

# CBK Metadata, Computable Phenotypes, Research Networks

Rachel Richesson, PhD  
Department of Learning Health Sciences  
University of Michigan Medical School

IMLS / Mobilizing Computable Biomedical Knowledge (MCBK) Training Online  
Pilot Class

January 3, 2021

# Posted Readings/Websites for Module

- Alper, BS, Flynn, A, Bray, BE, et al. Categorizing metadata to help mobilize computable biomedical knowledge. Learn Health Sys. 2021;e10271. <https://doi-org.proxy.lib.umich.edu/10.1002/lrh2.10271>
- Richesson R, Wiley LK, Gold S, Rasmussen L. Electronic Health Records–Based Phenotyping: Introduction. In: Rethinking Clinical Trials: A Living Textbook of Pragmatic Clinical Trials. Bethesda, MD: NIH Health Care Systems Research Collaboratory. Updated July 27, 2021. <https://rethinkingclinicaltrials.org/chapters/conduct/electronic-health-records-based-phenotyping/electronic-health-records-based-phenotyping-introduction/>
- Chapman M, Mumtaz S, Rasmussen LV, et al. Desiderata for the development of next-generation electronic health record phenotype libraries. Gigascience. 2021;10(9):giab059. doi:10.1093/gigascience/giab059 <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8434766/>
- OHDSI: <https://www.ohdsi.org/>
- MCBK: <https://mobilizecbk.med.umich.edu/>
- COVID-19 Knowledge Accelerator (COKA): <https://gps.health/covid-19-knowledge-accelerator-coka/> ; <https://confluence.hl7.org/pages/viewpage.action?pageId=97468919>
- PheKB: <https://phekb.org/>
- NLM Value Set Authority Center: <https://vsac.nlm.nih.gov/>

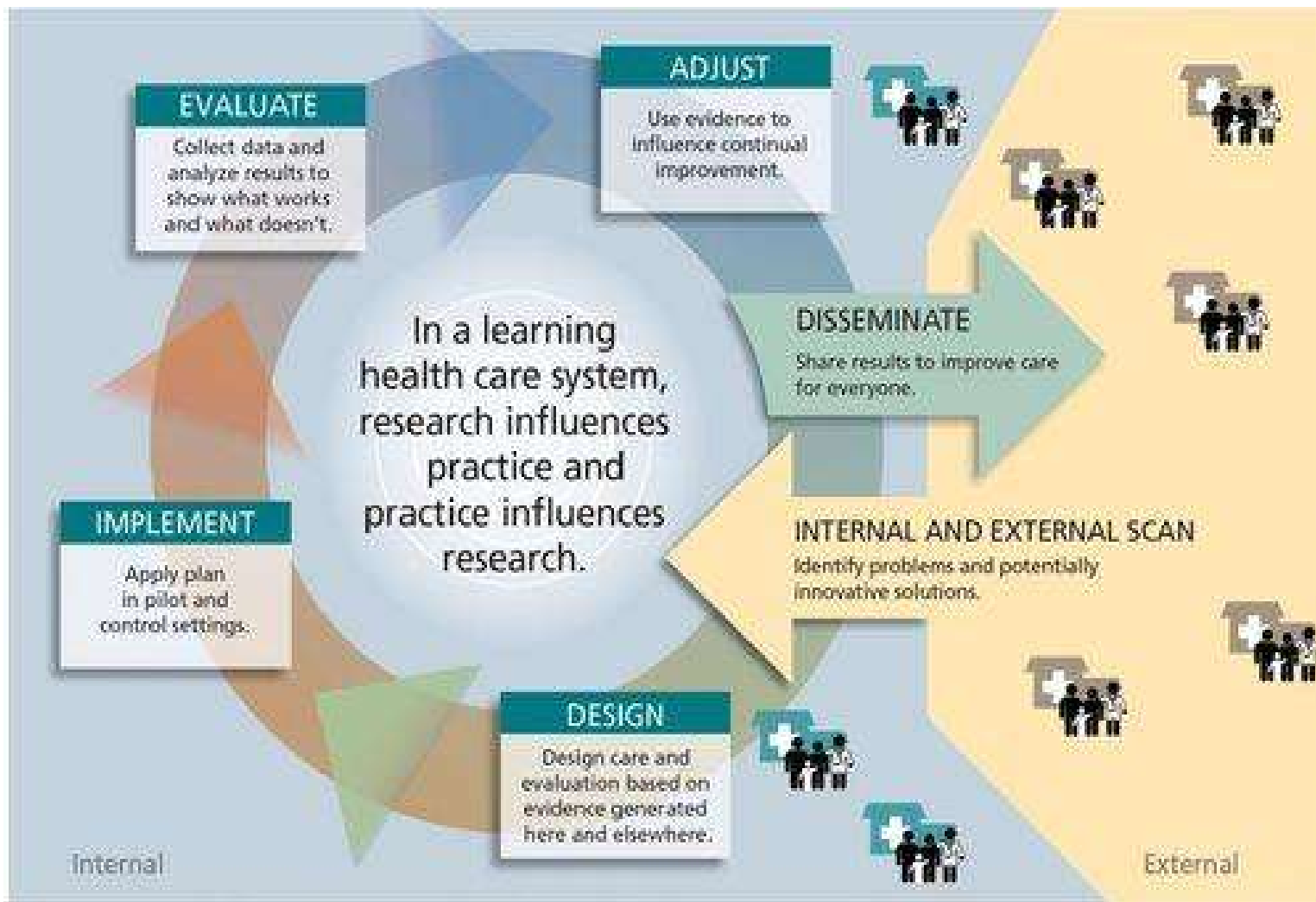
# Learning Objectives

- Describe the relevance of CBK to clinical care delivery, learning health systems, and health improvement
- List types of metadata categories that are important for managing CBK
- List 3 challenges for “mobilizing” CBK for action (in health systems)
- Describe role of research networks in developing and implementing CBK
- Describe how common data models (CDMs) and computable phenotypes support the development and application of CBK
- Identify features for libraries of CBK artifacts (e.g. computable phenotypes)
- Describe challenges for managing CBK at scale and highlight areas needing future development and research

# Outline

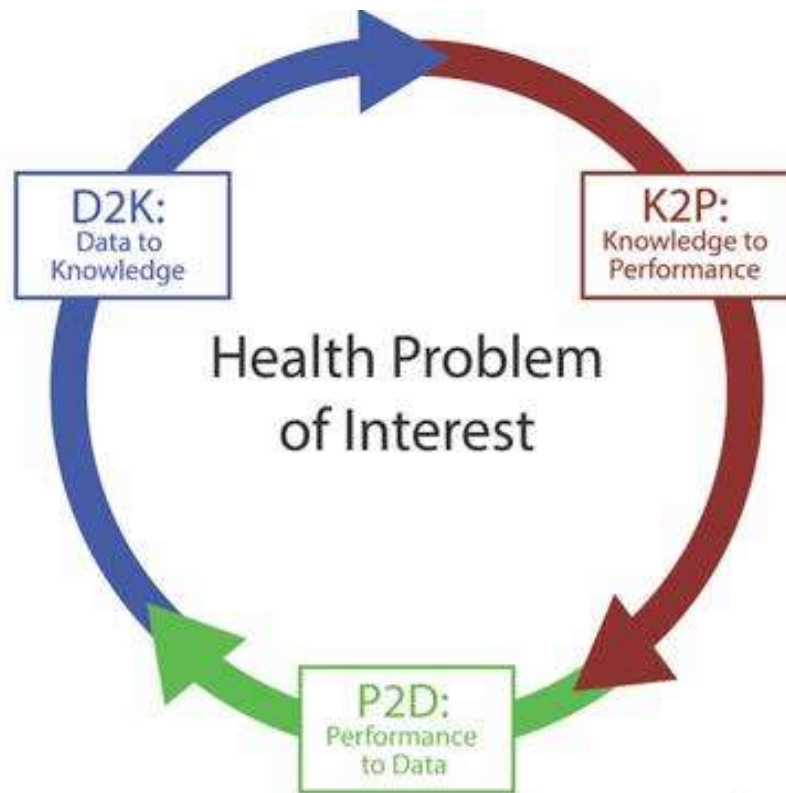
- Review – CBK, LHS, FAIR
- Metadata for CBK
- Mobilizing CBK for Action
  - Research Networks
  - Common Data Models
  - Computable Phenotypes
- Example: Desiderata for computable phenotype libraries
- Outstanding Challenges and Future Directions



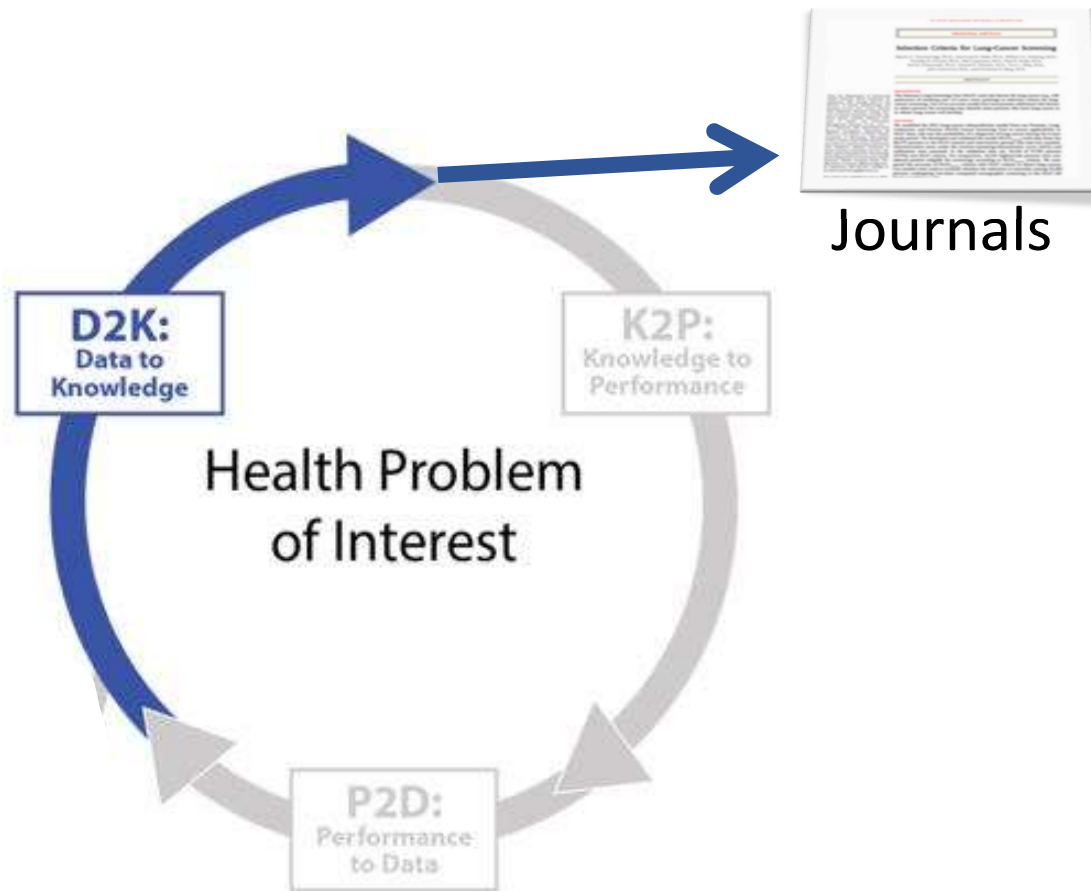


Thomas M. Maddox. Circulation. The Learning Healthcare System and Cardiovascular Care: A Scientific Statement From the American Heart Association, Volume: 135, Issue: 14, Pages: e826-e857, DOI: (10.1161/CIR.0000000000000480)

