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


- OHDSI: https://www.ohdsi.org/

- MCBK: https://mobilizecbk.med.umich.edu/


- 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
In a learning health care system, research influences practice and practice influences research.

EVALUATE
Collect data and analyze results to show what works and what doesn’t.

IMPLEMENT
Apply plan in pilot and control settings.

DESIGN
Design care and evaluation based on evidence generated here and elsewhere.

ADJUST
Use evidence to influence continual improvement.

DISSEMINATE
Share results to improve care for everyone.

INTERNAL AND EXTERNAL SCAN
Identify problems and potentially innovative solutions.
Better Health Requires This

Health Problem of Interest

D2K: Data to Knowledge

K2P: Knowledge to Performance

P2D: Performance to Data
Not Just This
Knowledge to Practice

Health Problem of Interest

D2K: Data to Knowledge

K2P: Knowledge to Performance

P2D: Performance to Data

Guidelines

Articles

Local Analytical Results
Knowledge should be FAIR*

- Findable
- Accessible
- Interoperable
- Reusable

*FAIR: [https://www.force11.org/group/fairgroup/fairprinciples](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/17idZFaz785807xQhHu9dfL9GR1WhFeqE?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
Arrest Assist - Reversible Causes of PEA Arrest Tool

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.)?
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.

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 trustable (FAIR+T). To help mobilize CBKs, we describe our efforts to outline metadata categories to make CBKs FAIR+T.

