health care systems research network

What it's good for, how it works, and how to use it in your research.

Introduction to the Virtual Data Warehouse

Aung Oakkar, KPMAS
Jennifer Robertson, HealthPartners
Sundeep Basra, KPMAS
Yonah Karp, KPWHRI

health care systems research network

Learning Objectives

- What the VDW is & isn't
- Why it exists & its history
- How it works
  - Technically
  - Functionally
  - Building and harmonization
- Workgroups & SIGs
  - Engage with the implementer & user communities
- Where it's documented
- Common pitfalls
- How it's governed
- Q & A / free-form consulting

health care systems research network

The VDW is…

- A collection of:
  - Data standards
  - Library code
  - Automated processes
  - In place at each of the 19 HCSRN sites
- This collection allows programs written at one HCSRN site:
  - To be run at all the other sites
  - Easily
  - With a bare minimum of site-specific customization.

health care systems research network

The VDW is *Not…*

- A centralized database
  - There's **no** repository where you can access all data at once
  - Individual-level data **stays** at originating sites
  - Summarized, de-identified data is **shared** with the **lead** site on a study

- A means for fully automating data-based research
  - VDW is *very* human-mediated

- A replacement for the local data you already know well at your site.
  - It **uses** a lot of this data, which is built into the VDW's detailed and specific data model.

health care systems research network

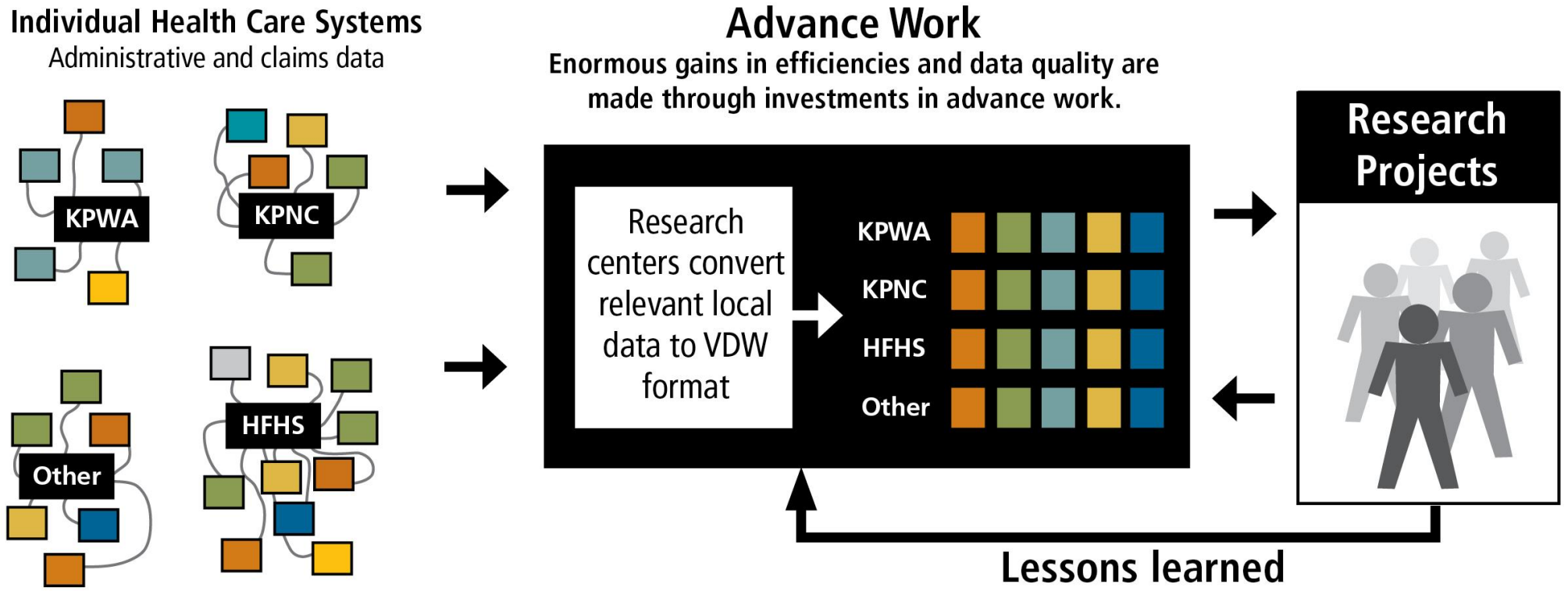
Why does it exist?

- In the earliest days of the (then) HMO Research Network (HMORN), member organizations had no data systems in common.
- Sites collaborating on research hired programmers at each participating site to implement a data pull specification from source systems.
- Post-project: discard the data (!)
- Around 2002 the Cancer Research Network (CRN) project uses Infrastructure $ to develop and build out data standards

health care systems research network

How It Works—Technically

The Virtual Data Warehouse:
A method for standardizing and pooling electronic health data for multi-site research



health care systems research network

How It Works—In Practice

- There's no formal process for taking in inquiries/solicitations. Nearly everything is generated from inside the HCSRN.
- The process is very human-mediated:
  - Investigators usually have the inspiration (and means, i.e.: grant $) for the research they want to conduct.
  - Then they usually need project groups sites
    - Project manager
    - Compliance expert
    - Programmer(s)
- Advice: find like-minded investigators at the sites you'd like to collaborate with and engage with them early.

Scientific Interest Groups (SIG)

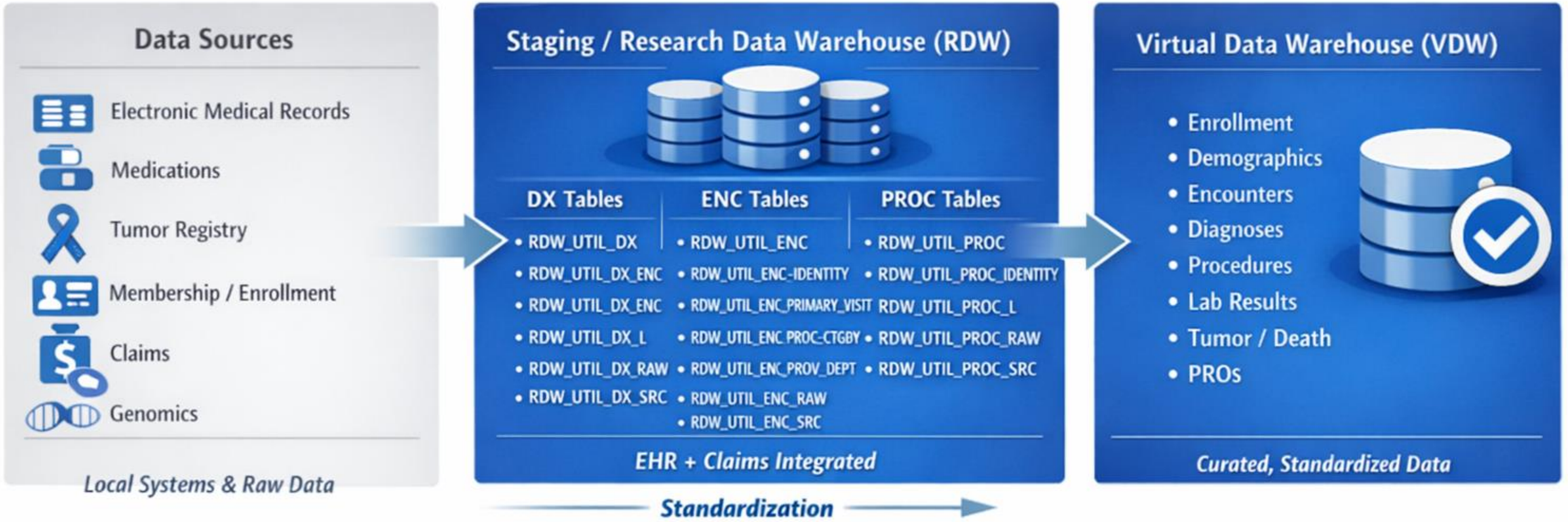
- AI in Health Care
- Aging
- Infectious Disease
- Patient Engagement In Research
- Pharmacy
- Mental Health Research Network
- Cancer