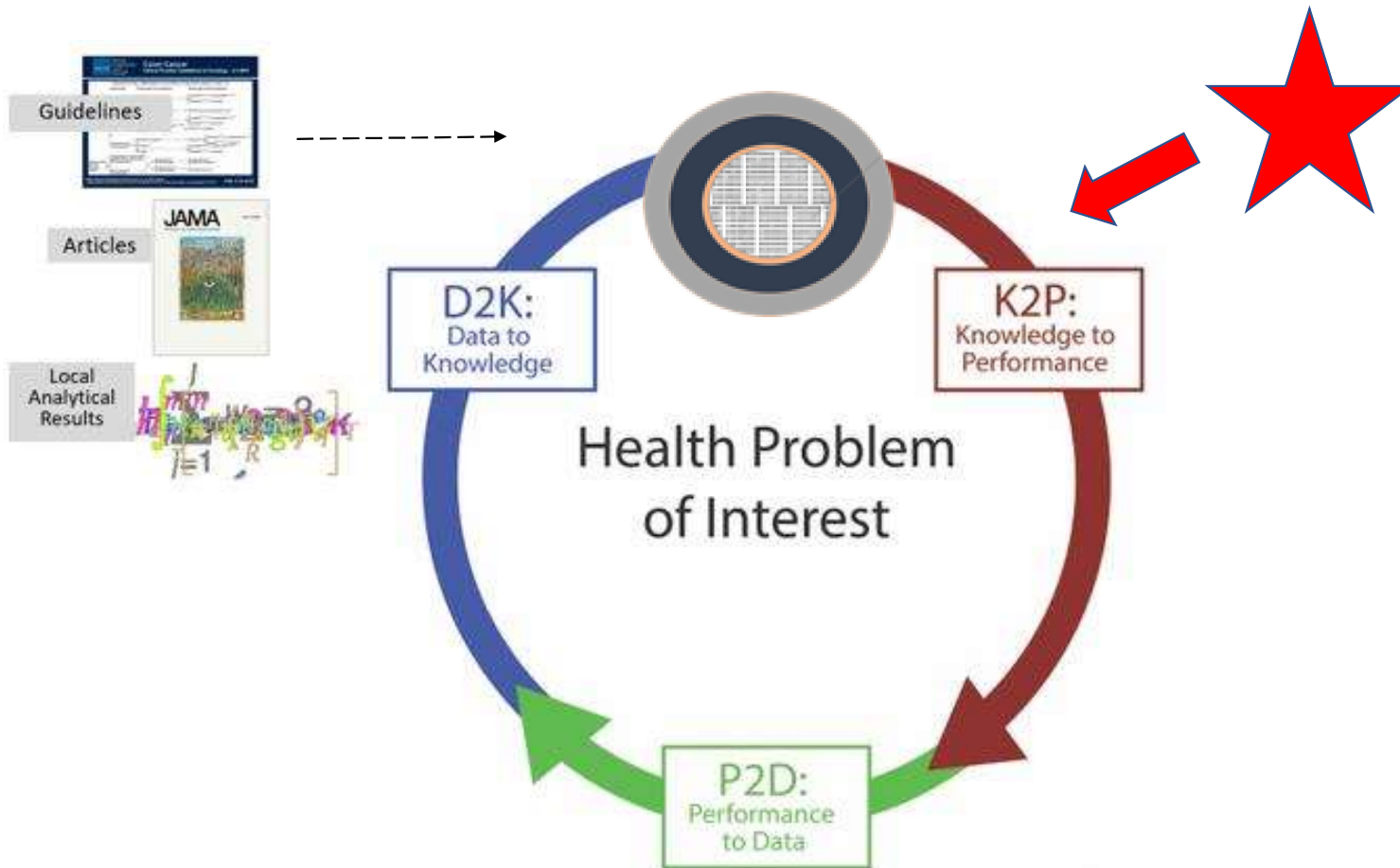
# Better Health Requires This



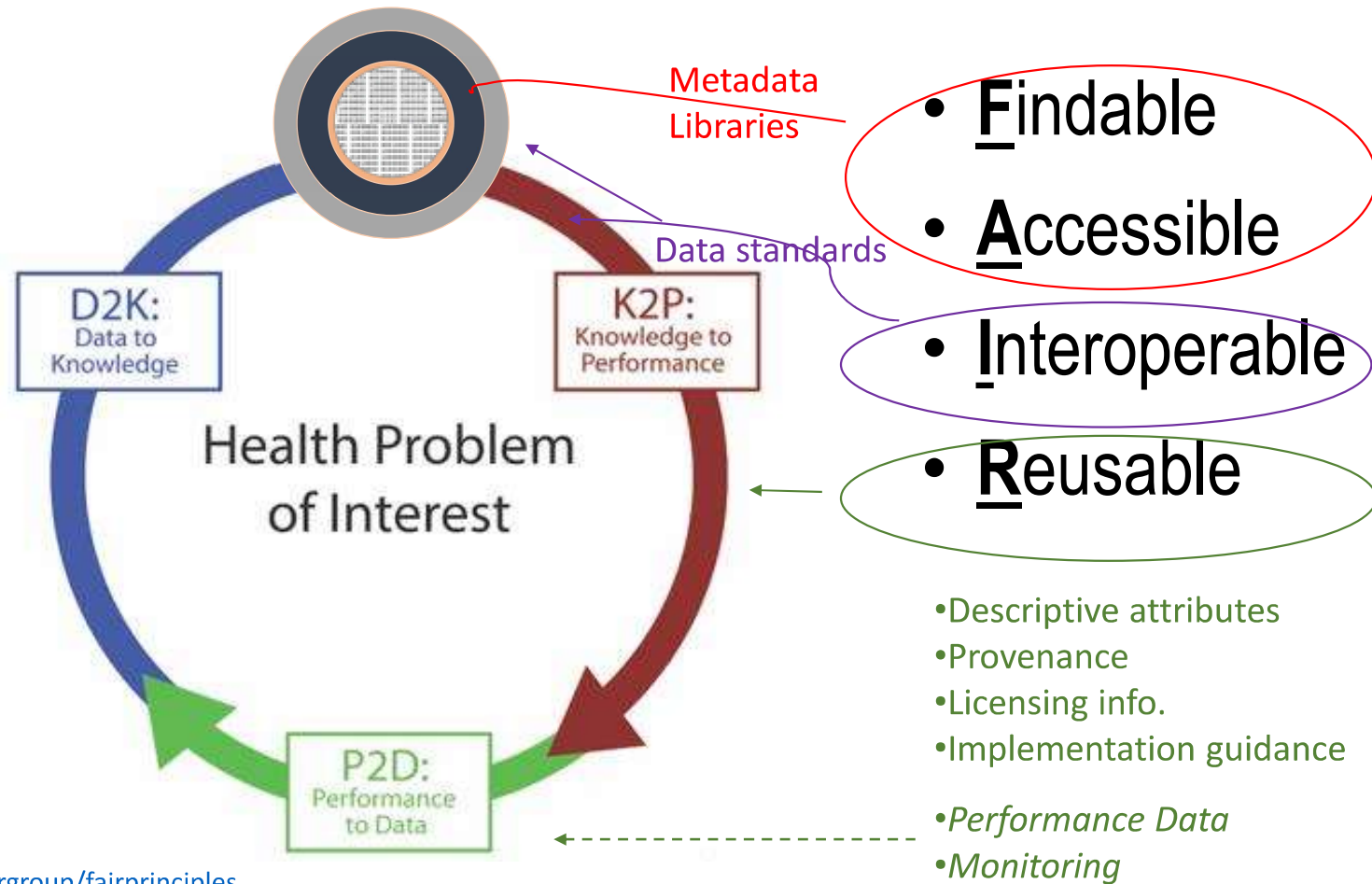
# Not Just This



# Knowledge to Practice



# Knowledge should be FAIR\*



\*FAIR: <https://www.force11.org/group/fairgroup/fairprinciples>

# ACTIVITY

- Finding, understanding, and using CBK...
- Instructions
  - Divide into 4 groups
  - Each group examine one CBK artifact, answer questions, and report back
  - Artifacts can be found here:  
<https://drive.google.com/drive/folders/17idZFaz785807xQhHu9dfL9GR1WhFegE?usp=sharing>
- Credit and appreciation to Dr. Allen Flynn, PhD, PharmD, UM Dept of LHS  
<https://medicine.umich.edu/dept/lhs/allen-flynn-phd-pharmd>

## BACKGROUND

An increasing quantity of biomedical knowledge is being expressed in computer-readable and computer-executable formats. An early example is [MYCIN](#), which used about 600 computer-executable rules to guide the diagnosis and treatment of blood infections. In addition to more advanced [rule-based systems](#), there are recent examples of computer-executable machine-learning models being developed and tested for accuracy in detecting and diagnosing disease in [images](#) or identifying [treatment problems](#).

## OVERALL TOPIC

As more biomedical knowledge used in laboratories, clinics, and homes comes in computer-readable and computer-executable formats, how are knowledge infrastructures changing? In other words, how are libraries, publishers, authors, and knowledge users adapting their tools and processes to handle ***computable biomedical knowledge***?

## GROUP TASK

Carefully examine the computable biomedical knowledge artifact at the link given and then answer the following questions.

# QUESTIONS for each knowledge artifact:

1. Which organization(s) is/are providing this computable biomedical knowledge (CBK) artifact online?

2. What is the purpose of the CBK artifact (ML model)? What is it for? What does it do?

3. What formats or programming languages are used to encode the CBK artifact? Can you tell?

4. Can you find instructions for deploying and using the CBK artifact? How does a person create it or run it or execute it? Can you tell?

5. What are your overall impressions about the knowledge infrastructure involved as users of this CBK artifact web page?

Grp 1: [Statin Use for the Primary Prevention of CVD in Adults: Clinician-Facing CDS Intervention](#)

Grp 2: [Tammemagi, 6 year Lung Cancer Risk Prediction Model for Screening](#)

Grp 3: [Deep EHR: Chronic Disease Prediction Using Medical Notes](#)

Grp 4: [Supervised Classification on liver-disorders – Run 8891972](#)





Featured Apps

All Apps

APPLICATION TYPE

CATEGORIES

Care Coordination

Clinical Research

Data Visualization

Disease Management

Genomics

Medication

Patient Engagement

Featured Apps 58

Sort: Name (A-Z) ▾



**Arrest Assist - Reversible Causes of PEA Arrest Tool**

View

MedStar Institute for Innovation (MI2)

A tool that searches a patient's medical history for reversible causes of PEA arrest. Great for hospital based code teams.

**Specialties:** Anesthesiology, Cardiology, Pulmonary **Designed for:** Clinicians

<https://apps.smarthealthit.org/apps/featured>

# METADATA

***What types of metadata are needed to describe CBK artifacts sufficiently to make them findable, accessible, interoperable, and re-usable (F.A.I.R.)?***

# Learning Health Systems

Open Access

TECHNICAL REPORT |  Open Access |  

## Categorizing metadata to help mobilize computable biomedical knowledge

Brian S. Alper  Allen Flynn, Bruce E. Bray, Marisa L. Conte, Christina Eldredge, Sigfried Gold, Robert A. Greenes, Peter Haug, Kim Jacoby, Gunes Koru, James McClay, Marc L. Sainvil ... [See all authors](#) 

First published: 09 May 2021 | <https://doi-org.proxy.lib.umich.edu/10.1002/lrh2.10271>

Brian S. Alper and Allen Flynn contributed equally to this article.

[Read the full text](#) >



PDF



TOOLS



SHARE

### Abstract

#### Introduction

Computable biomedical knowledge artifacts (CBKs) are digital objects conveying biomedical knowledge in machine-interpretable structures. As more CBKs are produced and their complexity increases, the value obtained from sharing CBKs grows. Mobilizing CBKs and sharing them widely can only be achieved if the CBKs are findable, accessible, interoperable, reusable, and trustworthy (FAIR+T). To help mobilize CBKs, we describe our efforts to outline metadata categories to make CBKs FAIR+T.

Alper, BS, Flynn, A, Bray, BE, et al. Categorizing metadata to help mobilize computable biomedical knowledge. *Learn Health Sys*. 2021;e10271. <https://doi-org.proxy.lib.umich.edu/10.1002/lrh2.10271>

**TABLE 1** List of metadata categories related to making CBKs and FAIR+T

Metadata category	Metadata elements in this category	Example predicates	Main principle supported	From
1. Type	Elements that classify CBKs by describing the <b>nature</b> of CBKs in some general way	[CBK] <i>is_a</i> {type}	FINDABLE	49,50
2. Domain	Elements relating CBKs to the <b>biomedical domains or topics</b> to which they belong	[CBK] <i>is_about</i> {domain}	FINDABLE	51,52
3. Purpose	Elements describing the <b>purposes</b> or circumscribing and limiting the <b>intended uses</b> of CBKs	[CBK] <i>has_purpose_of</i> ____ [CBK] <i>is_intended_to</i> ____ [CBK] <i>is_not_intended_to</i> ____	FINDABLE	53
4. Identification	Elements indicating <b>persistent identifiers</b> or <b>persistent unique identifiers</b> and <b>versions</b> assigned to CBKs	[CBK] <i>has_identifier</i> ____ [CBK] <i>has_name</i> ____ [CBK] <i>has_version</i> ____	FINDABLE	49,50
5. Location	Elements indicating the <b>physical or virtual locations</b> where CBKs can be accessed	[CBK] <i>has_location</i> {ADDRESS} [CBK] <i>is_located_at</i> {URL}	ACCESSIBLE	49,50
6. CBK-to-CBK relationships	Elements describing a <b>relationship between one CBK and some other CBK</b>	[CBK] <i>is_modification_of</i> [CBK] [CBK] <i>is_predecessor_of</i> [CBK] [CBK] <i>is_successor_of</i> [CBK] [CBK] <i>is_used_with</i> [CBK]	INTEROPERABLE	49,50
7. Technical	Elements to describe a wide array of <b>technical characteristics</b> of CBKs that need to be known to deploy, integrate, operate, and use them	[CBK] <i>has_file_type</i> ____ [CBK] <i>has_file_size</i> ____ [CBK] <i>has_dependency</i> ____ [CBK] <i>can be executed using</i> ____ [CBK] <i>has input</i> ____ [CBK] <i>has output</i> ____	INTEROPERABLE	54,55
8. Authorization and rights management	Elements describing <b>rights and responsibilities</b> pertaining to CBKs	[CBK] <i>is_available_to</i> [person] [CBK] <i>has_license</i> [license] [CBK] <i>copyright_held_by</i> [agent] [CBK] <i>has_disclaimer</i> [disclaimer]	REUSABLE	56

10. Integrity	Elements conveying <b>outputs from cryptographic functions</b> that allow CBK users to confirm CBK has not been tampered with	[CBK] <i>has_hash</i> [hash function output] [CBK] <i>uses_hash_function_type</i> [type]	REUSABLE	58
11. Provenance	Elements indicating <b>changes in ownership, custody, and status</b> during CBK lifecycles	[CBK] <i>is_owned_by</i> [agent] [CBK] <i>ownership_changed_on</i> [date] [CBK] <i>has status</i> [status] [CBK] <i>status_changed_on</i> [date] [CBK] <i>is_authored_by</i> [author] [CBK] <i>is_reviewed_by</i> [reviewer] [CBK] <i>is_endorsed_by</i> [endorser]	TRUSTABLE	59
<i>Two evidence categories</i>				
12. Evidential basis	Elements describing the <b>data upon which the claims in CBKs are based, the methods of obtaining and analyzing those data</b> , and the <b>strength</b> of the evidential basis of CBKs.	[CBK] <i>is_based_on_data_about</i> ____ [CBK] <i>is_based_on_data_collected_at</i> [place] [CBK] <i>is_based_on_data_collected_by</i> [agent] [CBK] <i>is_based_on_data_collected_on</i> [date] [CBK] <i>is_based_on_data_collected_for</i> ____ [CBK] <i>is_based_on_data_analysis_method_of</i> ____ [CBK] <i>is_based_on_data_analysis_results_of</i> ____ [CBK] <i>has_certainty_of_evidence</i> ____	TRUSTABLE	2,60-62
13. Evidence from use	Elements describing <b>data arising from CBK use, the methods of obtaining and analyzing those data</b> , and the <b>strength</b> of evidence about CBK use	[CBK] <i>use_is_evaluated_in</i> ____ [CBK] <i>use_is_associated_with</i> ____ [CBK] <i>use causes</i> ____ [CBK] <i>use_evidence_has_certainty_of</i> ____	TRUSTABLE	61-63



**TABLE 3** Research agenda for further CBK metadata exploration and analysis

Research agenda item	Brief description of research agenda item	Related metadata category
CBK typologies	A variety of different approaches have been taken to define the types and subtypes of CBKs. More work is needed to synthesize these efforts into coherent CBK typologies to support standards for CBK types.	Type
Schema for purpose metadata	There is an apparent need to formalize CBK purpose metadata. As complex artificial artifacts, all CBKs emerge from some human design process. It may be possible to create schema to convey the motivations and intents of CBK designers and of CBK users and others coherently and usefully.	Purpose
Schema for CBK-to-CBK relationships metadata	The many ways in which CBKs relate to one another are not clear. Work is needed to examine potential relationships between types of CBKs and actual relationships between existing CBKs.	CBK-to-CBK relationships
CBK lifecycles	The lifecycles of CBKs need to be better understood. Since CBK lifecycles may vary by CBK type, interactions between Provenance Metadata and Type Metadata need to be explored.	Provenance, Type, Preservation
CBK use outcomes	It is not clear which outcomes from using CBKs are of most interest to users. Studies of CBK user needs for evidence arising from use of CBKs are needed to better understand outcomes of interest.	Evidence from Use
Relationships between CBK metadata and the FAIR and trustability principles	Studies to test the hypotheses surfaced here that metadata from 13 categories can uphold the findability, accessibility, interoperability, reusability, and trustability of CBKs are needed.	All

Common Metadata Framework

## Navigation

[Project Title](#)  
[Project Description](#)  
[Project Actions](#)  
[Project Details](#)  
[Associated Resources](#)

Communicate

TY Share **Comment** Ask

Classify Rate Follow

Edit Project

Clone Project

Verify Project

**View JSON**

Add to Project

Exchange Data

Text View JSON View Usage View

**Project Title**

Common Metadata Framework

**Project Description**

Individuals from the COVID-19 Knowledge Accelerator (COKA) and Mobilizing Computable Biomedical Knowledge (MCBK) initiatives are contributing to specifications for a common metadata framework to facilitate making data Findable, Accessible, Interoperable and Reusable (FAIR) across systems that may use different standards for metadata specification.