<table>
<thead>
<tr>
<th>Metadata category</th>
<th>Metadata elements in this category</th>
<th>Example predicates</th>
<th>Main principle supported</th>
<th>From</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Type</td>
<td>Elements that classify CBKs by describing the nature of CBKs in some general way</td>
<td>[CBK] is_a [type]</td>
<td>FINDABLE</td>
<td>49,50</td>
</tr>
<tr>
<td>2. Domain</td>
<td>Elements relating CBKs to the biomedical domains or topics to which they belong</td>
<td>[CBK] is_about [domain]</td>
<td>FINDABLE</td>
<td>51,52</td>
</tr>
<tr>
<td>3. Purpose</td>
<td>Elements describing the purposes or circumscribing and limiting the intended uses of CBKs</td>
<td>[CBK] has_purpose_of _____</td>
<td>FINDABLE</td>
<td>53</td>
</tr>
<tr>
<td></td>
<td></td>
<td>[CBK] is_intended_to _____</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>[CBK] is_not_intended_to _____</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4. Identification</td>
<td>Elements indicating persistent identifiers or persistent unique identifiers and versions assigned to CBKs</td>
<td>[CBK] has_identifier _____</td>
<td>FINDABLE</td>
<td>49,50</td>
</tr>
<tr>
<td></td>
<td></td>
<td>[CBK] has_name _____</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>[CBK] has_version _____</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5. Location</td>
<td>Elements indicating the physical or virtual locations where CBKs can be accessed</td>
<td>[CBK] has_location [ADDRESS]</td>
<td>ACCESSIBLE</td>
<td>49,50</td>
</tr>
<tr>
<td></td>
<td></td>
<td>[CBK] is_located_at [URL]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6. CBK-to-CBK relationships</td>
<td>Elements describing a relationship between one CBK and some other CBK</td>
<td>[CBK] is_modification_of [CBK]</td>
<td>INTEROPERABLE</td>
<td>49,50</td>
</tr>
<tr>
<td></td>
<td></td>
<td>[CBK] is_predecessor_of [CBK]</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>[CBK] is_successor_of [CBK]</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>[CBK] is_used_with [CBK]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7. Technical</td>
<td>Elements to describe a wide array of technical characteristics of CBKs that need to be known to deploy, integrate, operate, and use them</td>
<td>[CBK] has_file_type _____</td>
<td>INTEROPERABLE</td>
<td>54,55</td>
</tr>
<tr>
<td></td>
<td></td>
<td>[CBK] has_file_size _____</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>[CBK] has_dependency _____</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>[CBK] can_be_executed_using _____</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>[CBK] has_input _____</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>[CBK] has_output _____</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8. Authorization and rights management</td>
<td>Elements describing rights and responsibilities pertaining to CBKs</td>
<td>[CBK] is_available_to [person]</td>
<td>REUSABLE</td>
<td>56</td>
</tr>
<tr>
<td></td>
<td></td>
<td>[CBK] has_license [license]</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>[CBK] copyright Held_by [agent]</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>[CBK] has_disclaimer [disclaimer]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10. Integrity</td>
<td>Elements conveying outputs from cryptographic functions that allow CBK users to confirm CBK has not been tampered with</td>
<td>[CBK] has_hash [hash function output]</td>
<td>REUSABLE 58</td>
<td></td>
</tr>
<tr>
<td>11. Provenance</td>
<td>Elements indicating changes in ownership, custody, and status during CBK lifecycles</td>
<td>[CBK] is_owned_by [agent]</td>
<td>TRUSTABLE 59</td>
<td></td>
</tr>
<tr>
<td></td>
<td>[CBK] ownership_changed_on [date]</td>
<td>[CBK] has status [status]</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>[CBK] status_changed_on [date]</td>
<td>[CBK] is_authored_by [author]</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>[CBK] is_reviewed_by [reviewer]</td>
<td>[CBK] is_endorsed_by [endorser]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Two evidence categories</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12. Evidential basis</td>
<td>Elements describing the data upon which the claims in CBKs are based, the methods of obtaining and analyzing those data, and the strength of the evidential basis of CBKs.</td>
<td>[CBK] is_based_on_data_about _____</td>
<td>TRUSTABLE 2.60-62</td>
<td></td>
</tr>
<tr>
<td></td>
<td>[CBK] is_based_on_data_collected_at [place]</td>
<td>[CBK] is_based_on_data_collected_by [agent]</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>[CBK] is_based_on_data_collected_on [date]</td>
<td>[CBK] is_based_on_data_collected_for _____</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>[CBK] is_based_on_data_analysis_method_of</td>
<td>[CBK] is_based_on_data_analysis_results_of</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>[CBK] has_certainty_of_evidence _____</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13. Evidence from use</td>
<td>Elements describing data arising from CBK use, the methods of obtaining and analyzing those data, and the strength of evidence about CBK use</td>
<td>[CBK] use_is_evaluated_in _____</td>
<td>TRUSTABLE 61-63</td>
<td></td>
</tr>
<tr>
<td></td>
<td>[CBK] use_is_associated_with _____</td>
<td>[CBK] use causes _____</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>[CBK] use_evidence_has_certainty_of _____</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Research agenda item</td>
<td>Brief description of research agenda item</td>
<td>Related metadata category</td>
<td></td>
<td></td>
</tr>
<tr>
<td>-----------------------------------------------------------</td>
<td>------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
<td>-------------------------------------------------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CBK typologies</td>
<td>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.</td>
<td>Type</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Schema for purpose metadata</td>
<td>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.</td>
<td>Purpose</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Schema for CBK-to-CBK relationships metadata</td>
<td>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.</td>
<td>CBK-to-CBK relationships</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CBK lifecycles</td>
<td>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.</td>
<td>Provenance, Type, Preservation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CBK use outcomes</td>
<td>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.</td>
<td>Evidence from Use</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Relationships between CBK metadata and the FAIR and trustability principles</td>
<td>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.</td>
<td>All</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Common Metadata Framework

Project Title
Common Metadata Framework

Project Description
Individuals from the COVID-19 Knowledge Accelerator (COKA) and Mobilizing Computable Biomedical Knowledge (MCKB) 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 (MCKB) 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.


The COVID-19 Knowledge Accelerator (COKA) Common Metadata Framework Working Group started February 2, 2021 and included some of the MCKB 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 MCKB Spreadsheet":

https://docs.google.com/spreadsheets/d/1ybUISVsOcT3J4J9DR-4LZNcD36ZGfEM/#gid=1835001094

https://fevir.net/resources/Project/29201
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,
- ESGC Pilot - EDMI 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 accomplis."

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

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!

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
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

- Office of the Chief Privacy Officer
- Office of the Chief Operating Officer
- Office of the Chief Scientist
- Office of Standards and Technology
- Office of the National Coordinator
- Office of Quality and Safety
- Office of Policy
- Office of Programs
- Office of Planning, Evaluation, and Analysis

Media Questions

Contact Peter Ashkenazi 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.Ashkenazi@hhs.gov
Telephone: (202) 260-6342
Fax: 202-690-0970

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 6.5" [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.
Unified Medical Language System

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.