HCSRN Scientific Interest and Work Groups:

https://www.hcsrn.org/scientific-interest-and-work-groups

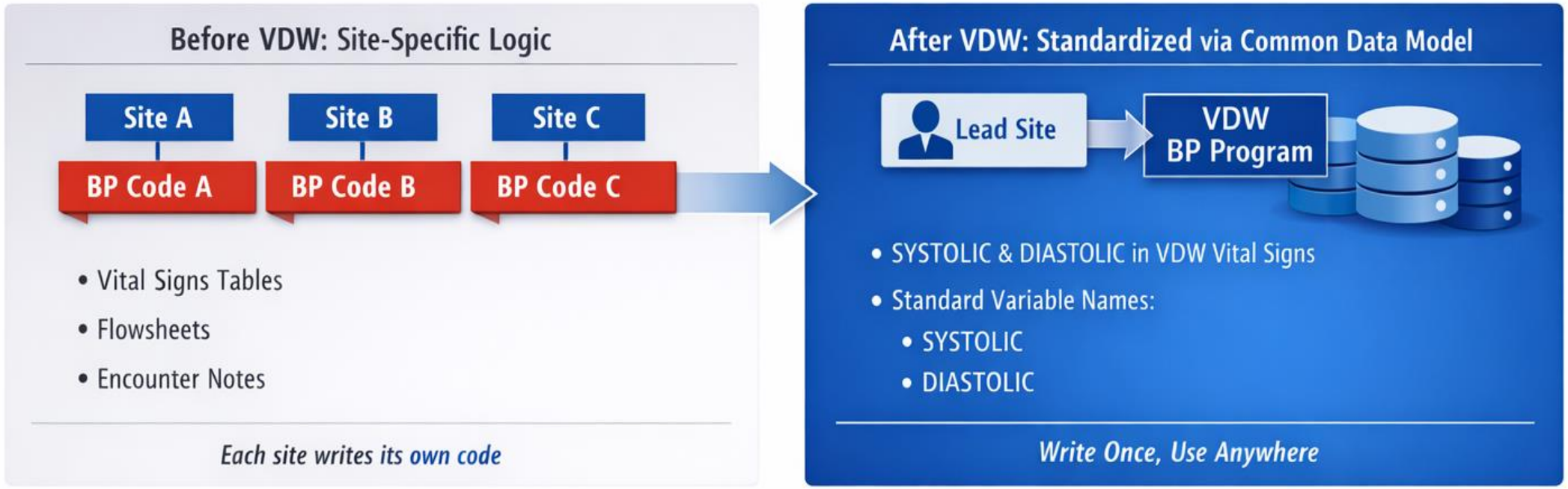
health care systems research network

Building Common VDW Data Warehouse



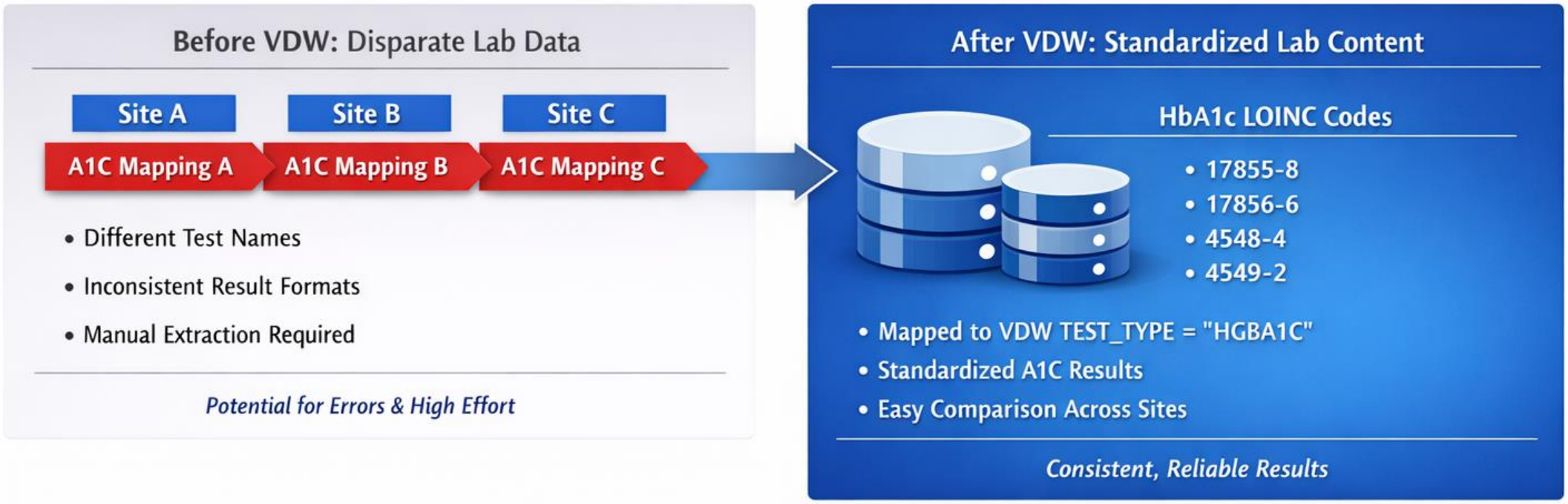
health care systems research network

Harmonization Example 1: Blood Pressure



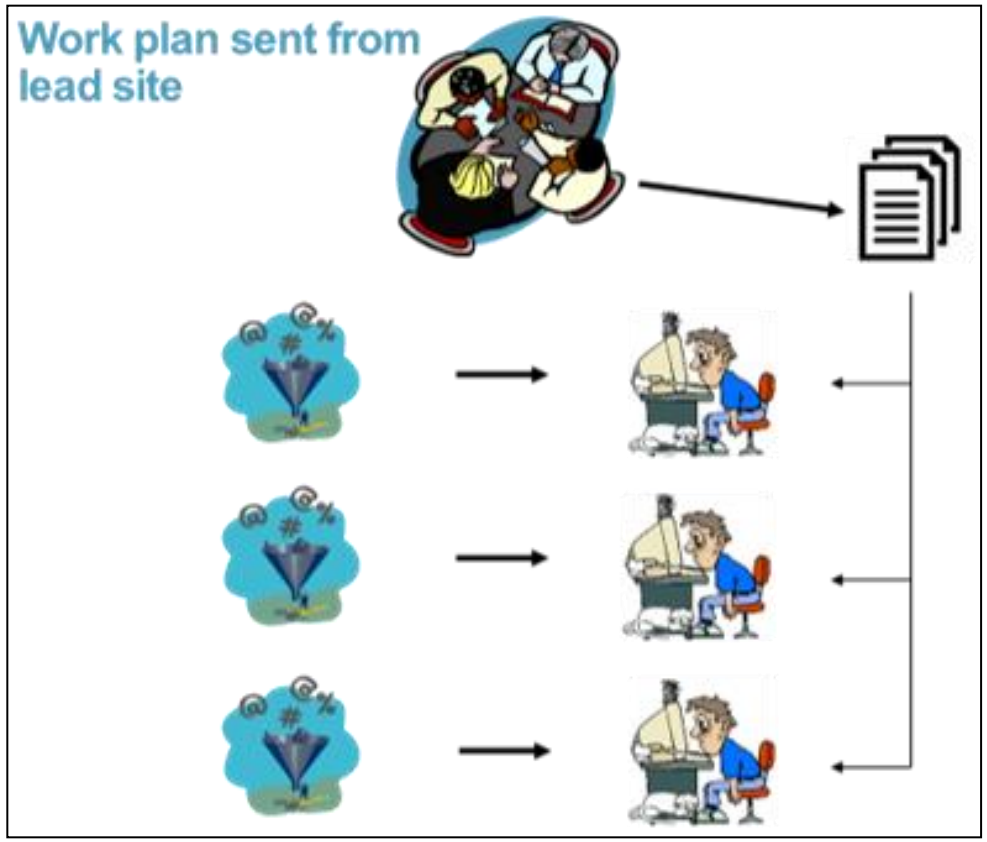
health care systems research network

Harmonization Example 2: HbA1c

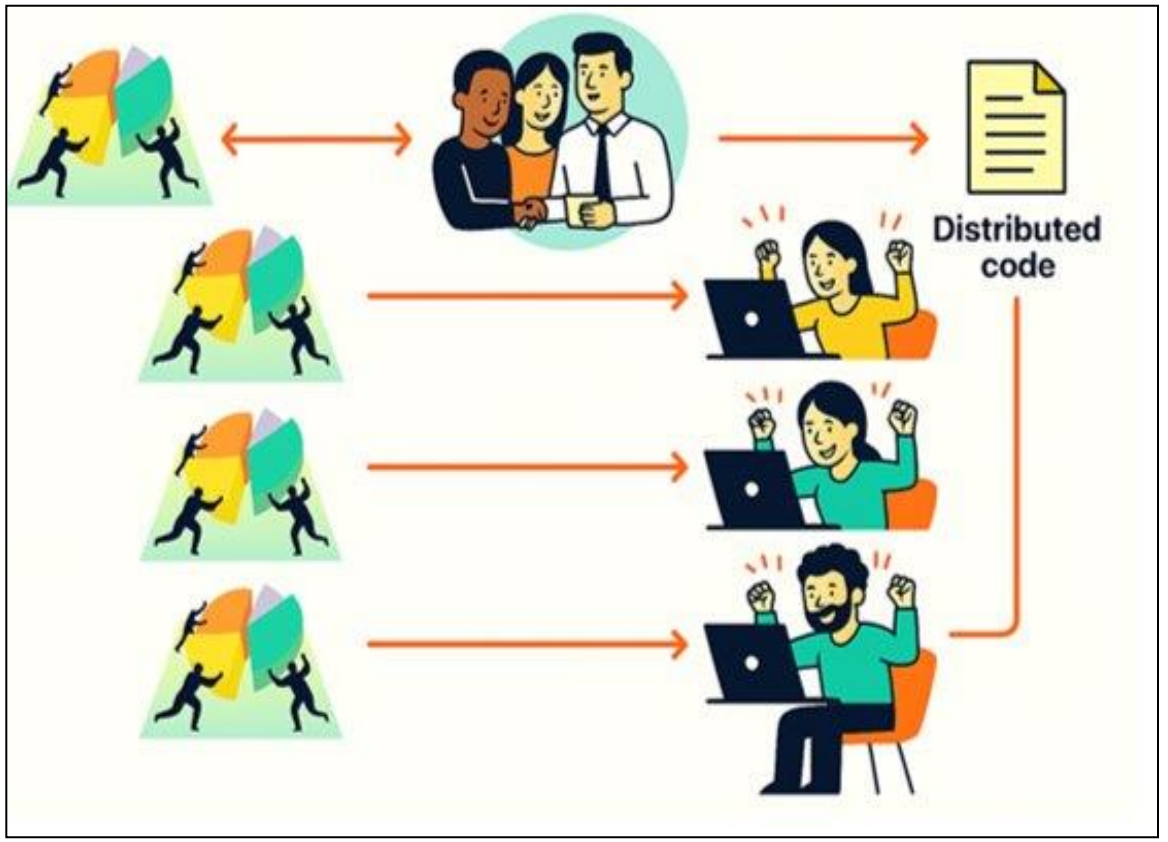


Optimized State

Instead of this...

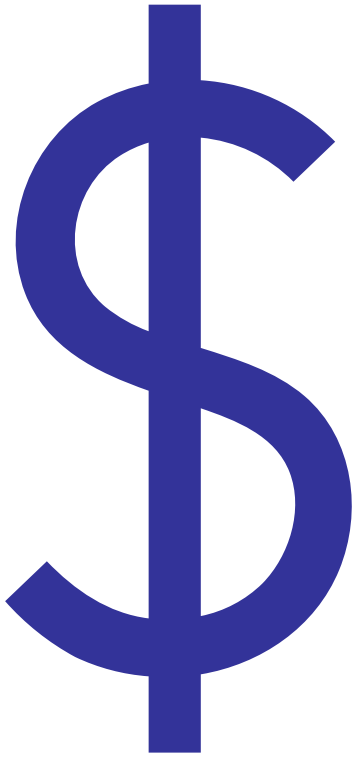


We have this!



health care systems research network

Cost of a research project pre-VDW



- Redundant
- Ad-hoc
- Potentially poorer quality
- Differing interpretations of workplan
- Data lessons lost/siloed
- Complicated sources
- Data discarded at end of project

= EXPENSIVE

health care systems research network

Why use, develop, and enhance the VDW?

- Data already curated from source systems to common data model standards and is updated regularly at each site
- Quicker to learn how to use the data
- Easier to write programs for studies
- Highest FTE at lead site, reduced FTE at contributing sites
- Standardized and tested macros and algorithms are shared across HCSRN for QA and routine data tasks