**Project Actions**

No actions.

**Project Details**

A working group within the Mobilizing Computable Biomedical Knowledge (MCBK) Standards Working Group completed a year of effort and published "Categorizing metadata to help mobilize computable biomedical knowledge" which identified 13 metadata categories to communicate the Findability, Accessibility, Interoperability, Reusability, and Trustability (FAIR+T) of knowledge artifacts.

<https://onlinelibrary.wiley.com/doi/10.1002/lrh2.10271>

The COVID-19 Knowledge Accelerator (COKA) Common Metadata Framework Working Group started February 2, 2021 and included some of the MCBK authors and some of the COKA participants.

We spent 4 months mapping FHIR Resource StructureDefinitions to FAIR+T principles and the 13 metadata categories to inform initial thinking about a common metadata framework.

We changed our approach in June and spent 5 months specifying elements for each of the 13 metadata categories (for a total of 134 elements). The details of this "first draft" of "Specifying metadata to help mobilize computable biomedical knowledge" can be found at "Specifying metadata to MCBK Spreadsheet":

<https://docs.google.com/spreadsheets/d/1yUiSVsOcTJ3J4J9DR-4LZNcD36ZGfEM/#gid=1635001094>

January 5, 2021

Dataset

Open Access

# Crosswalk of most used metadata schemes and guidelines for metadata interoperability

 Ojsteršek

## Related person(s)

 Corcho, Oscar;  Eriksson, Magnus;  Kurowski, Krzysztof;  van de Sanden, Mark;  Coppens, Frederik

This resource provides crosswalks among the most commonly used metadata schemes and guidelines to describe digital objects in Open Science, including:

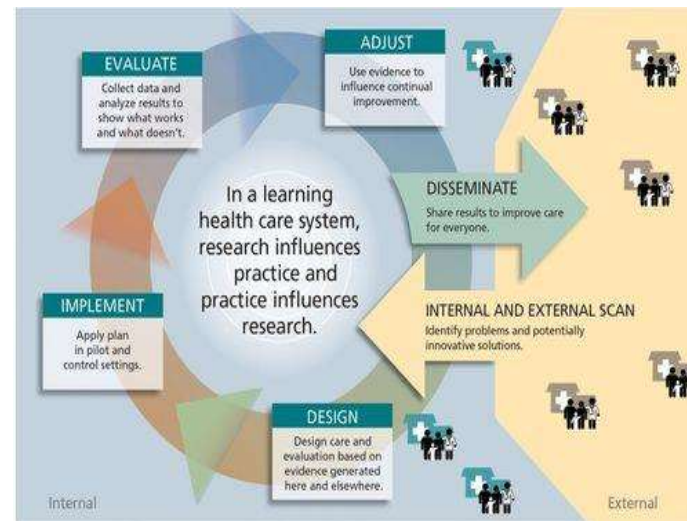
- RDA metadata IG recommendation of the metadata element set,
- EOSC Pilot - EDM metadata set,
- Dublin CORE Metadata Terms,
- Datacite 4.3 metadata schema,
- DCAT 2.0 metadata schema and DCAT 2.0 application profile,
- EUDAT B2Find metadata recommendation,
- OpenAIRE Guidelines for Data Archives,
- OpenAire Guidelines for literature repositories 4.0,
- OpenAIRE Guidelines for Other Research Products,
- OpenAIRE Guidelines for Software Repository Managers,
- OpenAIRE Guidelines for CRIS Managers,
- Crossref 4.4.2 metadata XML schema,
- Harvard Dataverse metadata schema,
- DDI Codebook 2.5 metadata XML schema,
- Europeana EDM metadata schema,
- Schema.org,
- Bioschemas,
- The PROV Ontology.

<https://zenodo.org/record/4420116#.YbjDJGDMJPY>



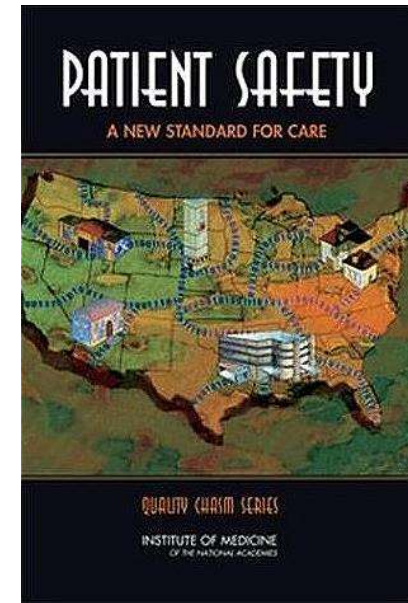
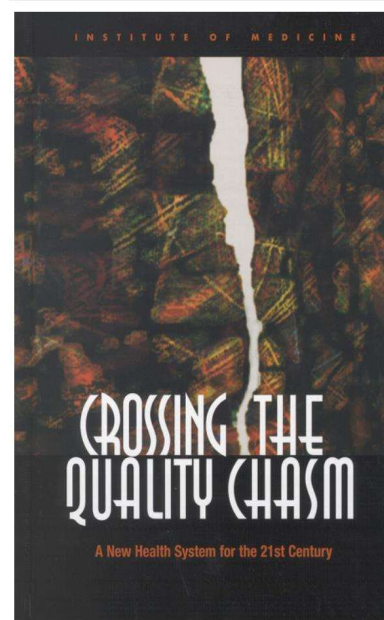
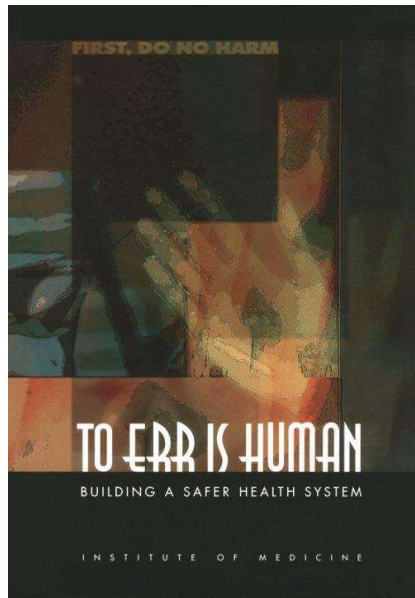
**Question:**

**What are challenges for mobilizing CBK (for Action)?**



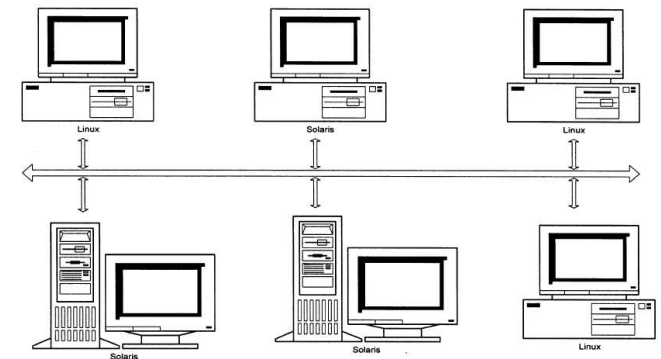
# Next topics

- Mobilizing CBK for Action
  - EHR Data
  - Research Networks
  - Common Data Models
  - Computable Phenotypes
- Example: Desiderata for computable phenotype libraries
- Outstanding Challenges and Future Directions



***"Interoperability must be addressed now, or else widespread adoption of stand-alone EHRs will be a fait accompli."***

David Brailer, MD, PhD, National Coordinator for Health Information Technology; Remarks at HIMSS 2005 Annual Conference, Feb 17, 2005



## Types of EHR data

- Diagnoses
- Problems
- Procedures
- Tests
- Lab results/values
- Family History
- Allergies
- Immunization
- Utilization
- Reports
- Notes



# Types of EHR data

- Diagnoses
- Problems
- Procedures
- Tests
- Lab results/values
- Family History
- Allergies
- Immunization
- Utilization
- Reports
- Notes

← Tweet

 **Traci Kurtzer MD**  
@Kurtzer\_MD

Replying to @DGlaucmflecken

Why is there one ICD10 code for pelvic pain R10.2 to cover so many different diagnoses & structures? Like it covers 28 different conditions and we wonder why women's health research is limited? As a comparison there are a myriad of codes for various falls i.e. off cliff or tree!

Approximate synonyms

- Acute pain in female pelvis
- Acute pelvic pain, female
- Burning sensation of vagina
- Burning sensation of vulva
- Chronic female pelvic pain syndrome
- Chronic pain in male pelvis
- Chronic pain in vagina
- Chronic pelvic pain of female
- Chronic pelvic pain syndrome
- Chronic pelvic pain, female
- Chronic vaginal pain
- Pain in female pelvis
- Pain in female perineum
- Pain in male pelvis
- Pain in male perineum
- Pain in pelvis
- Pain in round ligament in pregnancy
- Pain in vagina
- Pelvic pain
- Pelvic pain, female
- Pelvic pain, male
- Perineal pain, female
- Perineal pain, male
- Round ligament pain in pregnancy
- Vaginal burning
- W01 ☑ Fall on same level from slipping, tripping and stumbling
- W03 ☑ Other fall on same level due to collision with another person
- W04 ☑ Fall while being carried or supported by other persons
- W05 ☑ Fall from non-moving wheelchair, nonmotorized scooter and motorized mobility scooter
- W06 ☑ Fall from bed
- W07 ☑ Fall from chair
- W08 ☑ Fall from other furniture
- W09 ☑ Fall on and from playground equipment
- W10 ☑ Fall on and from stairs and steps
- W11 ☑ Fall on and from ladder
- W12 ☑ Fall on and from scaffolding
- W13 ☑ Fall from, out of or through building or structure
- W14 ☑ Fall from tree
- W15 ☑ Fall from cliff
- W16 ☑ Fall, jump or diving into water
- W17 ☑ Other fall from one level to another
- W18 ☑ Other slipping, tripping and stumbling

[https://twitter.com/Kurtzer\\_MD/status/1434236920071139338](https://twitter.com/Kurtzer_MD/status/1434236920071139338)

*What types of data are needed for patient-centered care – and are missing here?*

## Types of EHR data

- Diagnoses
- Problems
- Procedures
- Tests
- Lab results/values
- Family History
- Allergies
- Immunization
- Utilization
- Reports
- Notes
- Nursing; PT / OT / Diet / other
- Functioning and QOL
- Preferences
- SDOH
- Demographics



# Newsroom

<http://www.healthit.gov/newsroom/about-onc>

- Health IT in the News
- News Releases
- Events
- Fact Sheets
- Infographics
- Recent Updates
- About ONC

- ▶ [ONC Budget Documents and Performance Information](#)
- ▶ [Grants Management Advisories](#)
- ▶ [Grants Policy Statement](#)
- ▶ [ONC Program Information Notices \(PINS\)](#)
- ▶ [ONC Program Assistance Letters \(PALS\)](#)

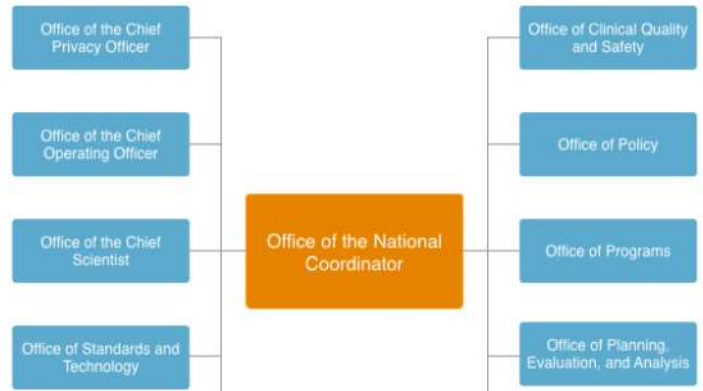
## About ONC

The Office of the National Coordinator for Health Information Technology (ONC) is at the forefront of the administration's health IT efforts and is a resource to the entire health system to support the adoption of health information technology and the promotion of nationwide health information exchange to improve health care. ONC is organizationally located within the Office of the Secretary for the U.S. Department of Health and Human Services (HHS).

ONC is the principal federal entity charged with coordination of nationwide efforts to implement and use the most advanced health information technology and the electronic exchange of health information. The position of National Coordinator was created in 2004, through an Executive Order, and legislatively mandated in the Health Information Technology for Economic and Clinical Health Act (HITECH Act) of 2009.

June 2014: [Statements of Organization, Functions, and Delegations of Authority: Office of the National Coordinator for Health Information Technology](#)

### ONC Organization



### Media Questions

Contact Peter Ashkenaz if you have media questions. Your queries will be addressed within one business day.

- [Go to ONC Speaker Request Form](#)
- [Get On-the-Ground Support](#)

**Email:**  
[Peter.Ashkenaz@hhs.gov](mailto:Peter.Ashkenaz@hhs.gov)  
 Telephone: (202) 260-6342  
 Fax: 202-690-6079

### Media Resources

- [About ONC](#)
- [Leadership Bios](#)
- [Federal Advisory Committee Act \(FACA\)](#)
- [3 Important Things to Know about Health IT](#)
  - [Size 2.75" x 7" \[PDF - 85 KB\]](#)
  - [Size 4.25" x 5.5" \[PDF - 93 KB\]](#)



## Interoperability Standards Advisory (ISA)

The Interoperability Standards Advisory (ISA) process represents the model by which the Office of the National Coordinator for Health Information Technology (ONC) will coordinate the identification, assessment, and determination of "recognized" interoperability standards and implementation specifications for industry use to fulfill specific clinical health IT interoperability needs.



### News & Updates

The comment period for ISA and submission period for Draft USCDI Version 3 will be open until September 30th at 11:59pm ET.

The public comment period for the Standards Version Advancement Process (SVAP) has been extended to May 2, 2022 to align with important standards development activities. Remember to log in or register to post a comment or to submit data elements and classes.

Please refer to the [Health IT Buzz Blog](#) for additional details.

#### About ISA

The ISA is frequently updated to include improvements made

#### ISA Structure

The ISA is organized and structured into four sections - read


#### Table of Contents

The Table of Contents of ISA's sections.