1. About the UMLS Resources
   • Metathesaurus: Semantic Network; SPECIALIST Lexicon and lexical programs; MetamorphoSys

2. Accessing UMLS Knowledge Sources
   • Metathesaurus license; Semantic Network; SPECIALIST Lexicon; DVD

3. Knowledge Source Server
   • Download files; searching; additional tools and resources

4. 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.

http://pcornet.org/
11 Clinical Data Research Networks and 18 Patient Powered Research Networks
PCORnet Distributed Research Network (DRN)
Multiple Networks Sharing Infrastructure

Mini-Sentinel

pcornet

NIH Distributed Research Network

Health Plan 1  Health Plan 2  Health Plan 3  Health Plan 4  Health Plan 5  Health Plan 6  Health Plan 7  Health Plan 8  Health Plan 9

Hospital 1  Hospital 2  Hospital 3  Hospital 4  Hospital 5  Hospital 6

Outpatient clinic 1  Outpatient clinic 2  Outpatient clinic 3

Patient network 1  Patient network 2  Patient network 3
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

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.

OMOP = The Observational Medical Outcomes Partnership

Comparisons of OMOP vs PCORnet


• [https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3900207/](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3900207/)
• [https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3824370/](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. Rusincovitch1, Abel N. Kho, MD, MS2, Jon E. Puro, MPA:HA3, Daniella Meeker, PhD4, Pedro Rivera, MSC5, Aaron A. Sorensen, MA5, Jeffrey S. Brown, PhD6, and Lesley H. Curtis, PhD7

1Duke Translational Medicine Institute, Durham, NC; 2Northwestern University Departments of Medicine and Preventive Medicine, Evanston, IL; 3OCHIN, Inc., Portland, OR; 4Department of Health, RAND Corporation, Santa Monica, CA; 5Temple University School of Medicine, Philadelphia, PA; 6Department of Population Medicine, Harvard Medical School and Harvard Pilgrim Health Care Institute, Boston, MA; 7Duke 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 domain, 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 identified questions for PCORnet-specific standards, 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\(^1\):

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

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

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

---

A comparison of phenotype definitions for diabetes mellitus

Rachael L Richesson,1 Shelley A Rusincovitch,2 Douglas Wixted,3 Bryan C Batch,4 Mark N Feinglos,5 Marie Lynn Miranda,3 W Ed Hammond,2,6 Robert M Califf,2,7 Susan E Spratt1

Abstract

Objective This study compared 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=175,503) 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 result) were computed.

Results Three phenotype definitions based heavily on ICD-9-CM codes identified 5–11% 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 ACR Registry, and diabetes-associated medications definitions, which have restricted no ICD-9-CM criteria, identified the smallest proportions of patients (7%). The demographic characteristics for all seven phenotype definitions were similar: 0.6–37% women, mean age range 56–75 years. The NYC ACR Registry definition had higher average patient counts (54) than the other definitions (range 44–80) and the reference population (20) over the 5-year observation period.

Conclusions Further research 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. 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 enhance the development of patient registries from healthcare data, and enable common inclusion criteria to support regional surveillance and the identification of rare disease communities. 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 (90%) 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, pancreatic disorders, and endocrine disorders. All types of diabetes manifest in high blood glucose, and laboratory values are the primary means for diagnosis and management.

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

Data from three domains (International Classification of Diseases, version 9, clinical modification (ICD-9-CM) code 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) 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.

