Defining/developing CDEs for use in AI/ML Applications for Clinical Research with Electronic Health Record Data

Advancing the Use and Development of Common Data Elements in Research Workshop

NIH | March 7, 2023

Katherine P. Liao, MD, MPH
Associate Professor of Medicine and Biomedical Informatics, Harvard Medical School
Division of Rheumatology, Inflammation, and Immunity, Brigham and Women's Hospital
VA Boston Healthcare System

BWH VA U.S. Department of Veterans Affairs

U.S. Department
of Veterans Affairs

Outline

- Example project w/ artificial intelligence (AI)/machine learning (ML) for clinical research
    - Treatment response in rheumatoid arthritis (RA)
- Application of common data elements (CDE)
    - Challenges
    - Potential solutions
    - Future directions

Rheumatoid arthritis (RA)

- Most common autoimmune inflammatory joint disease
- Numerous potent immunomodulatory treatments available
- Treatment in 2024 remains a trial-and-error approach
  - Joint destruction & disability
  - Side effects from ineffective therapies

BWH U.S. Department of Veterans Affairs

RA

Source: ACP Medicine © 2004 WebMD Inc.

AI/ML to study treatment response in RA

- RA randomized controlled trials (RCTs)
  - Gold standard for treatment effectiveness
- RCTs powered to answer specific set of clinical question(s)
  - Not powered for subgroup analyses
- Observational data w/ large population → potential to study RCT subgroup findings
  - Limited by confounding

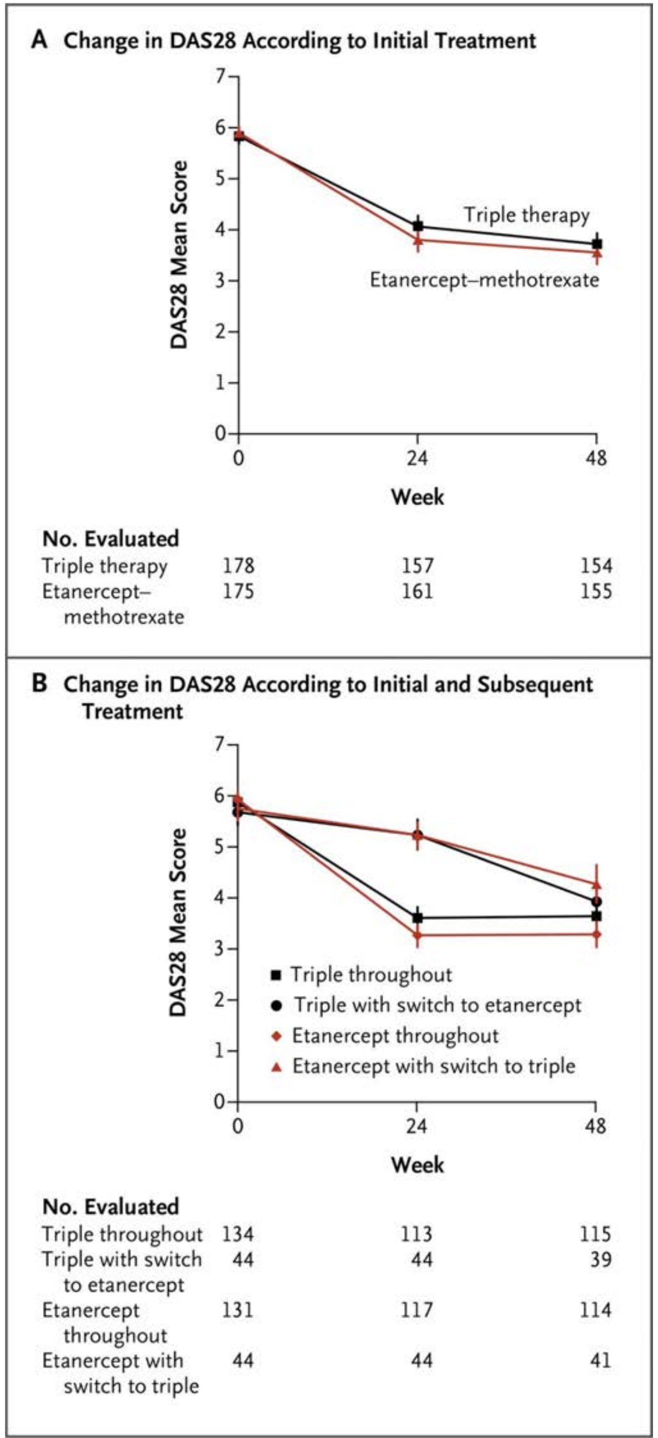
BWH VERITAS VA U.S. Department of Veterans Affairs

Therapies for active RA after methotrexate failure trial (RACAT)

- Findings: Triple therapy non-inferior to tumor necrosis factor (TNFi)

- RA treatment response defined by disease activity, e.g., DAS28

- Subgroup analyses
    - Same % switched at 24 weeks
    - Significant improvement in both groups
    - *Suggests some pts may have benefitted from one or the other therapy at outset*