<https://www.healthit.gov/isa/>



<http://www.nlm.nih.gov/research/umls/>



United States  
**National Library of Medicine**  
National Institutes of Health

Search NLM Web Site

[Go](#)

[NLM Home](#) | [Contact NLM](#) | [Site Map](#) | [FAQs](#)

---

## Unified Medical Language System

[UMLS Home](#)

---

[Home](#) > [Biomedical Research & Informatics](#) > [UMLS](#)

---

1/29/07: **UMLS 2007AA Release** now available for download from the UMLSKS. ••• 9/30/06 **Draft LOINC to CPT Mappings** now available for download from the UMLSKS. ••• New to the UMLS? [Register now.](#)

[About the UMLS Resources](#)  
Metathesaurus; Semantic Network; SPECIALIST Lexicon and lexical programs; MetamorphoSys

[Accessing UMLS Knowledge Sources](#)  
Metathesaurus license; Semantic Network; SPECIALIST Lexicon; DVD

[Knowledge Source Server](#)  
Download files; searching; additional tools and resources

[Documentation](#)

### Metathesaurus Source Vocabularies

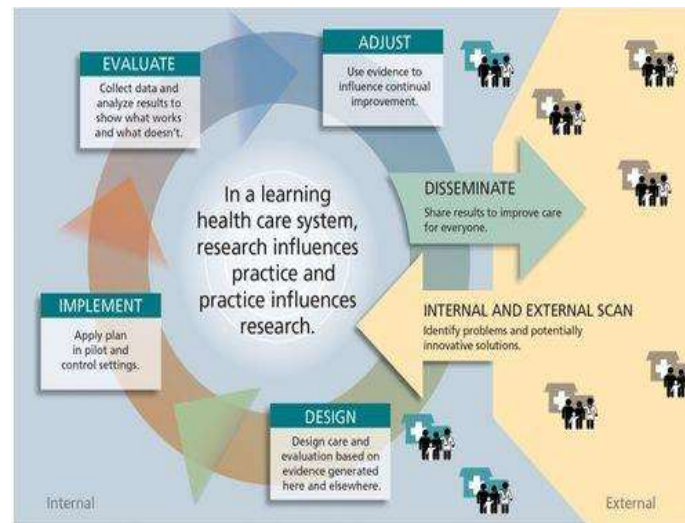
- [SNOMED CT](#)
- [LOINC](#)
- [RxNorm](#)
- [MeSH](#)
- [List of Sources](#)
- [Source FAQs](#)
- [Mappings](#)

### More Resources

- [Metathesaurus License](#)
- [Tools](#)
- [Learning Resources](#)
- [MetaMap Transfer \(MMTx\)](#)

**Question:**

**What is the role of research / healthcare networks in building and implementing CBK (at scale)?**

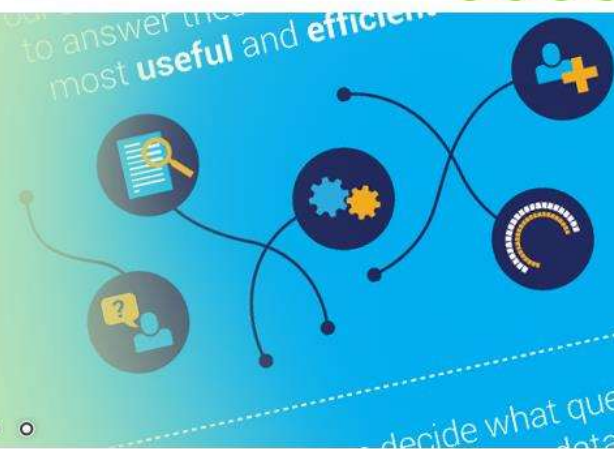


Networked Research  
and  
Common Data Models



## PCORnet – A Platform for More Efficient Health Research

Learn about this effort to harness the power of partnerships and data to improve patient outcomes



## PCORnet, the National Patient-Centered Clinical Research Network

PCORnet, the National Patient-Centered Clinical Research Network, is an innovative initiative of the [Patient-Centered Outcomes Research Institute \(PCORI\)](#). PCORnet will transform clinical research by engaging patients, care providers and health systems in collaborative partnerships that leverage health data to advance medical knowledge and improve health care. PCORnet will bring together health research and healthcare delivery, which have been largely separate endeavors. By doing so, this national health data network will allow us to explore the questions about conditions, care and outcomes that matter most to patients and their families.

PCORnet represents a unique opportunity to make a real difference in the lives of patients and their families. Until now, we have been unable to answer many most important questions affecting health and healthcare. But by combining the knowledge and insights of patients, caregivers, and researchers in a revolutionary network with carefully controlled access to rich sources of health data, we will be able to respond to patients' priorities and speed the creation of new knowledge to guide treatment on a national scale.

[PCORnet Blog](#)

[PCORnet Research](#)

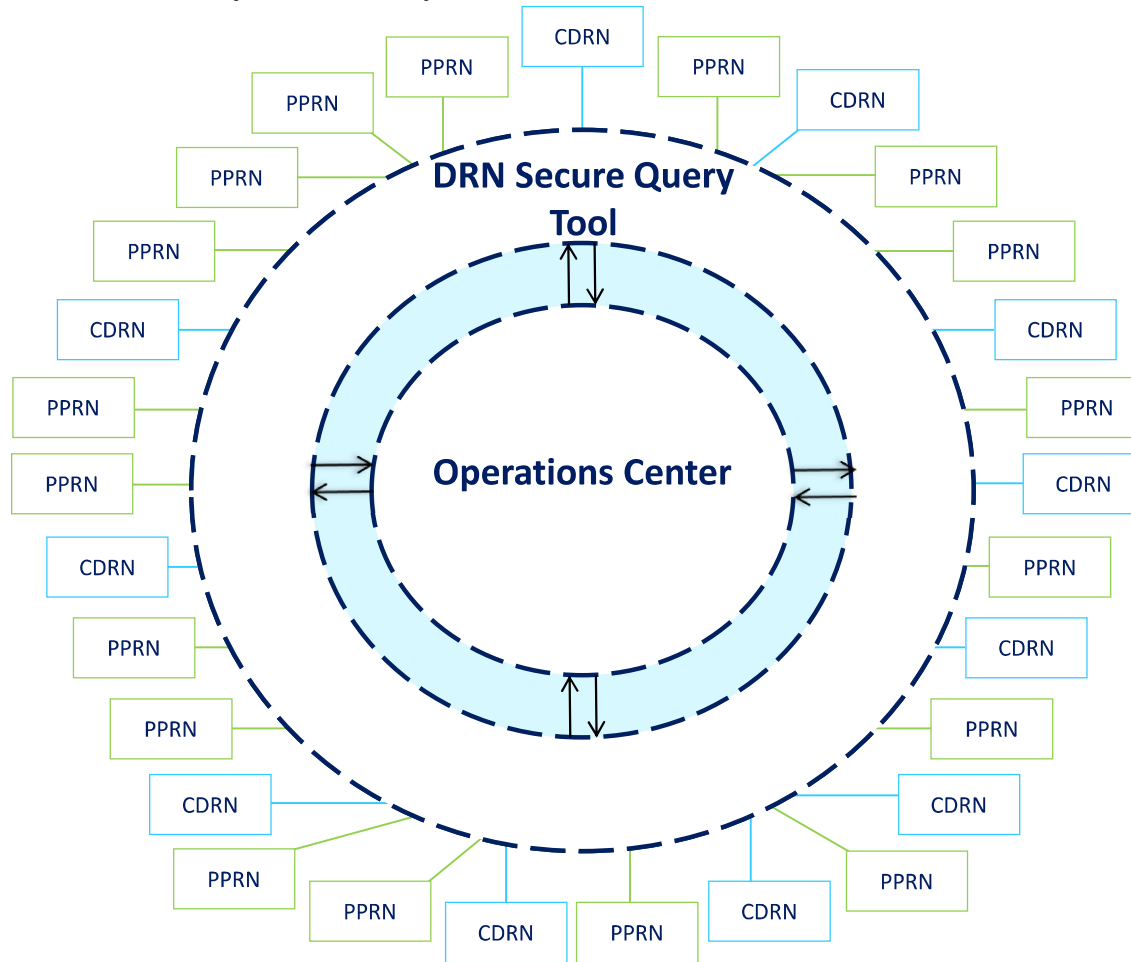
[PCORnet Partner Networks](#)

<http://pcornet.org/>

# 11 Clinical Data Research Networks and 18 Patient Powered Research Networks

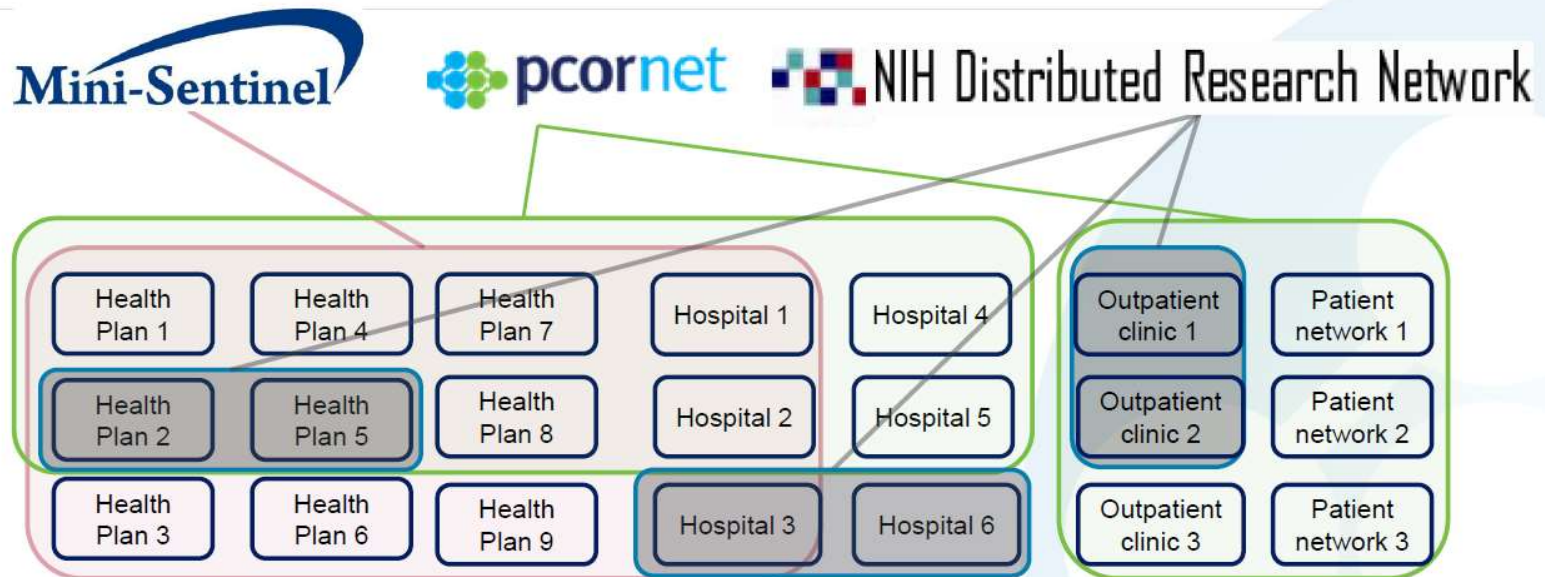


# PCORnet Distributed Research Network (DRN)





# Multiple Networks Sharing Infrastructure





# OHDSI

OBSERVATIONAL HEALTH DATA SCIENCES AND INFORMATICS

[Who We Are](#) [Who We Serve](#) [Data Standardization](#) [Software Tools](#) [Resources](#) [Join the Journey](#) [Events](#)

## Welcome to OHDSI!

The Observational Health Data Sciences and Informatics (or OHDSI, pronounced "Odyssey") program is a multi-stakeholder, interdisciplinary collaborative to bring out the value of health data through large-scale analytics. All our solutions are open-source.

OHDSI has established an international network of researchers and observational health databases with a central coordinating center housed at Columbia University.

Read more [about us](#), [about our goals](#), and how you can help support the OHDSI community.

[Join the Journey](#)

**OHDSI**  
OBSERVATIONAL HEALTH DATA SCIENCES AND INFORMATICS

**OHDSI SYMPOSIUM**  
October 20th 2015

**ABOUT THE EVENT**  
As the first OHDSI symposium, this event will provide those interested in observational research an opportunity to learn from OHDSI collaborators who will showcase their work and discuss the impact OHDSI will have on medical decision-making.

The day will also include hands-on demonstrations of the OHDSI tools that will revolutionize how medical evidence is generated.

**JOIN THE JOURNEY**  
Stakeholders who would benefit from this meeting include:

- Representatives from CMS, FDA, NIH, AHRQ and PCORI
- Representatives from academia, industry and healthcare start-ups
- Healthcare payers & providers

Presenters for the event will include influential leaders from biomedical informatics, epidemiology, computer sciences and statistics.