Diabetes-relevant medications are often included in phenotype definitions because medication data...
<table>
<thead>
<tr>
<th>Phenotype definitions:</th>
<th>Data domain criteria</th>
<th>ICD-9-CM 250.x0 and 250.x2 (excludes type 1 specific codes)</th>
<th>Expanded ICD-9-CM Codes (249.xx, 357.2, 362.0x, 366.41)</th>
<th>HbA1c</th>
<th>Fasting glucose</th>
<th>Random glucose</th>
<th>Abnormal OGTT</th>
<th>Diabetes-associated medications*</th>
</tr>
</thead>
<tbody>
<tr>
<td>ICD-9-CM 250.xx</td>
<td></td>
<td><img src="image" alt="Green" /></td>
<td><img src="image" alt="Red" /></td>
<td><img src="image" alt="Green" /></td>
<td><img src="image" alt="Red" /></td>
<td><img src="image" alt="Red" /></td>
<td><img src="image" alt="Red" /></td>
<td><img src="image" alt="Red" /></td>
</tr>
<tr>
<td>CMS CCW</td>
<td><img src="image" alt="Red" /></td>
<td><img src="image" alt="Red" /></td>
<td><img src="image" alt="Red" /></td>
<td><img src="image" alt="Red" /></td>
<td><img src="image" alt="Red" /></td>
<td><img src="image" alt="Red" /></td>
<td><img src="image" alt="Red" /></td>
<td><img src="image" alt="Red" /></td>
</tr>
<tr>
<td>NYC A\text{1c} Registry</td>
<td></td>
<td><img src="image" alt="Red" /></td>
<td><img src="image" alt="Red" /></td>
<td><img src="image" alt="Red" /></td>
<td><img src="image" alt="Red" /></td>
<td><img src="image" alt="Red" /></td>
<td><img src="image" alt="Red" /></td>
<td><img src="image" alt="Red" /></td>
</tr>
<tr>
<td>Diabetes-associated medications</td>
<td></td>
<td><img src="image" alt="Red" /></td>
<td><img src="image" alt="Red" /></td>
<td><img src="image" alt="Red" /></td>
<td><img src="image" alt="Red" /></td>
<td><img src="image" alt="Red" /></td>
<td><img src="image" alt="Red" /></td>
<td><img src="image" alt="Red" /></td>
</tr>
<tr>
<td>DDC</td>
<td><img src="image" alt="Red" /></td>
<td><img src="image" alt="Red" /></td>
<td><img src="image" alt="Red" /></td>
<td><img src="image" alt="Red" /></td>
<td><img src="image" alt="Red" /></td>
<td><img src="image" alt="Red" /></td>
<td><img src="image" alt="Red" /></td>
<td><img src="image" alt="Red" /></td>
</tr>
<tr>
<td>SUPREME-DM</td>
<td><img src="image" alt="Red" /></td>
<td><img src="image" alt="Red" /></td>
<td><img src="image" alt="Red" /></td>
<td><img src="image" alt="Red" /></td>
<td><img src="image" alt="Red" /></td>
<td><img src="image" alt="Red" /></td>
<td><img src="image" alt="Red" /></td>
<td><img src="image" alt="Red" /></td>
</tr>
<tr>
<td>eMERGE†</td>
<td></td>
<td><img src="image" alt="Red" /></td>
<td><img src="image" alt="Red" /></td>
<td><img src="image" alt="Red" /></td>
<td><img src="image" alt="Red" /></td>
<td><img src="image" alt="Red" /></td>
<td><img src="image" alt="Red" /></td>
<td><img src="image" alt="Red" /></td>
</tr>
</tbody>
</table>

*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 (i.e., 250.x1, 250.x3) results in the exclusion of the patient.

- Green = Sole criteria.
- Red = Optional criteria, one of many.
- Red with asterisk = Distinction made between inpatient and outpatient context.
- Red with double slashes = 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
• Ideally, *research and clinical definitions should be semantically equivalent.*
  
i.e., they should identify equivalent populations.
Multicenter Study Comparing Case Definitions Used to Identify Patients with Chronic Obstructive Pulmonary Disease

Valentin Prieto-Centurion¹, Andrew J. Rolle¹, David H. Au², Shannon S. Carson², Ashley G. Henderson³, Todd A. Lee⁴, Peter K. Lindemauer⁵, Mary A. McBurnie⁶, Richard A. Mularski⁷, Edward T. Naureckas⁸, William M. Vollmer⁷, Binoy J. Joese⁹, and Jerry A. Krishnan¹,⁹, on behalf of the CONCERT Consortium