health care systems research network

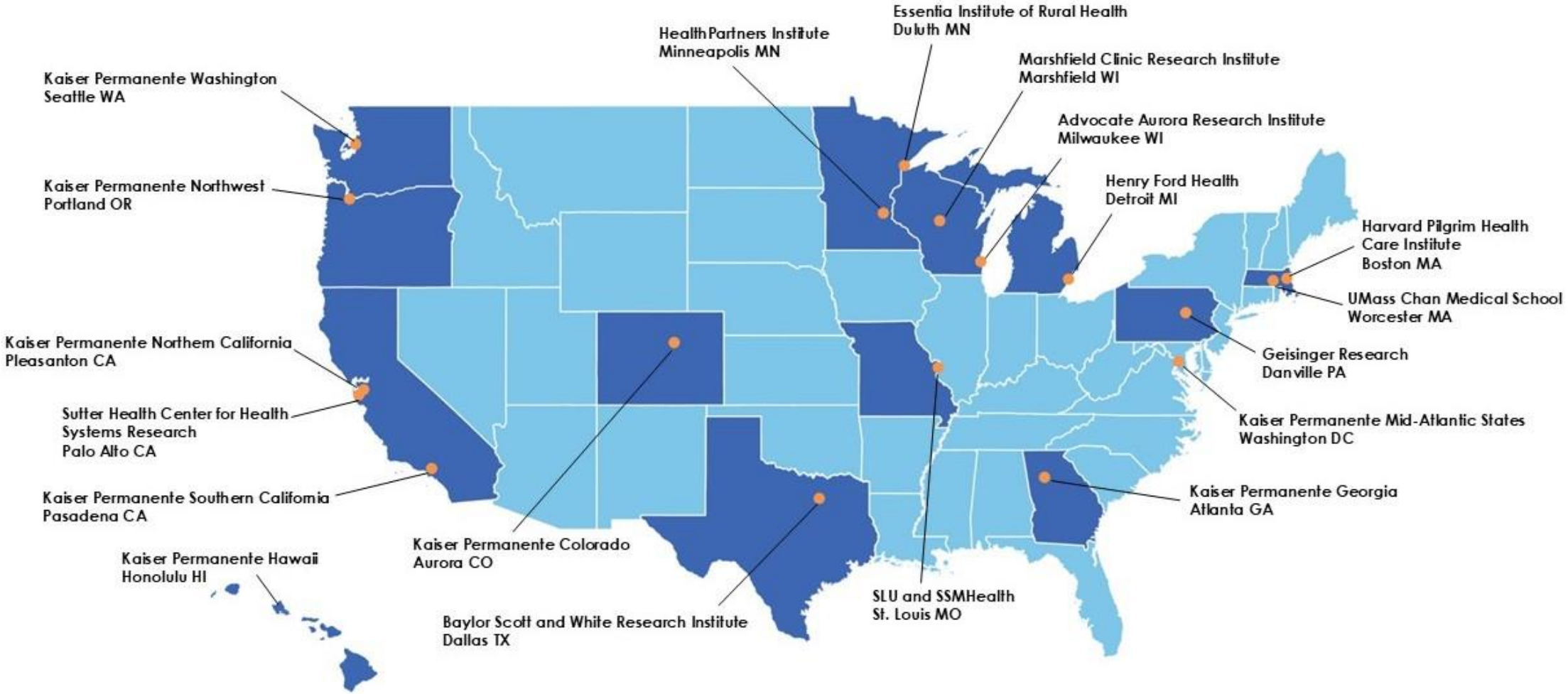
Why use, develop, and enhance the VDW?

- Scientific validity
  - Replication
  - Common definitions and decisions across projects and people
  - Standardization improves scientific rigor by improved consistency and reproducibility.
  - Standard code library to draw on (e.g. comorbidity indices, definitions of continuous enrollment)

At its best, the VDW is a *vehicle for preserving & cumulating the benefits/fruits of project-specific work.*

health care systems research network

HCSRN Sites



health care systems research network

What Data Do We Have?

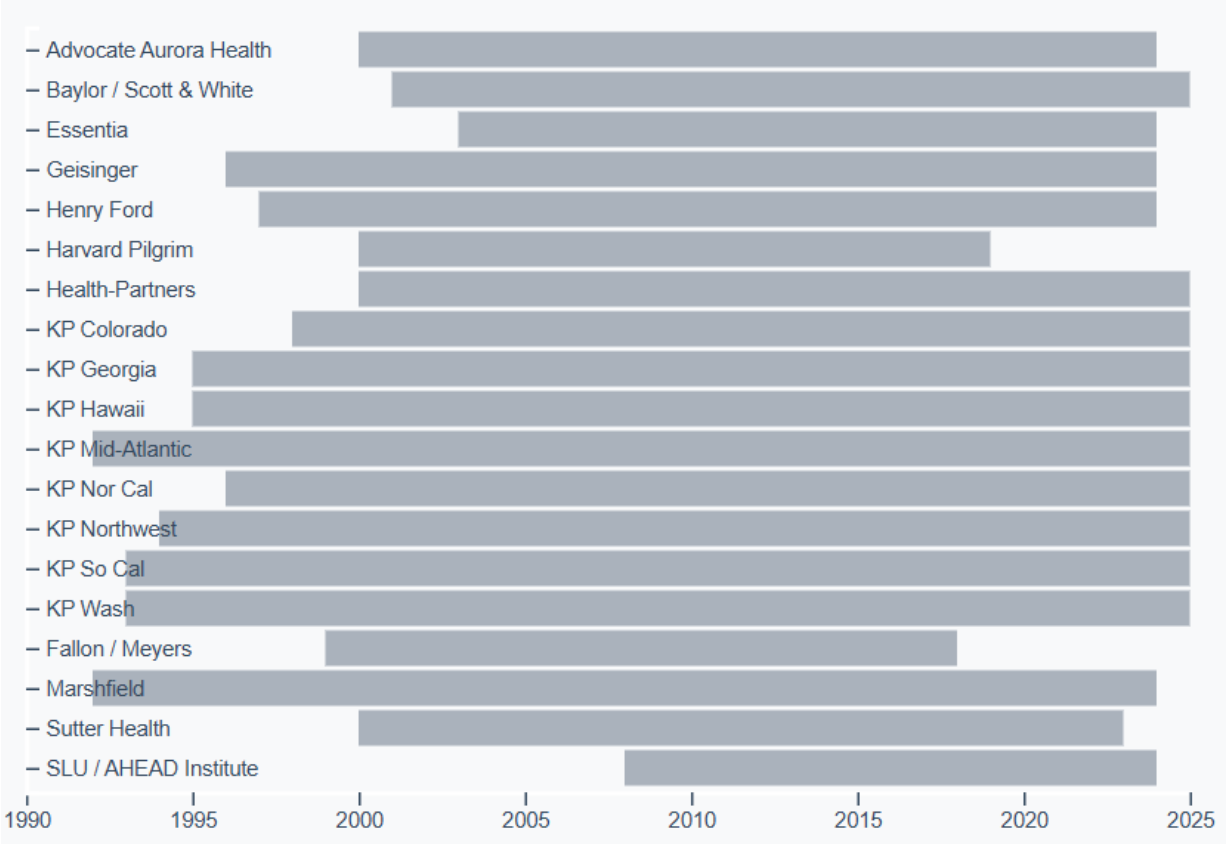
- We have 19 Active Organizations, all of which have implemented 1 or more VDW tables.
- There are 10 general content areas (comprising 26 distinct tables).

1. Enrollment/Demographics (3)
2. Utilization (6)
3. Pharmacy (2)
4. Lab Results (2)
5. Vital Signs (1)
6. Death (2)
7. Social History (1)
8. Tumor (2)
9. Census (3)
10. Patient Reported Outcomes (4)