Washington Hilton, 1919 Connecticut Ave NW  
Washington, DC 20009

<http://www.ohdsi.org/>

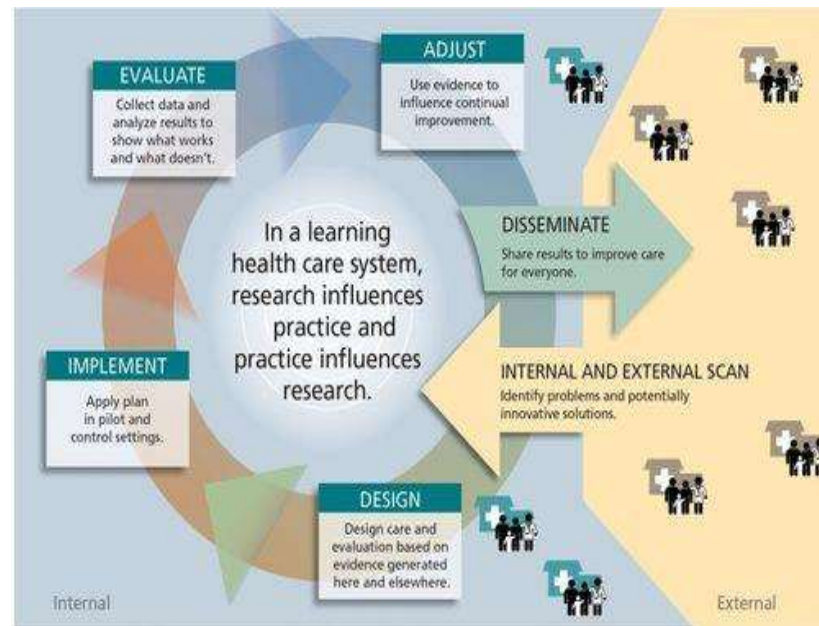
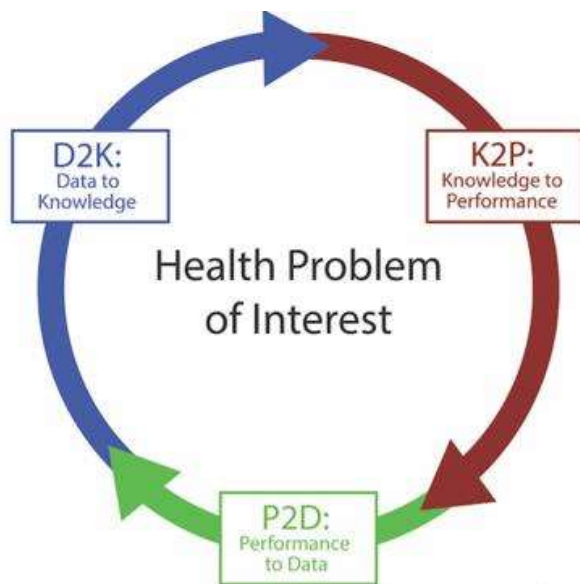


## Other networks

- *NIH Collaboratory Distributed Research Network (DRN)*
- *the High Value Healthcare Collaborative (HVHC)*
- *the Health Care Systems Research Network*
- *the Observational Health Data Sciences and Informatics (OHDSI) program*

## Re-cap:

**Describe role of research networks in developing and implementing CBK.**

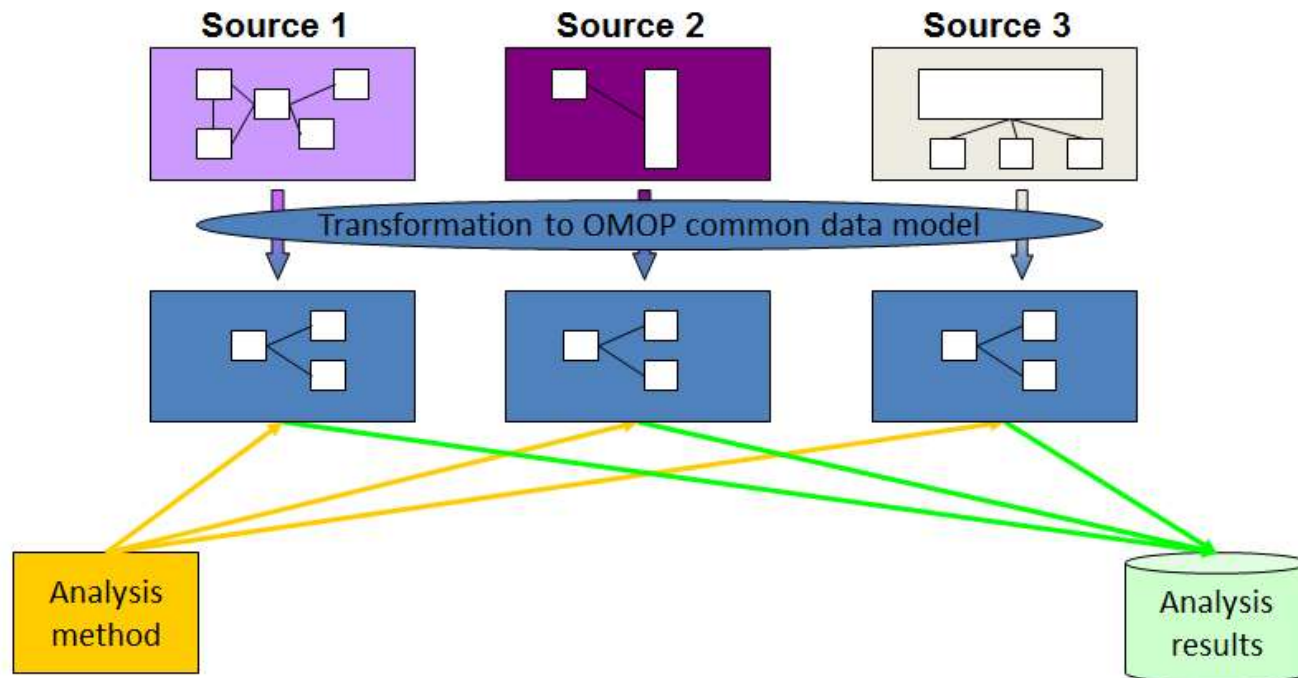


# Common Data Models (CDM)

- Allows for the systematic analysis of disparate observational databases.
- Approach is to transform data from disparate databases into a common format (data model), and then perform systematic analyses using a library of standard analytic routines, based on the common format.
- ***Why do we need a CDM?***
- Observational databases differ in both purpose and design.
- Have different logical organization and physical formats, and the terminologies used vary.

<http://www.ohdsi.org/data-standardization/the-common-data-model/>

# OMOP Common Data Model

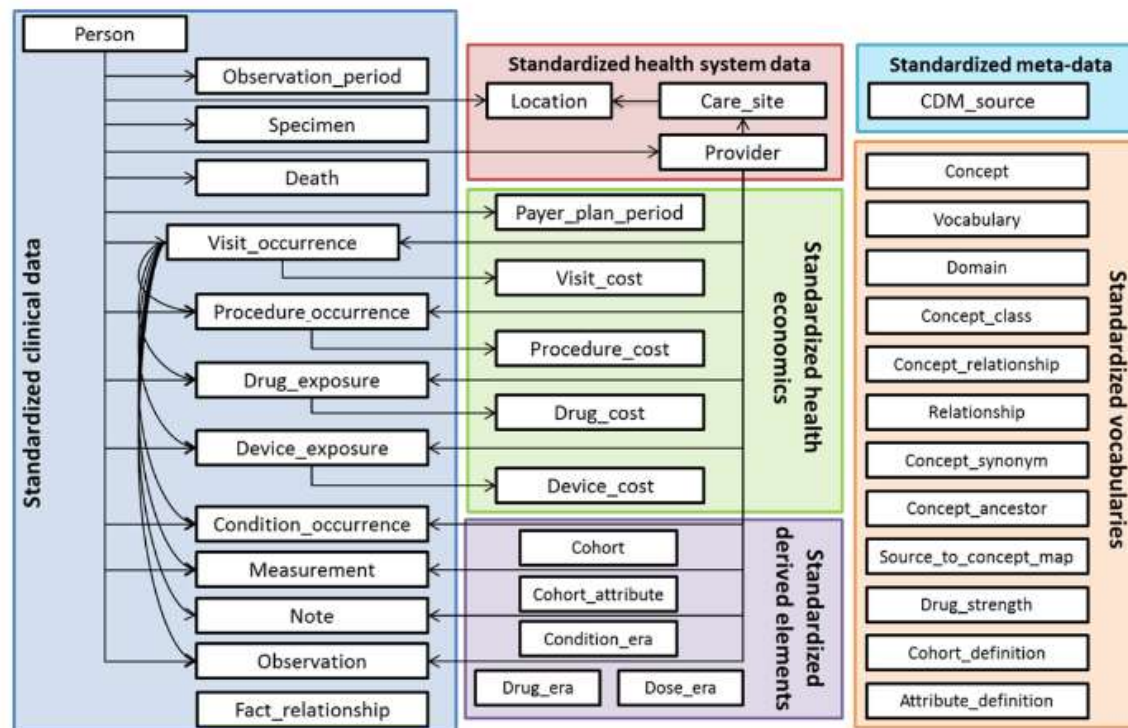


OMOP = The Observational Medical Outcomes Partnership

Source: <http://www.ohdsi.org/data-standardization/the-common-data-model/>

# OMOP Common Data Model Specifications Version 5.0

October 14, 2014



<https://github.com/OHDSI/CommonDataModel/blob/master/OMOP%20CDM%20v5.pdf> ;  
[http://www.ohdsi.org/web/wiki/doku.php?id=documentation:cdm:details\\_of\\_the\\_model](http://www.ohdsi.org/web/wiki/doku.php?id=documentation:cdm:details_of_the_model)

# PCORnet Common Data Model v3.0

New to v3.0

DEMOGRAPHIC
PATID
BIRTH_DATE
BIRTH_TIME
SEX
HISPANIC
RACE
BIOBANK_FLAG

Fundamental basis

ENROLLMENT
PATID
ENR_START_DATE
ENR_END_DATE
CHART
ENR_BASIS

DISPENSING
DISPENSINGID
PATID
PRESCRIBINGID (optional)
DISPENSE_DATE
NDC
DISPENSE_SUP
DISPENSE_AMT

DEATH
PATID
DEATH_DATE
DEATH_DATE_IMPUTE
DEATH_SOURCE
DEATH_MATCH_CONFIDENCE

DEATH_CONDITION
PATID
DEATH_CAUSE
DEATH_CAUSE_CODE
DEATH_CAUSE_TYPE
DEATH_CAUSE_SOURCE
DEATH_CAUSE_CONFIDENCE

Data captured from processes associated with healthcare delivery

VITAL
VITALID
PATID
ENCOUNTERID (optional)
MEASURE_DATE
MEASURE_TIME
VITAL_SOURCE
HT
WT
DIASTOLIC
SYSTOLIC
ORIGINAL_BMI
BP_POSITION
SMOKING
TOBACCO
TOBACCO_TYPE

CONDITION
CONDITIONID
PATID
ENCOUNTERID (optional)
REPORT_DATE
RESOLVE_DATE
ONSET_DATE
CONDITION_STATUS
CONDITION
CONDITION_TYPE
CONDITION_SOURCE

PRO_CM
PRO_CM ID
PATID
ENCOUNTERID (optional)
PRO_ITEM
PRO_LOINC
PRO_DATE
PRO_TIME
PRO_RESPONSE
PRO_METHOD
PRO_MODE
PRO_CAT

Data captured within multiple contexts: healthcare delivery, registry activity, or directly from patients

ENCOUNTER
ENCOUNTERID
PATID
ADMIT_DATE
ADMIT_TIME
DISCHARGE_DATE
DISCHARGE_TIME
PROVIDERID
FACILITY_LOCATION
ENC_TYPE
FACILITYID
DISCHARGE_DISPOSITION
DISCHARGE_STATUS
DRG
DRG_TYPE
ADMITTING_SOURCE

DIAGNOSIS
DIAGNOSISID
PATID
ENCOUNTERID
ENC_TYPE (replicated)
ADMIT_DATE (replicated)
PROVIDERID (replicated)
DX
DX_TYPE
DX_SOURCE
PDX

PROCEDURES
PROCEDURESID
PATID
ENCOUNTERID
ENC_TYPE (replicated)
ADMIT_DATE (replicated)
PROVIDERID (replicated)
PX_DATE
PX
PX_TYPE
PX_SOURCE

Data captured from healthcare delivery, direct encounter basis

LAB_RESULT_CM
LAB_RESULT_CM ID
PATID
ENCOUNTERID (optional)
LAB_NAME
SPECIMEN_SOURCE
LAB_LOINC
PRIORITY
RESULT_LOC
LAB_PX
LAB_PX_TYPE
LAB_ORDER_DATE
SPECIMEN_DATE
SPECIMEN_TIME
RESULT_DATE
RESULT_TIME
RESULT_QUAL
RESULT_NUM
RESULT_MODIFIER
RESULT_UNIT
NORM_RANGE_LOW
NORM_MODIFIER_LOW
NORM_RANGE_HIGH
NORM_MODIFIER_HIGH
ABN_IND

PRESCRIBING
PRESCRIBINGID
PATID
ENCOUNTERID (optional)
RX_PROVIDERID
RX_ORDER_DATE
RX_ORDER_TIME
RX_START_DATE
RX_END_DATE
RX_QUANTITY
RX_REFILLS
RX_DAYS_SUPPLY
RX_FREQUENCY
RX BASIS
RXNORM_CUI

PCORNET_TRIAL
PATID
TRIALID
PARTICIPANTID
TRIAL_SITEID
TRIAL_ENROLL_DATE
TRIAL_END_DATE
TRIAL_WITHDRAW_DATE
TRIAL_INVITE_CODE

Associations with PCORnet clinical trials

HARVEST
NETWORKID
NETWORK_NAME
DATAMARTID
DATAMART_NAME
DATAMART_PLATFORM
CDM_VERSION
DATAMART_CLAIMS
DATAMART_EHR
BIRTH_DATE_MGMT
ENR_START_DATE_MGMT
ENR_END_DATE_MGMT
ADMIT_DATE_MGMT
DISCHARGE_DATE_MGMT
PX_DATE_MGMT
RX_ORDER_DATE_MGMT
RX_START_DATE_MGMT
RX_END_DATE_MGMT
DISPENSE_DATE_MGMT
LAB_ORDER_DATE_MGMT
SPECIMEN_DATE_MGMT
RESULT_DATE_MGMT
MEASURE_DATE_MGMT
ONSET_DATE_MGMT
REPORT_DATE_MGMT
RESOLVE_DATE_MGMT
PRO_DATE_MGMT
REFRESH_DEMOGRAPHIC_DATE
REFRESH_ENROLLMENT_DATE
REFRESH_ENCOUNTER_DATE
REFRESH_DIAGNOSIS_DATE
REFRESH_PROCEDURES_DATE
REFRESH_VITAL_DATE
REFRESH_DISPENSING_DATE
REFRESH_LAB_RESULT_CM_DATE
REFRESH_CONDITION_DATE
REFRESH_PRO_CM_DATE
REFRESH_PRESCRIBING_DATE
REFRESH_PCORNET_TRIAL_DATE
REFRESH_DEATH_DATE
REFRESH_DEATH_CAUSE_DATE

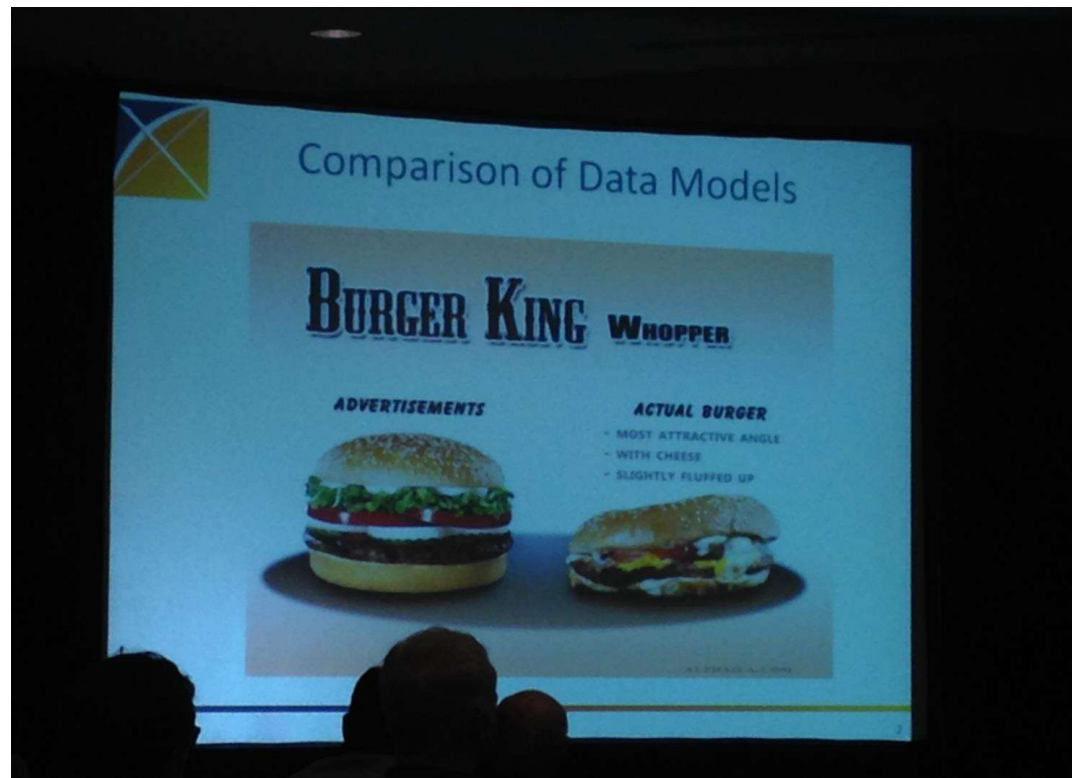
Process-related data

<http://www.pcornet.org/resource-center/pcor-net-common-data-model/>

**Bold font** indicates fields that cannot be null due to primary key definitions or record-level constraints.







“best slide ever” – from AMIA CRI Summit, 2015-03-27 Panel,  
Parsa Mirhaji; Shawn N. Murphy; Christian G. Reich; Keith Marsolo.  
“Tug of Ontologies: How Many Information Models Does It Take to Weave a  
Nationwide Clinical Data Research Network?”