¹Division of Pulmonary, Critical Care, Sleep and Allergy and ²Department of Pharmacy Systems, Outcomes and Policy, University of Illinois at Chicago, Chicago, Illinois; ³University of Washington/VA Puget Sound, Seattle, Washington; ⁴Division of Pulmonary and Critical Care Medicine, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina; ⁵Department of Medicine and Center for Quality of Care Research, Baystate Medical Center, Springfield, Massachusetts; ⁶Tufts University School of Medicine, Boston, Massachusetts; ⁷The Center for Health Research, Kaiser Permanente, Portland, Oregon; ⁸Section of Pulmonary and Critical Care, University of Chicago Medicine, Chicago, Illinois; and ⁹Population 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>
<thead>
<tr>
<th>Characteristic</th>
<th>Total Sample (n = 998)</th>
<th>Clinical Trial Reference Standard</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Yes(\text{a} (n = 560))</td>
<td>No(\text{b} (n = 438))</td>
</tr>
<tr>
<td>Comorbid conditions, %</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cardiovascular disease</td>
<td>76</td>
<td>74</td>
<td>78</td>
</tr>
<tr>
<td>Hypertension</td>
<td>65</td>
<td>63</td>
<td>69</td>
</tr>
<tr>
<td>Heart failure</td>
<td>18</td>
<td>16</td>
<td>22</td>
</tr>
<tr>
<td>Coronary artery disease</td>
<td>23</td>
<td>22</td>
<td>24</td>
</tr>
<tr>
<td>Myocardial infarction</td>
<td>19</td>
<td>18</td>
<td>20</td>
</tr>
<tr>
<td>Stroke</td>
<td>15</td>
<td>14</td>
<td>15</td>
</tr>
<tr>
<td>Depression</td>
<td>42</td>
<td>36</td>
<td>50</td>
</tr>
<tr>
<td>Arthritis</td>
<td>36</td>
<td>33</td>
<td>41</td>
</tr>
<tr>
<td>Diabetes</td>
<td>28</td>
<td>22</td>
<td>34</td>
</tr>
<tr>
<td>Cancer history</td>
<td>23</td>
<td>26</td>
<td>19</td>
</tr>
<tr>
<td>Anemia</td>
<td>28</td>
<td>26</td>
<td>30</td>
</tr>
<tr>
<td>Kidney disease</td>
<td>20</td>
<td>18</td>
<td>21</td>
</tr>
<tr>
<td>Dementia</td>
<td>2</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>Dyspnea at rest (Borg), %</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0, no dyspnea</td>
<td>52</td>
<td>54</td>
<td>50</td>
</tr>
<tr>
<td>0.5–2, slight</td>
<td>38</td>
<td>38</td>
<td>37</td>
</tr>
<tr>
<td>(\geq 3), moderate to very severe</td>
<td>10</td>
<td>7</td>
<td>13</td>
</tr>
<tr>
<td>Spirometry, post-bronchodilator, %</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(\text{FEV}_1)/(\text{FVC}) (\leq 70)%</td>
<td>61</td>
<td>100</td>
<td>11</td>
</tr>
<tr>
<td>(\text{FEV}_1)/(\text{predicted}) (\leq 80)%</td>
<td>72</td>
<td>86</td>
<td>55</td>
</tr>
<tr>
<td>6-minute-walk distance, %</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Distance walked (\leq 350) m</td>
<td>53</td>
<td>52</td>
<td>54</td>
</tr>
</tbody>
</table>

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.

\(\text{a} (A + D + E + G)\) and \(\text{b} (B + C + F)\) in Figure 2.
Table 3. Characteristics Associated with Meeting the Clinical Trial Reference Standard

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>Odds Ratio (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Race (vs. white)</td>
<td></td>
</tr>
<tr>
<td>Black</td>
<td>0.37 (0.26–0.53)*</td>
</tr>
<tr>
<td>Other</td>
<td>0.52 (0.27–1.00)</td>
</tr>
<tr>
<td>Education (vs. high school or less)</td>
<td></td>
</tr>
<tr>
<td>College/professional degree</td>
<td>0.38 (0.26–0.56)*</td>
</tr>
<tr>
<td>Some college</td>
<td>0.68 (1.06–2.03)*</td>
</tr>
<tr>
<td>BMI, kg/m² (vs. normal)</td>
<td></td>
</tr>
<tr>
<td>&lt;18.5 (underweight)</td>
<td>4.00 (1.27–12.50)*</td>
</tr>
<tr>
<td>25–29.99 (overweight)</td>
<td>0.87 (0.58–1.30)</td>
</tr>
<tr>
<td>≥30 (obese)</td>
<td>0.51 (0.35–0.75)*</td>
</tr>
<tr>
<td>Depression (yes vs. no)</td>
<td>0.53 (0.40–0.71)*</td>
</tr>
<tr>
<td>Diabetes (yes vs. no)</td>
<td>0.67 (0.48–0.93)*</td>
</tr>
<tr>
<td>Cancer (yes vs. no)</td>
<td>1.47 (1.05–2.08)*</td>
</tr>
</tbody>
</table>