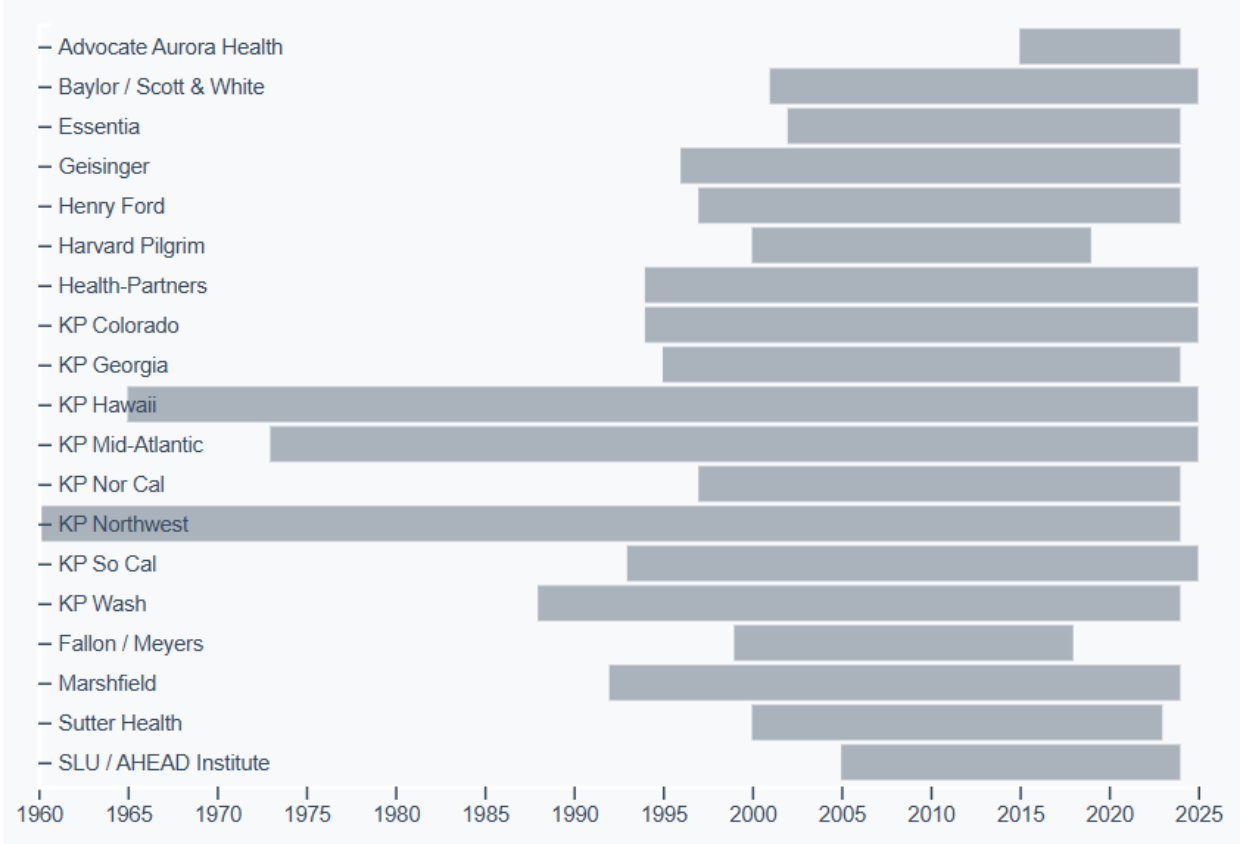
health care systems research network

Different Sites' Implementations Vary in Extent

Encounters



Enrollments



health care systems research network

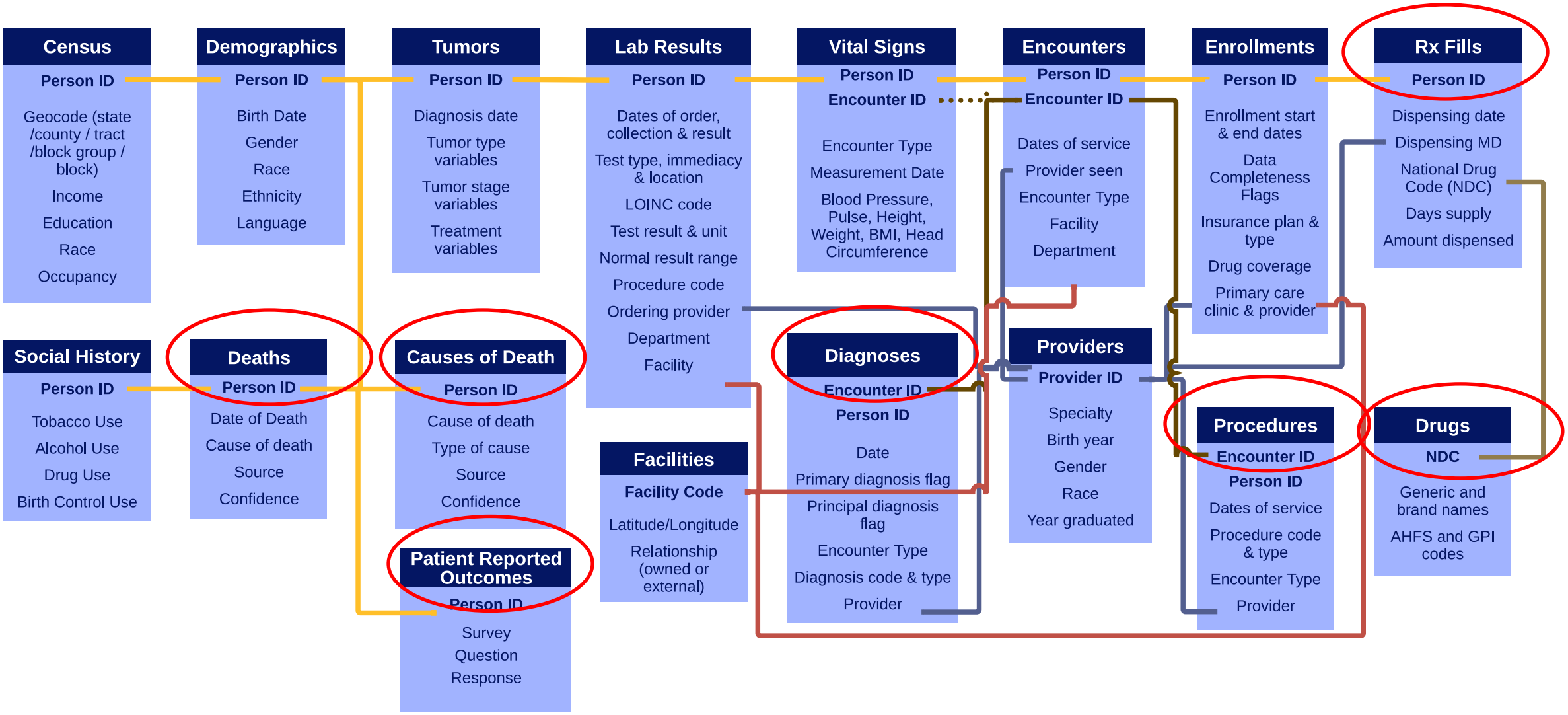
Implementation Overview Dashboard

Implementation Overview

Data Area	Baylor / Scott & White	Essentia	UMass	Geisinger	Harvard Pilgrim	Health-Partners
Demographics Patient/Enrollee Lookup	2001 - Present Q	2002 - 2024 M	1999 - 2018 SA	2001 - 2024 Q	2000 - 6/2019 A	1990 - Present M
Languages Patient Languges	2001 - Present Q	N/A - N/A M	1999 - 2018 SA	2001 - 2024 Q	2000 - 6/2019 A	2004 - Present M
Enrollments Records of Medical Coverages or Patient Affiliations	2001 - Present Q	2002 - 2024 M	1999 - 2018 SA	1996 - 2024 Q	2000 - 6/2019 A	1994 - Present M
Utilization Encounters between Patients and Medical Personnel	2001 - Present Q	2003 - 2024 M	1999 - 2018 SA	1996 - 2024 Q	2000 - 6/2019 A	2000 - Present M
Diagnoses Diagnoses made at Encounters	2001 - Present Q	2003 - 2024 M	1999 - 2018 SA	1996 - 2024 Q	2000 - 6/2019 A	2000 - Present M
Procedures Procedures performed at Encounters	2001 - Present Q	2003 - 2024 M	1999 - 2018 SA	1997 - 2024 Q	2000 - 6/2019 A	2000 - Present M