# Comparisons of OMOP vs PCORnet

- <http://forums.ohdsi.org/t/omop-data-model-alternatives/406>  
(several posters listed and good / recent discussion)
- <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3900207/>
- <https://www.ncbi.nlm.nih.gov/pubmed/23774519>
- <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3824370/>

## Pragmatic Data Domain Selection for a National Distributed Research Network: The PCORnet Common Data Model Strategy

Shelley A. Rusincovitch<sup>1</sup>, Abel N. Kho, MD, MS<sup>2</sup>, Jon E. Puro, MPA:HA<sup>3</sup>,  
Daniella Meeker, PhD<sup>4</sup>, Pedro Rivera, MSCS<sup>3</sup>, Aaron A. Sorensen, MA<sup>5</sup>,  
Jeffrey S. Brown, PhD<sup>6</sup>, and Lesley H. Curtis, PhD<sup>7</sup>

<sup>1</sup>Duke Translational Medicine Institute, Durham, NC; <sup>2</sup>Northwestern University Departments of Medicine and Preventive Medicine, Evanston, IL; <sup>3</sup>OCHIN, Inc., Portland, OR; <sup>4</sup>Department of Health, RAND Corporation, Santa Monica, CA; <sup>5</sup>Temple University School of Medicine, Philadelphia, PA; <sup>6</sup>Department of Population Medicine, Harvard Medical School and Harvard Pilgrim Health Care Institute, Boston, MA; <sup>7</sup>Duke University School of Medicine, Department of Medicine, Durham, NC

### Abstract

*The PCORnet Common Data Model (CDM) is the foundation for the PCORI national distributed research network. We describe our experiences in assessing potential data domains and making decisions for inclusion in the CDM, including modeling attributes, dimensions of assessment, and lessons learned.*

### Introduction and Background

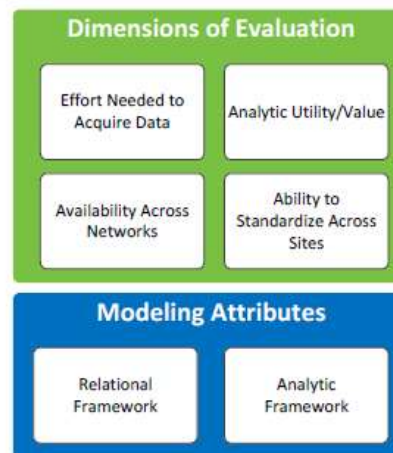
The PCORnet Common Data Model (CDM) specifies the data foundation for the national distributed research network under development by the Patient-Centered Outcomes Research Institute (PCORI). The PCORnet CDM is developed with a phase-based approach, with each phase incorporating new concepts and data tables to support distributed clinical research (observational and interventional). The first version of the CDM established six tables reflecting key patient-level data captured routinely within healthcare delivery and billing systems. In order to establish priorities for subsequent CDM development, it was necessary to establish a method of assessing new concepts and making decisions for inclusion to serve the functional, pragmatic focus of the initiative.

### Methods

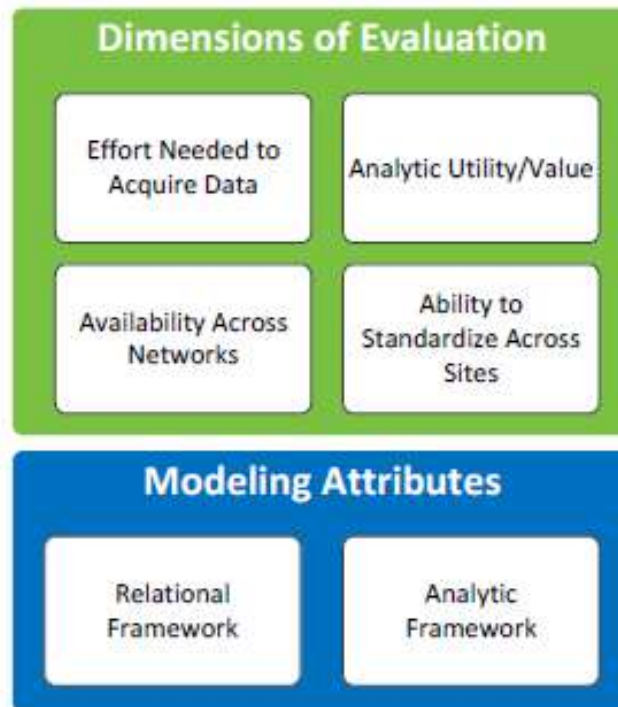
The assessment was organized by data domains; i.e., the high-level concepts of data organization based upon existing data sources, workflows, and processes. Our assessment included best practices established by existing data models and advice from external experts for specific topics. We chose four dimensions for assessment: Effort to acquire data; analytic value of data; ability to standardize data; and availability of data. Each of these dimensions was classified using a simple high, moderate, or low ranking. The CDM Working Group (CDM WG), initially convened in 3 meetings during the summer of 2014 to evaluate and prioritize new data domains for the CDM.

### Results

During development and modeling of domains we paid close attention to PCORnet specific requirements, such as



**Figure 1.** Overview of the data domain evaluation and modeling elements.



**Figure 1.** Overview of the data domain evaluation and modeling elements.

# Key Points

- Research networks and collaborations have formed and have potential to generate evidence.
- Common data models are being used.
- These data models developed from with data that is widely available in EHRs; many gaps exist.
- **Future full of opportunities to leverage and expand these networks and (data, models) to facilitate evidence and discovery on a national scale.**

# Computable Phenotype Definition

- Specifications for identifying patients or populations with a given characteristic or condition of interest from EHRs using data that are routinely collected in EHRs or ancillary data sources.
- EHR-based condition definition

# Example

Diabetes defined as<sup>1</sup>:

- one inpatient discharge diagnosis (ICD-9-CM 250.x, 357.2, 366.41, 362.01-362.07)

ICD-9  
codes

or any combination of two of the following events occurring within 24 months of each other:

- A1C  $\geq$  6.5% (48 mmol/mol)
- fasting plasma glucose  $\geq$  126 mg/dl (7.0 mmol/L)
- random plasma glucose  $\geq$  200 mg/dl (11.1 mmol/L)
- 2-h 75-g OGTT  $\geq$  200 mg/dl
- outpatient diagnosis code (same codes as inpatient)
- anti-hyperglycemic medication dispense (see details below)
- NDC in associated list
- **...etc., etc...**

Lab  
codes

Medication  
codes

1. Nichols GA, Desai J, Elston Lafata J, et al. Construction of a Multisite DataLink Using Electronic Health Records for the Identification, Surveillance, Prevention, and Management of Diabetes Mellitus: The SUPREME-DM Project. Prev Chronic Dis. 2012;9:110311.



## A comparison of phenotype definitions for diabetes mellitus

Rachel L Richesson,<sup>1</sup> Shelley A Rusincovitch,<sup>2</sup> Douglas Wixted,<sup>3</sup> Bryan C Batch,<sup>4</sup> Mark N Feinglos,<sup>4</sup> Marie Lynn Miranda,<sup>5</sup> W Ed Hammond,<sup>2,6</sup> Robert M Califf,<sup>3,7</sup> Susan E Spratt<sup>4</sup>

► Additional material is published online only. To view please visit the journal online (<http://dx.doi.org/10.1136/amiajnl-2013-001952>).

<sup>1</sup>Duke University School of Nursing, Durham, North Carolina, USA

<sup>2</sup>Applied Informatics Research, Duke Health Technology Solutions, Durham, North Carolina, USA

<sup>3</sup>Duke Translational Medicine Institute, Durham, North Carolina, USA

<sup>4</sup>Division of Endocrinology, Metabolism and Nutrition, Department of Medicine, Duke University School of Medicine, Durham, North Carolina, USA

<sup>5</sup>Department of Pediatrics, School of Natural Resources and Environment, University of Michigan, Ann Arbor, Michigan, USA

<sup>6</sup>Duke Center for Health Informatics, Durham, North Carolina, USA

<sup>7</sup>Division of Cardiology, Department of Medicine, Duke University School of Medicine, Durham, North Carolina, USA

**Correspondence to** Dr Rachel L Richesson, Duke University School of Nursing, 2007 Pearson Building, 307 Trent Drive, Durham, NC 27710, USA; [rachel.richesson@duke.edu](mailto:rachel.richesson@duke.edu)

Received 19 April 2013

Revised 30 July 2013

Accepted 20 August 2013

Published Online First

11 September 2013

**To cite:** Richesson RL, Rusincovitch SA, Wixted D, et al. *J Am Med Inform Assoc* 2013;20:e319–e326.

### ABSTRACT

**Objective** This study compares the yield and characteristics of diabetes cohorts identified using heterogeneous phenotype definitions.

**Materials and methods** Inclusion criteria from seven diabetes phenotype definitions were translated into query algorithms and applied to a population (n=173 503) of adult patients from Duke University Health System. The numbers of patients meeting criteria for each definition and component (diagnosis, diabetes-associated medications, and laboratory results) were compared.

**Results** Three phenotype definitions based heavily on ICD-9-CM codes identified 9–11% of the patient population. A broad definition for the Durham Diabetes Coalition included additional criteria and identified 13%. The electronic medical records and genomics, NYC A1c Registry, and diabetes-associated medications definitions, which have restricted or no ICD-9-CM criteria, identified the smallest proportions of patients (7%). The demographic characteristics for all seven phenotype definitions were similar (56–57% women, mean age range 56–57 years). The NYC A1c Registry definition had higher average patient encounters (54) than the other definitions (range 44–48) and the reference population (20) over the 5-year observation period. The concordance between populations returned by different phenotype definitions ranged from 50 to 86%. Overall, more patients met ICD-9-CM and laboratory criteria than medication criteria, but the number of patients that met abnormal laboratory criteria exclusively was greater than the numbers meeting diagnostic or medication data exclusively.

**Discussion** Differences across phenotype definitions can potentially affect their application in healthcare organizations and the subsequent interpretation of data.

**Conclusions** Further research focused on defining the clinical characteristics of standard diabetes cohorts is important to identify appropriate phenotype definitions for health, policy, and research.

### INTRODUCTION

The ability to identify people with diabetes across healthcare organizations by using a common definition has value for clinical quality, health improvement, and research. Registries have been shown to improve care in diabetes, and are the cornerstone of the chronic disease care model.<sup>1–2</sup> Standard phenotype definitions can enable direct comparison of population characteristics, risk factors, and complications, allowing decision makers to identify and target patients for interventions demonstrated in similar

populations. Furthermore, standard phenotype definitions can streamline the development of patient registries from healthcare data, and enable consistent inclusion criteria to support regional surveillance and the identification of rare disease complications. An understanding of the populations generated from various phenotype definitions will inform standard methods for identifying diabetes cohorts, facilitate the rapid generation of patient registries and research datasets with uniform sampling criteria, and enable comparative and aggregate analysis. This descriptive study presents and compares the size and characteristics of patient populations retrieved using different phenotype definitions adopted from prominent diabetes registries and research networks, a large community intervention program in our county, and federal reporting standards.

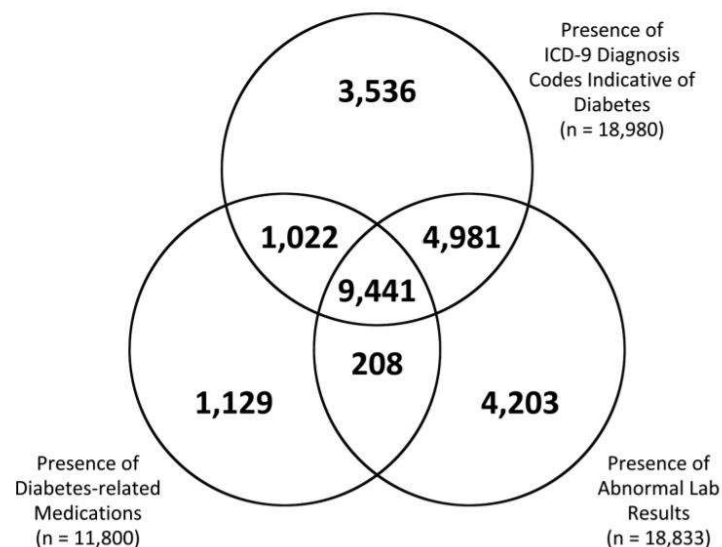
### BACKGROUND AND SIGNIFICANCE Diabetes diagnosis and management

Diabetes is a complex disease with multiple subtypes associated with different etiologies, diagnostic indicators, and clinical management strategies. Type 2 diabetes mellitus (T2DM) is the most common (95%) type of diabetes in the USA and can be treated with diet and exercise, oral medication, or insulin. Type 1 diabetes mellitus (T1DM) is less common and requires treatment with insulin. Rare types of diabetes result from drug interactions, genetic defects of beta cell or insulin action function, pancreatic disorders, and inherited endocrine disorders. All types of diabetes manifest in high blood glucose, and laboratory values are the primary means for diagnosis and management.<sup>3</sup>

### Diabetes-relevant data available for electronic health record-based phenotyping

Data from three domains (International Classification of Disease, revision 9, clinical modification (ICD-9-CM) coded diagnoses, laboratory test results, and medication data) in varying combinations and thresholds constitute most phenotype definitions used for diabetes cohort identification. The ICD-9-CM coding system has more than 20 broad codes (and scores of higher precision codes) suggestive or indicative of diabetes (presented in the diabetes phenotype definition shared on Phenotype KnowledgeBase),<sup>4</sup> and is a critical component of most queries and phenotypes. However, ICD-9-CM has been shown to be insufficient for capturing etiology, subtypes, or all cases of diabetes.<sup>5–7</sup>

Diabetes-related medications are often included in phenotype definitions because medication data



**Figure 1** Overlap of diabetes cohorts identified from different categories of phenotype eligibility criteria; n=24 520 patients identified by criteria from any of the three categories.