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
<table>
<thead>
<tr>
<th>Name</th>
<th>Type</th>
<th>Code System</th>
<th>Steward</th>
<th>OID</th>
</tr>
</thead>
<tbody>
<tr>
<td>Complications of Pregnancy, Childbirth and the Fetus</td>
<td>SNOMEDCT</td>
<td>NCQA</td>
<td>2.16.840.1.113883.1003.111.11.1023</td>
<td></td>
</tr>
<tr>
<td>Complications of Pregnancy, Childbirth and the Fetus</td>
<td>SNOMEDCT</td>
<td>NCQA</td>
<td>2.16.840.1.113883.1003.111.11.1022</td>
<td></td>
</tr>
<tr>
<td>Conditions Possibly Justifying Elective Delivery Pt Grouping</td>
<td>ICD-10-CM</td>
<td>The Joint Commission</td>
<td>2.16.840.1.113883.117.1.1.1.2.280</td>
<td></td>
</tr>
<tr>
<td>Conditions Possibly Justifying Elective Delivery Pt Grouping</td>
<td>ICD-10-CM</td>
<td>The Joint Commission</td>
<td>2.16.840.1.113883.117.1.1.1.2.294</td>
<td></td>
</tr>
<tr>
<td>Conditions Possibly Justifying Elective Delivery Pt Grouping</td>
<td>ICD-10-CM</td>
<td>The Joint Commission</td>
<td>2.16.840.1.113883.117.1.1.1.2.292</td>
<td></td>
</tr>
<tr>
<td>Conditions Possibly Justifying Elective Delivery Pt Grouping</td>
<td>ICD-10-CM</td>
<td>The Joint Commission</td>
<td>2.16.840.1.113883.117.1.1.1.2.295</td>
<td></td>
</tr>
<tr>
<td>Degeneration of Macula and Posterior Pole</td>
<td>Grouping</td>
<td>ICD-10-CM</td>
<td>AMA-PCP</td>
<td>2.16.840.1.113883.3.529.1.1453</td>
</tr>
<tr>
<td>Degeneration of Macula and Posterior Pole</td>
<td>Extensional</td>
<td>SNOMEDCT</td>
<td>AMA-PCP</td>
<td>2.16.840.1.113883.3.520.2.1643</td>
</tr>
<tr>
<td>Delivery</td>
<td>Extensional</td>
<td>ICD-10-CM</td>
<td>Optum</td>
<td>2.16.840.1.113883.3.520.2.1643</td>
</tr>
<tr>
<td>Delivery · Diagnoses</td>
<td>Grouping</td>
<td>ICD-10-CM</td>
<td>Optum</td>
<td>2.16.840.1.113883.3.520.2.1643</td>
</tr>
<tr>
<td>Delivery · Diagnoses</td>
<td>Extensional</td>
<td>ICD-10-CM</td>
<td>Optum</td>
<td>2.16.840.1.113883.3.520.2.1643</td>
</tr>
<tr>
<td>Diabetes</td>
<td>Extensional</td>
<td>ICD-10-CM</td>
<td>NCQA</td>
<td>2.16.840.1.113883.3.464.1003.103.111.1002</td>
</tr>
<tr>
<td>Diabetes</td>
<td>Grouping</td>
<td>ICD-10-CM</td>
<td>NCQA</td>
<td>2.16.840.1.113883.3.464.1003.103.111.1001</td>
</tr>
<tr>
<td>Diabetes</td>
<td>Extensional</td>
<td>SNOMEDCT</td>
<td>NCQA</td>
<td>2.16.840.1.113883.3.464.1003.103.111.1001</td>
</tr>
<tr>
<td>Diabetes</td>
<td>Extensional</td>
<td>SNOMEDCT</td>
<td>CPT</td>
<td>2.16.840.1.113883.3.464.1003.103.111.1001</td>
</tr>
</tbody>
</table>
What is the Phenotype KnowledgeBase?