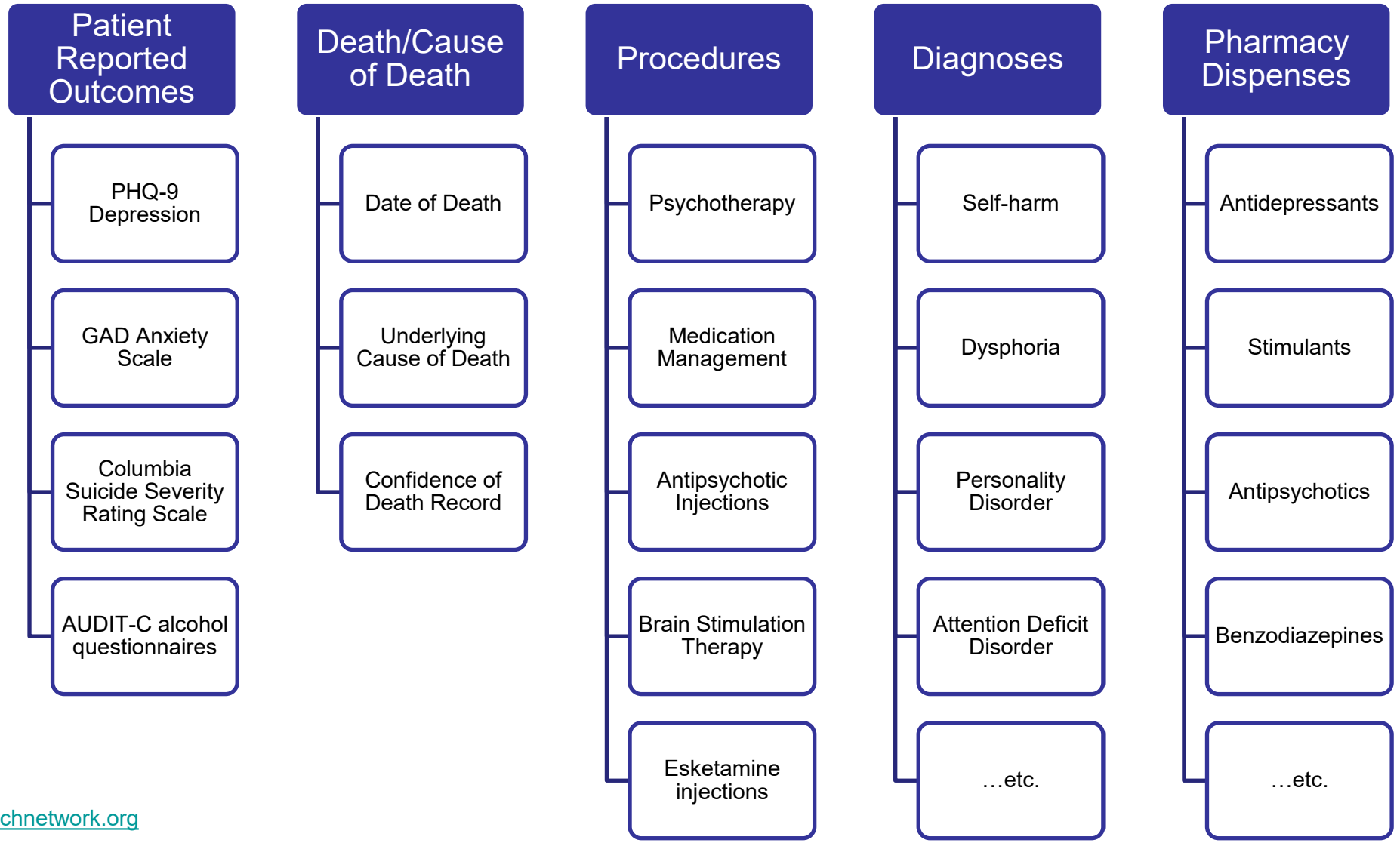
health care systems research network

Data Areas



health care systems research network

Mental/Behavioral Health Research Key tables of interest



Source: mhresearchnetwork.org

health care systems research network

Mental/Behavioral Health Resources

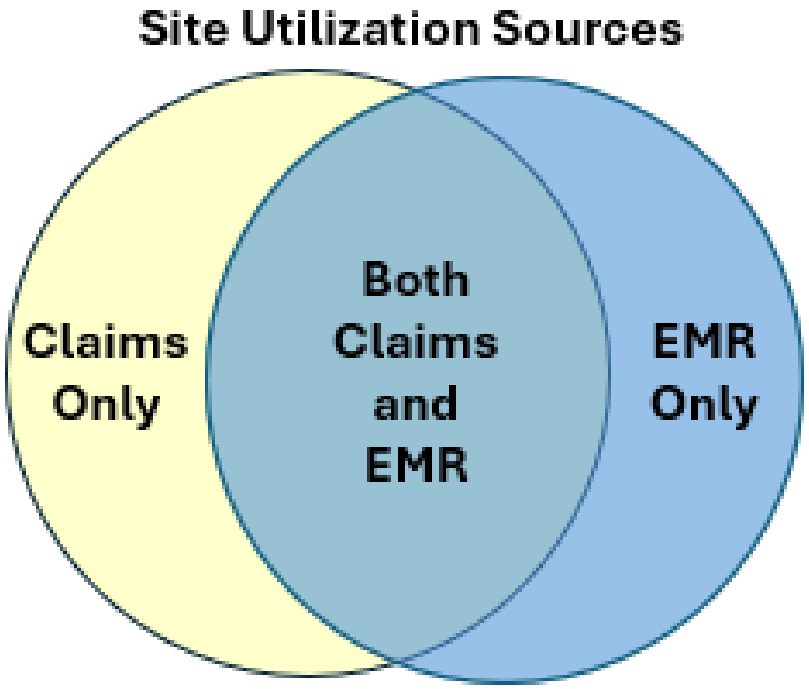
- Mental Health Research Network
  - Part of the larger HCSRN
  - Consortium of 14 research centers
  - mhresearchnetwork.org
  - Curated lists are specific to the use of the VDW and available codes

health care systems research network

We encourage sites to use "all the data that fits"

- Specs are generally laid out to accommodate all compatible data sources, including:
  - Legacy (pre-modern EMR) internal systems
  - Claims (rx, facility, professional)
  - EMR
  - Other business/administrative data as appropriate for the subject area

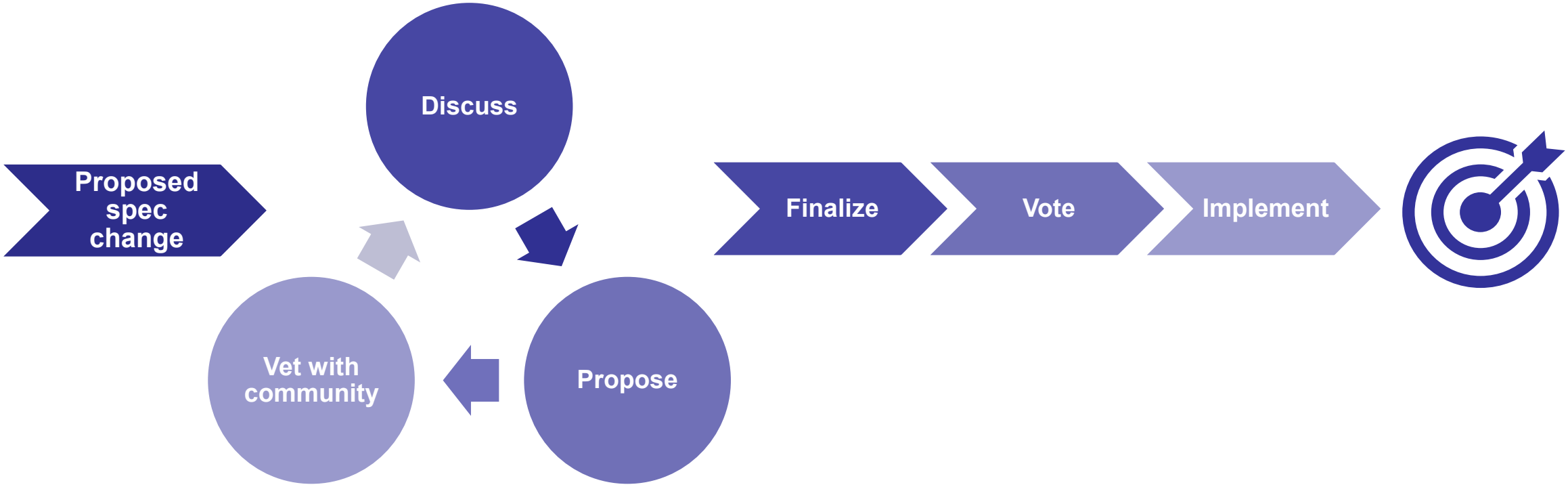
- Many specs include metadata fields to indicate provenance (e.g., SOURCE_DATA in the Encounters Table)



health care systems research network

Robust Change Management

- As research topics, methods, and workflows evolve, so must the VDW
- The VDW works **as a collaborative at all levels** to understand needs and create new specs and implementation guidelines.