Richesson RL, Rusincovitch SA, Wixted D, et al. A comparison of phenotype definitions for diabetes mellitus. *J Am Med Inform Assoc*. 2013;20(e2):e319–e326. doi:10.1136/amiajnl-2013-001952

**Table 1** Data domain criteria used in selected phenotype definitions

Phenotype definitions:	Data domain criteria							
	ICD-9-CM 250.xx	ICD-9-CM 250.x0 and 250.x2 (excludes type 1 specific codes)	Expanded ICD-9-CM Codes (249.xx, 357.2, 362.0x, 366.41)	HbA1c	Fasting glucose	Random glucose	Abnormal OGTT	Diabetes-associated medications*
ICD-9-CM 250.xx	●							
CMS CCW	▲*//		▲*//					
NYC A1c Registry				●				
Diabetes-associated medications								●
DDC		▲	▲	▲//	▲//	▲//	▲//	▲
SUPREME-DM	▲*//		▲*//	▲//	▲//	▲//	▲	▲
eMERGE†		●*//		▲	▲	▲		▲

\*Medications vary by phenotype definition and are listed for each in the supplementary appendix (available online only).

†The eMERGE phenotype definition consists of five case scenarios with varying combinations of criteria. Any instance of type 1 specific codes (ie, 250.x1, 250.x3) results in the exclusion of the patient.

●=Sole criteria.

▲=Optional criteria, one of many.

\*=Distinction made between inpatient and outpatient context.

// = Distinction made for multiple instances and/or time points.

CMS CCW, Centers for Medicare and Medicaid Services Chronic Condition Data Warehouse; DDC, Durham Diabetes Coalition; eMERGE, electronic medical records and genomics; HbA1c, hemoglobin A1c; ICD-9-CM, International Classification of Disease, revision 9, clinical modification; NYC, New York City; OGTT, oral glucose tolerance test; SUPREME-DM, Surveillance, Prevention, and Management of Diabetes Mellitus.



# Benefits from Standard Phenotypes...

- Development and conduct of new multi-site studies (interventional and observational)
  - Efficiencies of re-using executable phenotype code
- Comparability of EHR-derived data sets
- Comparison of study results and aggregation of evidence
- Reporting of data sets or results (e.g., ClinicalTrials.gov, NIH)
- Description of research populations in medical journals

RESEARCH



HEALTH  
CARE

- Ideally, ***research and clinical definitions should be semantically equivalent.***  
i.e., they should identify equivalent populations.

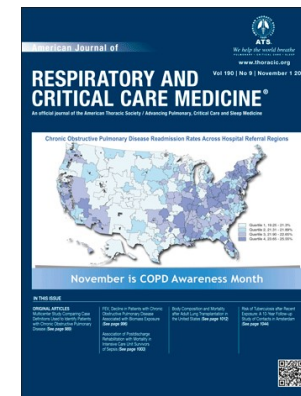
## ORIGINAL ARTICLE

# Multicenter Study Comparing Case Definitions Used to Identify Patients with Chronic Obstructive Pulmonary Disease

Valentin Prieto-Centurion<sup>1</sup>, Andrew J. Rolle<sup>1</sup>, David H. Au<sup>2</sup>, Shannon S. Carson<sup>3</sup>, Ashley G. Henderson<sup>3</sup>, Todd A. Lee<sup>4</sup>, Peter K. Lindenauer<sup>5,6</sup>, Mary A. McBurnie<sup>7</sup>, Richard A. Mularski<sup>7</sup>, Edward T. Naureckas<sup>8</sup>, William M. Vollmer<sup>7</sup>, Binoy J. Joese<sup>9</sup>, and Jerry A. Krishnan<sup>1,9</sup>; on behalf of the CONCERT Consortium

<sup>1</sup>Division of Pulmonary, Critical Care, Sleep and Allergy and <sup>4</sup>Department of Pharmacy Systems, Outcomes and Policy, University of Illinois at Chicago, Chicago, Illinois; <sup>2</sup>University of Washington/VA Puget Sound, Seattle, Washington; <sup>3</sup>Division of Pulmonary and Critical Care Medicine, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina; <sup>5</sup>Department of Medicine and Center for Quality of Care Research, Baystate Medical Center, Springfield, Massachusetts; <sup>6</sup>Tufts University School of Medicine, Boston, Massachusetts; <sup>7</sup>The Center for Health Research, Kaiser Permanente, Portland, Oregon; <sup>8</sup>Section of Pulmonary and Critical Care, University of Chicago Medicine, Chicago, Illinois; and <sup>9</sup>Population Health Sciences Program, University of Illinois Hospital and Health Sciences System, Chicago, Illinois

Am J Respir Crit Care Med. 2014 Nov 1;190(9):989-95.  
doi: 10.1164/rccm.201406-1166OC.



**Table 2.** Clinical Characteristics of Patients Who Met and Did Not Meet the Clinical Trial Reference Standard

Characteristic	Total Sample (n = 998)	Clinical Trial Reference Standard		P Value
		Yes* (n = 560)	No† (n = 438)	
Comorbid conditions, %				
Cardiovascular disease	76	74	78	0.15
Hypertension	66	63	69	0.03
Heart failure	18	16	22	0.01
Coronary artery disease	23	22	24	0.66
Myocardial infarction	19	18	20	0.43
Stroke	15	14	15	0.95
Depression	42	36	50	<0.0001
Arthritis	36	33	41	0.006
Diabetes	28	22	34	<0.0001
Cancer history	23	26	19	0.02
Anemia	28	26	30	0.17
Kidney disease	20	18	21	0.30
Dementia	2	2	3	0.15
Dyspnea at rest (Borg), %				
0, no dyspnea	52	54	50	0.02
0.5–2, slight	38	38	37	
≥3, moderate to very severe	10	7	13	
Spirometry, post-bronchodilator, %				
FEV <sub>1</sub> /FVC <70%	61	100	11	<0.0001
FEV <sub>1</sub> <80% predicted	72	86	55	<0.0001
6-minute-walk distance, %				
Distance walked <350 m	53	52	54	0.67

Patients who met the trial reference standard are more likely to have airflow obstruction by spirometry but report being less dyspneic. Patients who met the reference standard also have different prevalence of comorbidities. For example, they are more likely to have hypertension, heart failure, and depression. Data for 6-minute-walk distance missing in 9% patients (9% and 10%) and dyspnea scores missing in 8% patients (8% and 9%) in those who met and did not meet the clinical trial reference standard, respectively.

\* $(A + D + E + G)$  and † $(B + C + F)$  in Figure 2.

**Table 3.** Characteristics Associated with Meeting the Clinical Trial Reference Standard

Characteristics	Odds Ratio (95% CI)
Race (vs. white)	
Black	0.37 (0.26–0.53)*
Other	0.52 (0.27–1.00)
Education (vs. high school or less)	
College/professional degree	0.38 (0.26–0.56)*
Some college	0.68 (1.06–2.03)*
BMI, kg/m <sup>2</sup> (vs. normal)	
<18.5 (underweight)	4.00 (1.27–12.50)*
25–29.99 (overweight)	0.87 (0.58–1.30)
≥30 (obese)	0.51 (0.35–0.75)*
Depression (yes vs. no)	0.53 (0.40–0.71)*
Diabetes (yes vs. no)	0.67 (0.48–0.93)*
Cancer (yes vs. no)	1.47 (1.05–2.08)*

*Definition of abbreviations:* BMI = body mass index; CI = confidence interval.

Clinical trial reference standard (*A + D + E + G*) versus others (*B + C + F*) in Figure 2. Multivariable logistic regression model that included characteristics listed in Tables 1 and 2 (characteristics significantly associated with meeting the trial reference standard). Results indicate that patients who are black (vs. white), with college or higher (vs. high school or less) education, obese (vs. normal weight), with depression, or diabetes are less likely to meet the trial reference standard. Patients with a history of cancer and underweight patients (vs. normal weight) are more likely to meet the trial reference standard. Hosmer-Lemeshow goodness-of-fit test (*P* value = 0.17) demonstrates adequate model fit.

\**P* < 0.05.

# Lots of Phenotypes

- >75 phenotype/cohort definitions



- ~40 public (92 private)



- 19 in PCORnet



# Other Sources for Phenotypes

- Clinical Classifications Software , “AHRQ Bundles”
- CMS Chronic Conditions Warehouse
- Quality Net (CMS and Joint Commission)
- Mini-Sentinel (FDA)
- SHARPN
- .....
- Multi-site registries
- Research networks





Welcome

[Search Value Sets](#)

[Download](#)

[Help](#)

[Apply Filters](#) [Clear Filters](#)

Search the NLM Value Set Repository

Query:

[Search](#)

Narrow search results by selecting from pull-down menus below:

CMS eMeasure (NQF

Number)

[Select](#)

Quality Data Model

Category

[Select](#)

Steward

[Select](#)

Meaningful Use Measures

[Select](#)

Code System

[Select](#)

[Search Results](#)

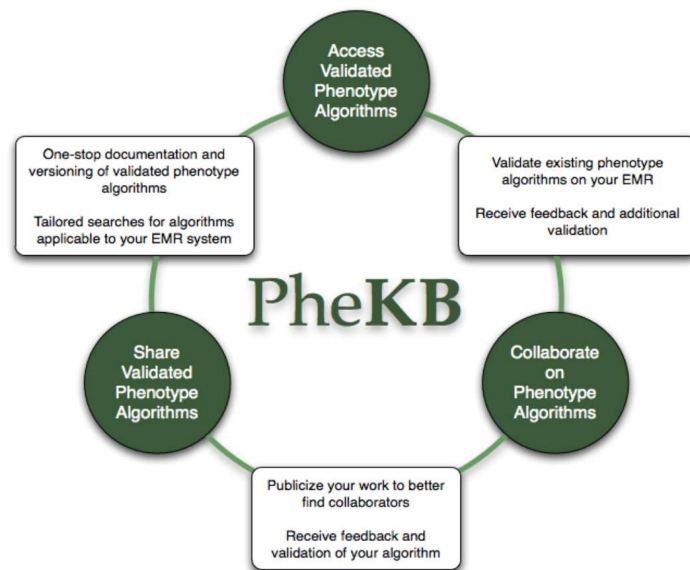
[Value Set Details](#)

[Export Search Results \(Excel\)](#)

Matched Value Sets

<input type="checkbox"/>	Name	Type	Code System	Steward	OID
<input type="checkbox"/>	Complications of Pregnancy, Childbirth and the Puerperium	Extensional	SNOMEDCT	NCQA	<a href="#">2.16.840.1.113883.3.464.1003.111.11.1023</a>
<input type="checkbox"/>	Complications of Pregnancy, Childbirth and the Puerperium	Extensional	ICD10CM	NCQA	<a href="#">2.16.840.1.113883.3.464.1003.111.11.1022</a>
<input type="checkbox"/>	Conditions Possibly Justifying Elective Delivery Prior to 37 Weeks Gestation	Grouping	ICD10CM SNOMEDCT	The Joint Commission	<a href="#">2.16.840.1.113883.3.117.1.7.1.286</a>
<input type="checkbox"/>	Conditions Possibly Justifying Elective Delivery Prior to 37 Weeks Gestation	Extensional	ICD9CM	The Joint Commission	<a href="#">2.16.840.1.113883.3.117.1.7.1.394</a>
<input type="checkbox"/>	Conditions Possibly Justifying Elective Delivery Prior to 37 Weeks Gestation	Extensional	ICD10CM	The Joint Commission	<a href="#">2.16.840.1.113883.3.117.1.7.1.393</a>
<input type="checkbox"/>	Conditions Possibly Justifying Elective Delivery Prior to 37 Weeks Gestation	Extensional	SNOMEDCT	The Joint Commission	<a href="#">2.16.840.1.113883.3.117.1.7.1.395</a>
<input type="checkbox"/>	Degeneration of Macula and Posterior Pole	Grouping	ICD10CM ICD9CM SNOMEDCT	AMA-PCPI	<a href="#">2.16.840.1.113883.3.526.3.1453</a>
<input type="checkbox"/>	Degeneration of Macula and Posterior Pole	Extensional	SNOMEDCT	AMA-PCPI	<a href="#">2.16.840.1.113883.3.526.2.1643</a>
<input type="checkbox"/>	Delivery	Extensional	ICD10CM	Optum	<a href="#">2.16.840.1.113762.1.4.1078.3</a>
<input type="checkbox"/>	Delivery - Diagnosis	Grouping	ICD10CM ICD9CM SNOMEDCT	Optum	<a href="#">2.16.840.1.113883.3.67.1.101.1.278</a>
<input type="checkbox"/>	Delivery ICD9Dx	Extensional	ICD9CM	Optum	<a href="#">2.16.840.1.113883.3.67.1.101.1.83</a>
<input type="checkbox"/>	Diabetes	Extensional	ICD10CM	NCQA	<a href="#">2.16.840.1.113883.3.464.1003.103.11.1002</a>
<input type="checkbox"/>	Diabetes	Grouping	ICD10CM ICD9CM SNOMEDCT	NCQA	<a href="#">2.16.840.1.113883.3.464.1003.103.12.1001</a>
<input type="checkbox"/>	Diabetes	Extensional	ICD9CM	NCQA	<a href="#">2.16.840.1.113883.3.464.1003.103.11.1001</a>
<input type="checkbox"/>	Diabetes	Extensional	SNOMEDCT	NCQA	<a href="#">2.16.840.1.113883.3.464.1003.103.11.1003</a>
<input type="checkbox"/>	Diabetes	Extensional	CPT	NCQA	<a href="#">2.16.840.1.113883.3.464.1003.103.11.1004</a>

## What is the Phenotype KnowledgeBase?



purposefully integrated tools and standards that guide the user in efficiently navigating each of these stages from early stage development to public sharing and reuse. PheKB

Health Data is becoming an increasing important source for clinical and genomic research. Researchers create and iteratively refine algorithms using structured and unstructured data to better identify cohorts of subjects within the health data.

