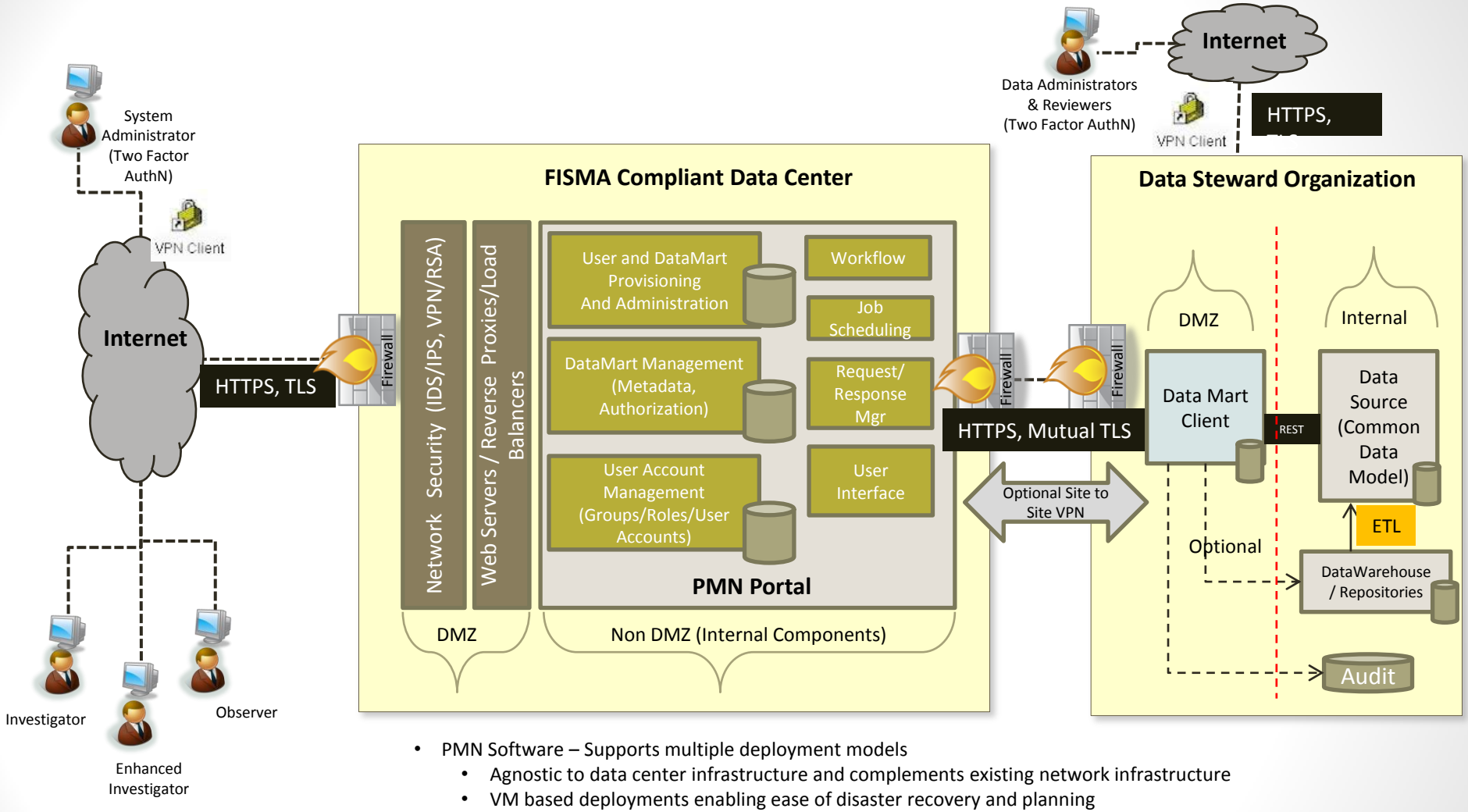
Current Networks

Data Steward	Funder				
	AHRQ		FDA	ONC	
	SPAN	PEAL	Mini-Sentinel	MDPHnet	HMORNnet
HMO Research Network (# sites in each network)	✓ (11)	✓ (4)	✓ (13)		✓ (7)
Vanderbilt		✓	✓		
Aetna			✓		
Humana			✓		
Optum (United Healthcare)			✓		
WellPoint (HealthCore)			✓		
Massachusetts League of Community Health Centers				✓	
AtriusHealth				✓	
Beth Israel Deaconess Medical Center			✓ (Query Health Pilot)		

Distributed Data / Distributed Analysis

- Data stewards keep and analyze their own data
- Standardize the data using a common data model
- Distribute code to stewards for local execution
- Provide results, not data, to requestor
- All activities audited and secure

System Architecture – Deployment Overview



- PMN Software – Supports multiple deployment models
  - Agnostic to data center infrastructure and complements existing network infrastructure
  - VM based deployments enabling ease of disaster recovery and planning
  - Seamless overlay of VPN Connections (Remote Access, Site to Site, Two Factor User Authentication)
  - Supports consolidation of remote sites into the data center for central management (Data Steward Components can be hosted in a central data center similar to the PMN Portal)
  - Secure End to End connection (Encrypted Transport using X.509 certificates)
  - Supports industry standard RBAC configuration for users
  - Supports Data Source provisioning based on RBAC and additional data source specific metadata
  - Queries distributed using a PULL model instead of PUSH model

Design Features

- **Any data model from any source**
- Flexible and secure distributed querying
  - Execution of custom analytic code
  - Menu-driven queries
- Role-based access control
- Data steward autonomy
- Query execution options range from fully automated to manual
- Auditing
- Software-enabled governance

Implementation Features

- **Secure, private multi-center research network**
- Open source application
- Data stewards maintain control of their data
- Flexible governance, access control, permissions, auditing
- Mature documentation and set-up procedures
- Scalable: easy to add new data, new partners
- Interoperable with other networks using same networking platform (PopMedNet)

Security Features

- FISMA compliant tier III data center
- 3rd-party secure audit completed
- Passed multiple independent security audits and penetration tests

National Standards

- The networking platform (PopMedNet) is a key component of the ONC's QueryHealth Initiative
- ONC national standard for distributed querying
  - QueryHealth Initiative uses PMN as the distributed querying platform for policy and governance
- Standards & Interoperability (S&I) Framework: http://wiki.siframework.org/Home

Governance (proposed)

- Data stewards retain control of their data
- All activities are opt-in
- Data stewards can choose to be full partners in the design and implementation of research
- Data steward costs must be reimbursed
  - Includes amortizing cost of maintaining data in query-able form
- TBD: A board of representatives to engage NIH leadership

Operations

- Each data steward designates a single contact for new queries
- Each data steward uses its own process for deciding whether to participate in any activity

Fine print

- Current resources will support ~20 sites
- Using existing data resources is fast; developing new ones is slow
  - Most current resources have extensive claims data, and limited EHR data
- Using existing analysis tools is fast; developing new ones is slow
- Ability to query multiple sites requires
  - Each site's data to be in the same format
  - Consistent definitions of variables

Timeline

- General querying capability begins July 2013 for organizations participating in existing networks

Thank you!