health care systems research network

Change Management Comments Requested

For Scientific Users

- How do you expect these changes to affect your current program of research?
- How valuable is this information to you?
- Are the new variables proposed:
  - conceptually coherent?
  - clearly defined?
  - likely to be useful?

For Implementers

- Can you obtain the information needed for the new variables?
- Do you have other useful information in your source data that would not be well accommodated by the proposed changes?
- Are the new variables proposed
  - conceptually coherent?
  - clearly defined?
  - likely to be useful?
- Can you think of a better way of storing this data for research use?

health care systems research network

Documentation & Resources

Public-facing (hcsrn.org)

VDW RESOURCES FOR SCIENTISTS

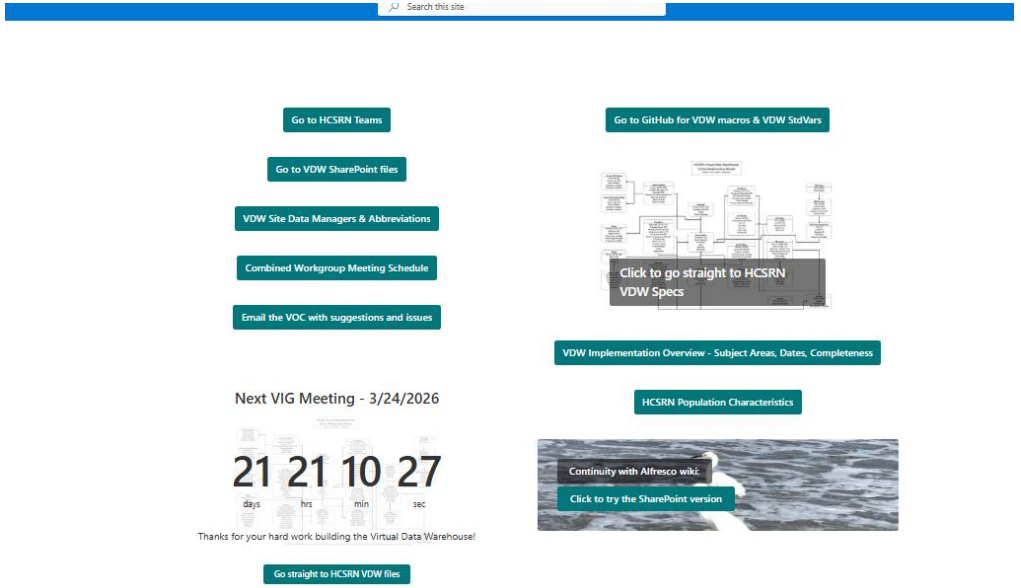
- The VDW and How to Use It – An introduction to the HCSRN Virtual Data Warehouse (VDW) for Investigators. This includes an overview of the distributed data model, its data areas and structure, and steps for using it.
- Examples of VDW Projects – A great deal of research can be conducted using only VDW data. More commonly VDW data supplements other data gathered from subjects (e.g., surveys, research specimens). A list of example projects is provided.
- Virtual Data Warehouse Q&A – Archived 2011 Virtual Data Warehouse (VDW) orientation presentated by VDW Operations Committee staff. Some content will be out of date.
- Orientation to the HCSRN and Virtual Data Warehouse – Orientation slides from the 2018 HCSRN Conference, with an overview of the HCSRN itself, and a closer look at the contents and structure of the Virtual Data Warehouse. Includes information about the distributed data network as well as the querying process and research opportunities afforded by a common data model
- Presentation: Using the Virtual Data Warehouse – Archived presentation by Cancer Research Network (CRN) staff entitled, Using the Virtual Data Warehouse (some content will be out of date)

VDW DATA MODEL AND METADATA

- VDW Data Model – Detailed data specifications for the VDW data model. Version 5
- Figure – VDW Data Structure – This figure illustrates the primary data domains of the Virtual Data Warehouse and provides some detail about variables under each domain.
- Figure – VDW Data Areas – This simple figure illustrates the primary data domains of the Virtual Data Warehouse.
- Figure – VDW Infographic – Figure depicting VDW data standardization, pooling, and user-feedback loop. Includes header with title, "The VDW and How It Works" and a contextual statement. Four site acronyms shown in left panel.

Privacy & Cookies Policy

Members-only



health care systems research network

Public Documentation

- HCSRN.org's
  - Data Resources page: Data Resources | HCSRN Collaborative Research Resources
  - VDW Data Model: VDW Data Model | HCSRN Collaborative Research Resources
- Includes:
  - Orientation to HCSRN and VDW
  - Examples of VDW Projects
  - Data Model
  - Specifications

VDW Data Model

health care systems research network

Public Documentation (cont)

- We also have several assets on github, including our standard macro library and documentation.
- Finally, we have the Implementation Overview page: Virtual Data Warehouse Implementations which shows:
  - Sites
  - Data Areas
  - How far back (and forward) in time each site's implementation goes.
  - How often each site updates its VDW files
  - Lightweight, unofficial data specifications

Implementation Overview

Data Area	Baylor / Scott & White	Essentia	UMass	Geisinger	Harvard Pilgrim	Health-Partners
Demographics Patient/Enrollee Lookup	2001 - Present Q	2002 - 2024 M	1999 - 2018 SA	2001 - 2024 Q	2000 - 6/2019 A	1990 - Present M
Languages Patient Languges	2001 - Present Q	N/A - N/A M	1999 - 2018 SA	2001 - 2024 Q	2000 - 6/2019 A	2004 - Present M
Enrollments Records of Medical Coverages or Patient Affiliations	2001 - Present Q	2002 - 2024 M	1999 - 2018 SA	1996 - 2024 Q	2000 - 6/2019 A	1994 - Present M
Utilization Encounters between Patients and Medical Personnel	2001 - Present Q	2003 - 2024 M	1999 - 2018 SA	1996 - 2024 Q	2000 - 6/2019 A	2000 - Present M
Diagnoses Diagnoses made at Encounters	2001 - Present Q	2003 - 2024 M	1999 - 2018 SA	1996 - 2024 Q	2000 - 6/2019 A	2000 - Present M
Procedures Procedures performed at Encounters	2001 - Present Q	2003 - 2024 M	1999 - 2018 SA	1997 - 2024 Q	2000 - 6/2019 A	2000 - Present M

health care systems research network

Members-Only Documentation/Resources

- HCSRN VDW SharePoint: https://hcsrnvdw.sharepoint.com/
  - The home of the Official VDW Specifications
  - E-mail addresses for leadership & workgroups
  - Reports of interest
    - Cross-site QAs (e.g., enrollment)
    - HCSRN Population Characteristics
  - QA Issue Tracker

- HCSRN VDW Microsoft Teams Instance
  - Channels devoted to different data areas
  - General channel
  - Excellent place to engage with (other) implementers, to ask questions about what you'll find at the sites.

health care systems research network

Common Pitfalls Using VDW

- Biggest is assuming that all HCSRN sites are like yours.
  - Utilization comes completely from claims, or from non-claims.
  - We only treat people we insure (or we insure people at all!)
  - Vast majority of inpatient stays are at external hospitals.
  - We have our own Tumor registry
  - We run the EPIC EMR, and have the Clarity reporting system.
  - Chemo data winds up in Procedures data, not Pharmacy.