The Phenotype Knowledgebase website, PheKB, is a collaborative environment to building and validating electronic algorithms to identify characteristics of patients within health data. PheKB was functionally designed to enable such a workflow and has

### Most Recent Phenotypes

HIV
Functional seizures
RxNorm RxCUI codes for Cancer Therapies
Type 1 Diabetes
Body Mass Index (BMI)














<https://phekb.org/>

Activity – explore PheKB



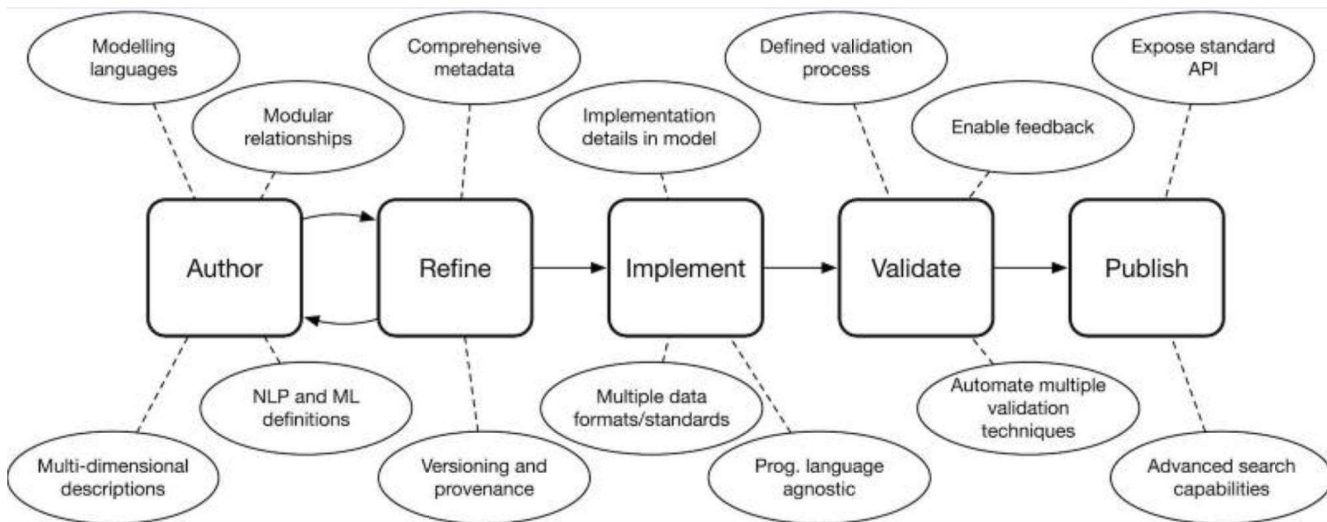
REVIEW

## Desiderata for the development of next-generation electronic health record phenotype libraries

Martin Chapman <sup>1,\*</sup>, Shahzad Mumtaz <sup>2</sup>, Luke V. Rasmussen <sup>3</sup>,  
Andreas Karwath <sup>4</sup>, Georgios V. Gkoutos <sup>4</sup>, Chuang Gao <sup>2</sup>,  
Dan Thayer <sup>5</sup>, Jennifer A. Pacheco <sup>3</sup>, Helen Parkinson <sup>6</sup>, Rachel  
L. Richesson <sup>7</sup>, Emily Jefferson <sup>2</sup>, Spiros Denaxas <sup>8</sup> and Vasa Curcin <sup>1</sup>

<sup>1</sup>Department of Population Health Sciences, King's College London, London, SE1 1UL, UK; <sup>2</sup>Health Informatics Centre (HIC), University of Dundee, Dundee, DD1 9SY, UK; <sup>3</sup>Feinberg School of Medicine, Northwestern University, Chicago, IL 60611, USA; <sup>4</sup>Institute of Cancer and Genomic Sciences, University of Birmingham, Birmingham, B15 2TT, UK; <sup>5</sup>SAIL Databank, Swansea University, Swansea, SA2 8PP, UK; <sup>6</sup>European Molecular Biology Laboratory, European Bioinformatics Institute, Hinxton, CB10 1SD, UK; <sup>7</sup>Department of Learning Health Sciences, University of Michigan Medical School, MI 48109, USA and <sup>8</sup>Institute of Health Informatics, University College London, London, NW1 2DA, UK

\*Correspondence address. Martin Chapman, 3.07 Addison House, Guy's Campus, King's College London, London, SE1 1UL, UK. E-mail: [martin.chapman@kcl.ac.uk](mailto:martin.chapman@kcl.ac.uk)  <http://orcid.org/0000-0002-5242-9701>



**Table 1:**

Phenotype definition formats

<b>Format</b>	<b>Description</b>	<b>Example</b>	<b>Category</b>
Code list	A set of codes that must exist in a patient's health record in order to include them within a phenotype cohort	COVID-19 ICD-10 code "U07.1"	Rule-based
Simple data elements	Formalizing the relationship between code-based data elements using logical connectives	COVID-19 ICD-10 code "U07.1" AND ICD-11 code "RA01.0"	Rule-based
Complex data elements	Formalizing the relationship between complex data elements, such as those derived via NLP	Patient's blood pressure reading >140 OR patient notes contain "high BP"	Rule-based
Temporal	Prefix rules with temporal qualifiers	Albumin levels increased by 25% over 6 hours, high blood pressure reading has to occur during hospitalization	Rule-based
Trained classifier	Use rule-based definitions as the basis for constructing a classifier for future (or additional) cohorts	A <i>k</i> -fold cross-validated classifier capable of identifying patients with COVID-19	Probabilistic



**Table 2:**

## Phenotype validation mechanisms

<b>Mechanism</b>	<b>Description</b>	<b>Example</b>
Disease registries	Compare the phenotype cohort with those present in the registry	Comparison of a diabetes phenotype cohort with those patients present in a diabetes registry (e.g., T1D exchange)
Chart review	Compare the phenotype cohort with the patients identified by manual review of medical records	Comparison with a diabetes gold standard, produced by double manual review of patient medical records
Cross-EHR concordance	Compare percentage of cases identified by a phenotype across different sources, and identify any overlap	Comparison of the percentage of patients identified by a diabetes phenotype in primary and secondary care EHRs, and the identification of any case overlap
Risk factors	Compare the magnitude of the phenotype cohort with standard risk calculations	Comparison with the output of a Cox hazards model
Prognosis	Compare the magnitude of the phenotype cohort with external prognosis models	Comparison with a survival analysis
Genetic associations	Compare whether the presence of a patient in a phenotype cohort is consistent with their genetic profile	A patient is more likely to be a valid member of a diabetes cohort if they have the HLA-DR3 gene



# Desiderata (14)

- Support modelling languages
- Support NLP-based and machine learning-based definitions
- Support multi-dimensional descriptions
- Support versioning and data provenance
- Support modular relationships between phenotypes
- Communicate implementation information in the model
- Support tooling that provides multiple programming language implementations
- Support tooling that provides connectivity with multiple data standards
- Support a defined validation process
- Automate multiple validation techniques
- Enable feedback
- Expose a standard API
- Offer advanced search capabilities
- Include comprehensive metadata

Sections: [modelling](#), [logging](#), implementation, [validation](#), and [sharing and warehousing](#)

# Real-World Evidence



## **Real-world data (RWD) and real-world evidence (RWE) are playing an increasing role in health care decisions.**

- FDA uses RWD and RWE to monitor postmarket safety and adverse events and to make regulatory decisions.
- The health care community is using these data to support coverage decisions and to develop guidelines and decision support tools for use in clinical practice.
- Medical product developers are using RWD and RWE to support clinical trial designs (e.g., large simple trials, pragmatic clinical trials) and observational studies to generate innovative, new treatment approaches.

The 21st Century Cures Act, passed in 2016, places additional focus on the use of these types of data to support regulatory decision making, including approval of new indications for approved drugs. Congress defined RWE as data regarding the usage, or the potential benefits or risks, of a drug derived from sources other than traditional clinical trials. FDA has expanded on this definition as discussed below.

### **Why is this happening now?**

The use of computers, mobile devices, wearables and other biosensors to gather and store huge amounts of health-related data has been rapidly accelerating. This data holds potential to allow us to better design and conduct clinical trials and studies in the health care setting to answer questions previously though infeasible. In addition, with the development of sophisticated, new analytical capabilities, we are better able to analyze these data and apply the results of our analyses to medical product development and approval.

<https://www.fda.gov/science-research/science-and-research-special-topics/real-world-evidence>

# Conclusions

- CBK is complex and dynamic (life cycle)
- CBK representation not standardized
- CBK will interact with \*data\* - hence data standards are relevant
- Interaction with data can differ across settings and time
- Context is important to represent – and challenging
- Computable phenotypes are one example of CBK
- Phenotype metadata and libraries will continue to evolve



## *A multi-stakeholder movement to mobilize computable knowledge*

A screenshot of the MCBK website homepage. The page features the MCBK logo in the top left corner and a navigation menu with links for 'HOME', 'JOIN MCBK', 'NEWS & EVENTS', and 'WORKGROUPS'. The main content area contains two paragraphs of text: 'We believe every decision affecting the health of individuals and populations should be informed by the best available knowledge.' and 'Mobilizing Computable Biomedical Knowledge is an international community from academia, the sciences, and government working together to ensure that biomedical knowledge in computable form is findable, accessible, interoperable, and reusable.' Below the text is a 'WATCH VIDEO' button. At the bottom of the page is a banner with a collage of images showing diverse people in professional settings, and a call to action: 'Join us as we mobilize health knowledge on a global scale.'

**MCBK**  
Mobilizing Computable Biomedical Knowledge

HOME JOIN MCBK NEWS & EVENTS WORKGROUPS

We believe every decision affecting the health of individuals and populations should be informed by the best available knowledge.

Mobilizing Computable Biomedical Knowledge is an international community from academia, the sciences, and government working together to ensure that biomedical knowledge in computable form is findable, accessible, interoperable, and reusable.

WATCH VIDEO

Join us as we mobilize health knowledge on a global scale.

**#MobilizeCBK**  
**Mobilizecbk.org**

**Mobilizing Computable Biomedical Knowledge (CBK):  
A Manifesto**

**Preamble**

Knowledge has the potential to improve health care, the health of individuals, and the health of populations. Every decision affecting health should be informed by the best available knowledge. For moral and ethical reasons, it is imperative that each and every member of society has access to what is known at the time they are making health-related choices and decisions.

It is no longer sufficient to represent knowledge in the form of printed words and static pictures. The increasingly rapid rate of scientific discovery needs knowledge representations that are more agile and amenable to scalability and mass action. This in turn can enable the continuous cycles of discovery and improvement envisioned as Learning Health Systems.

Contemporary digital technology enables knowledge to be represented in *computable* forms expressed in machine-executable code. Computable knowledge unleashes the potential of information technology to generate and deliver useful information—and particularly, decision-specific advice—to individuals and organizations with great speed on a world-wide scale. It is essential to take full advantage of these capabilities, while continuing established practices that validate knowledge, preserve it, and ensure that it can be trusted.

There is work to do to mobilize best available health knowledge for the greater good. To begin, biomedical knowledge in computable form must be made interoperable using open standards, and widely available so that it can be used to immediately impact health.

It is time for action on a global scale.

**Computable Biomedical Knowledge**

Computable Biomedical Knowledge is the result of an analytic and/or deliberative process about human health, or affecting human health, that is explicit, and therefore can be represented and reasoned upon using logic, formal standards, and mathematical approaches.

**Vision**

We are dedicated to:

*Mobilizing biomedical knowledge that can support action toward improving human health. This should be done using computable formats that can be shared and integrated into health information systems and applications.*

*Efficiently and equitably serving the learning and knowledge needs of all participants, as well as the public good. This will work to significantly reduce health disparities.*

# MCBK Manifesto

*Ensuring that the knowledge properly reflects the best and most current evidence and science. This will ensure that knowledge can be trusted for use to improve health and health care.*

*Achieving this through evolution of an open Computable Biomedical Knowledge ecosystem dedicated to achieving the FAIR principles: making Computable Biomedical Knowledge easily findable, universally accessible, highly interoperable, and readily reusable.\* The current interest in making data "FAIR" should be matched by equally intense interest in making knowledge "FAIR".*

**Mechanisms of Activity**

We believe that all of the following are important:

- The CBK Concept
  - Sustain the Computable Biomedical Knowledge ecosystem through public-private partnerships.
  - Establish broadly-based participatory governance of the ecosystem.
  - Make the ecosystem diverse and inclusive.
  - Explore the sciences of Computable Biomedical Knowledge collaboratively.
  - Be agile to reflect the increasingly rapid changes in knowledge.
- The CBK Technical System
  - Enable the ecosystem with open standards.
  - Build and uphold trust in Computable Biomedical Knowledge through the ecosystem.
  - Ensure robust and unbiased methods to support transparency and expose the currency, validity and provenance of Computable Biomedical Knowledge.
  - Implement the highest standards of privacy and security for all stakeholders.
  - Enable a pipeline that transitions knowledge from human-readable to fully computable through successive stages.
- The CBK Use/User System
  - Ensure the safe and effective use of Computable Biomedical Knowledge through the ecosystem.
  - Generate value for the creators of the knowledge, the users of the knowledge, and the general public.
  - Engender equity in health and in knowledge accessibility

\*: Wilkinson MD, Dumontier M, Aalbersberg LJ, Appleton G, Axton M, Baak A, Blomberg N, Bolten JW, da Silva Santos LB, Bourne PE, Bouwman J. The FAIR Guiding Principles for scientific data management and stewardship. Scientific data: 2016;3.

[www.MobilizeCBK.org](http://www.MobilizeCBK.org)

# Workgroups & Co-Chairs

- **Standards & Infrastructure**

*Bruce Bray  
Jamie McCusker*

- **Sustainability for Mobilization & Inclusion**

*Jerry Perry  
Terrie Wheeler*

- **Policy & Coordination to Ensure Quality & Trust**

*Jodyn Platt  
Blackford Middleton*

# Questions? Follow-up?

Rachel Richesson

[richessr@med.umich.edu](mailto:richessr@med.umich.edu)



# Learning Objectives

- Describe the relevance of CBK to clinical care delivery, learning health systems, and health improvement
- List types of metadata categories that are important for managing CBK
- List 3 challenges for “mobilizing” CBK for action (in health systems)
- Describe role of research networks in developing and implementing CBK
- Describe how common data models (CDMs) and computable phenotypes support the development and application of CBK
- Identify features for libraries of CBK artifacts (e.g. computable phenotypes)
- Describe challenges for managing CBK at scale and highlight areas needing future development and research