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 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

https://phekb.org/
Activity – explore PheKB
Desiderata for the development of next-generation electronic health record phenotype libraries

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

1Department of Population Health Sciences, King's College London, London, SE1 1UL, UK; 2Health Informatics Centre (HIC), University of Dundee, Dundee, DD1 9SY, UK; 3Feinberg School of Medicine, Northwestern University, Chicago, IL 60611, USA; 4Institute of Cancer and Genomic Sciences, University of Birmingham, Birmingham, B15 2TT, UK; 5SAIL Databank, Swansea University, Swansea, SA2 8PP, UK; 6European Molecular Biology Laboratory, European Bioinformatics Institute, Hinxton, CB10 1SD, UK; 7Department of Learning Health Sciences, University of Michigan Medical School, MI 48109, USA and 8Institute 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 http://orcid.org/0000-0002-5242-9701

### Table 1:

Phenotype definition formats

<table>
<thead>
<tr>
<th>Format</th>
<th>Description</th>
<th>Example</th>
<th>Category</th>
</tr>
</thead>
<tbody>
<tr>
<td>Code list</td>
<td>A set of codes that must exist in a patient’s health record in order to include them within a phenotype cohort</td>
<td>COVID-19 ICD-10 code “U07.1”</td>
<td>Rule-based</td>
</tr>
<tr>
<td>Simple data elements</td>
<td>Formalizing the relationship between code-based data elements using logical connectives</td>
<td>COVID-19 ICD-10 code “U07.1” AND ICD-11 code “RA01.0”</td>
<td>Rule-based</td>
</tr>
<tr>
<td>Complex data elements</td>
<td>Formalizing the relationship between complex data elements, such as those derived via NLP</td>
<td>Patient’s blood pressure reading &gt;140 OR patient notes contain “high BP”</td>
<td>Rule-based</td>
</tr>
<tr>
<td>Temporal</td>
<td>Prefix rules with temporal qualifiers</td>
<td>Albumin levels increased by 25% over 6 hours, high blood pressure reading has to occur during hospitalization</td>
<td>Rule-based</td>
</tr>
<tr>
<td>Trained classifier</td>
<td>Use rule-based definitions as the basis for constructing a classifier for future (or additional) cohorts</td>
<td>A k-fold cross-validated classifier capable of identifying patients with COVID-19</td>
<td>Probabilistic</td>
</tr>
<tr>
<td>Mechanism</td>
<td>Description</td>
<td>Example</td>
<td></td>
</tr>
<tr>
<td>------------------------</td>
<td>-----------------------------------------------------------------------------</td>
<td>-------------------------------------------------------------------------</td>
<td></td>
</tr>
<tr>
<td>Disease registries</td>
<td>Compare the phenotype cohort with those present in the registry</td>
<td>Comparison of a diabetes phenotype cohort with those patients present in a diabetes registry (e.g., T1D exchange)</td>
<td></td>
</tr>
<tr>
<td>Chart review</td>
<td>Compare the phenotype cohort with the patients identified by manual review of medical records</td>
<td>Comparison with a diabetes gold standard, produced by double manual review of patient medical records</td>
<td></td>
</tr>
<tr>
<td>Cross-EHR concordance</td>
<td>Compare percentage of cases identified by a phenotype across different sources, and identify any overlap</td>
<td>Comparison of the percentage of patients identified by a diabetes phenotype in primary and secondary care EHRs, and the identification of any case overlap</td>
<td></td>
</tr>
<tr>
<td>Risk factors</td>
<td>Compare the magnitude of the phenotype cohort with standard risk calculations</td>
<td>Comparison with the output of a Cox hazards model</td>
<td></td>
</tr>
<tr>
<td>Prognosis</td>
<td>Compare the magnitude of the phenotype cohort with external prognosis models</td>
<td>Comparison with a survival analysis</td>
<td></td>
</tr>
<tr>
<td>Genetic associations</td>
<td>Compare whether the presence of a patient in a phenotype cohort is consistent with their genetic profile</td>
<td>A patient is more likely to be a valid member of a diabetes cohort if they have the HLA-DR3 gene</td>
<td></td>
</tr>
</tbody>
</table>
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 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](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

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.

#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 increasing rate of scientific discovery demands 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 unlocks the potential of information technology to generate and deliver useful information—and, in particular, 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.

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

Achieving this through evolution of an open Computable Biomedical Knowledge ecosystem designed to achieving the FAIR principles: making Computable Biomedical Knowledge findable, accessible, interoperable, and 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 science of Computable Biomedical Knowledge collaboratively.
  - Be able 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 Ecosystem 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.
  - Enhance equity in health and knowledge accessibility.

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
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