- Assuming all variables will be populated over all time (e.g., Gender Identity & Sexual Orientation in Demographics are pretty sparse).

- Assuming uniformity of data across sites.
  - VDW does some assessment, cleaning & normalization, but not as frequent as most people assume.

health care systems research network

More pitfalls

- Assuming <<some data element you need>> is available in VDW.
  - Study the specifications!

- In general:
  - Go slow.
  - Explore.
  - Take nothing for granted.
  - Don’t hesitate to consult with Yonah and other staff.

health care systems research network

How is VDW Governed?

- The VDW Operations Committee (VOC) is charged with maintaining and setting priorities for the VDW
  - VOC includes:
    1. Board Liaison: Dr. Stacey Honda (KP Hawaii)
    2. Coordinator: Celeste Machen (KP Northwest)
    3. Technical Lead: Yonah Karp (KP Washington)
  - Meets weekly to discuss governance issues, others involved:
    - Executive Director: Deb Hamlin (Cap Hill)
    - Association Manager: Joe Moreno (Cap Hill)
- Subject Area Workgroups
  - One per data area, each with a scientific lead and a technical lead
  - Steward the relevant specifications—correct errors & consider/propose improvements.
  - Consult with implementers and users when there are questions
  - Generate QA packages and periodically collate/report results.
  - Meets monthly or bi-monthly.

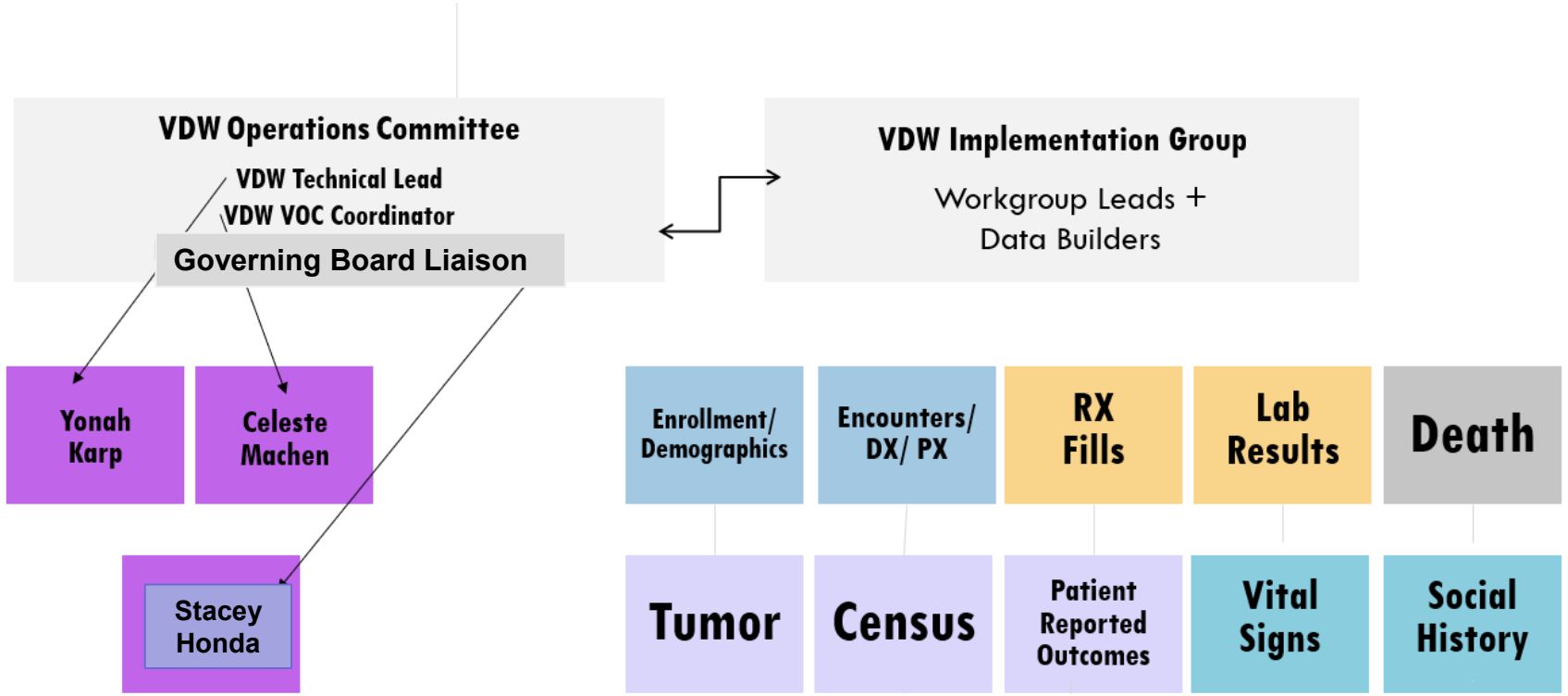
health care systems research network

How is the VDW governed? (Cont)

- Each site designates a Site Data Manager, who can answer questions about that site's implementations.
- The VDW Implementation Group (VIG) is the set of Site Data Managers and data developers from all sites
  - A working group (and occasionally, support group) for implementers.
  - Charged with vetting technical/data issues, including especially the feasibility of proposed spec changes.
- Where there is interest, new working groups can be chartered.
  - Volunteer-led, not centrally funded
  - If you hear yourself say "Someone should…"—consider that you are someone.

health care systems research network

VDW Governance Schematic



VDW MANAGEMENT

health care systems research network

Communications

- Conference Calls
  - VIG Conference Calls: Fourth Tuesdays @ 12:00-1:00 pm Pacific Time
  - Workgroups: various times
- E-Mail Lists
  - One for VIG
  - One per Workgroup
  - Email the VOC chr_vdw_voc@kpchr.org to be added
- Teams Channels
  - Also, one per Workgroup
  - Plus, other general channels
- We need your expertise!
  - Nearly all of this is volunteer effort
  - Please consider joining us on workgroups

health care systems research network

Thank You!

QUESTIONS?

health care systems research network

Spec Change Management

- Someone sees the need for a change to a VDW spec.
- The idea gets discussed on the relevant workgroup's calls.
- Someone writes a Change Proposal that
  - Sets out the problem
  - Sets out the proposed change (add/remove fields, split a table out into 2 tables, etc.)
  - Describes how the change will address the problem.
- The Proposal is posted to the MSTeams Channel, and the link e-mailed out to the VIG group, who are asked to forward to their local users & investigators.

health care systems research network

More Pitfalls

- More “like me” assumptions:
  - We run the EPIC EMR, and have the Clarity reporting system.
  - Chemo data winds up in Procedures data, not Pharmacy.
- Assuming all variables will be populated over all time (e.g., Gender Identity & Sexual Orientation in Demographics are pretty sparse).
- Assuming uniformity of data across sites.
  - VDW does some assessment, cleaning & normalization, but not as much as most people assume.