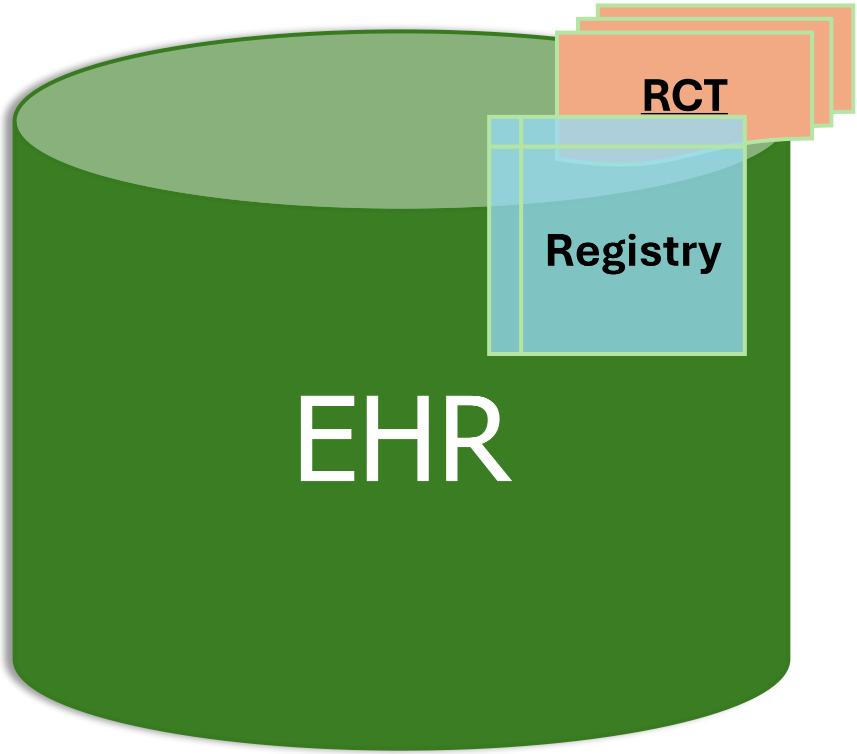


O'Dell et al., NEJM 2013

Data sources, data types, and CDEs

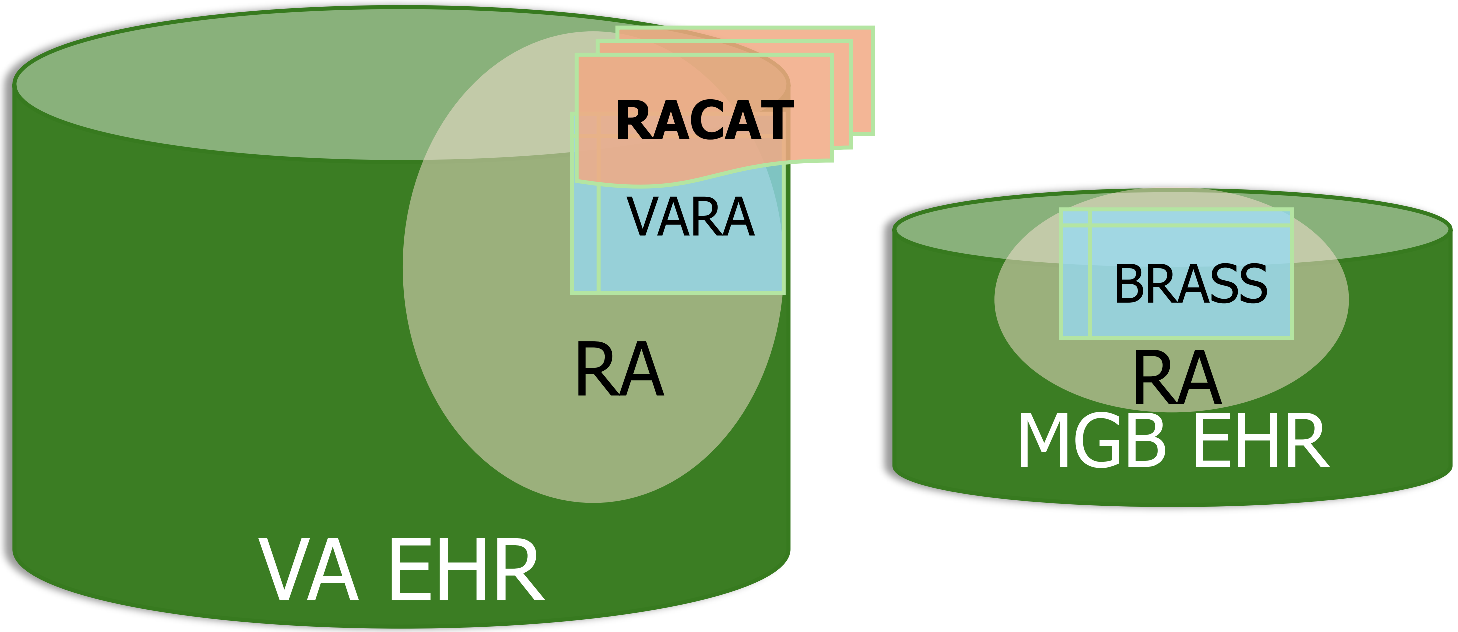
- Test causal inference & machine learning (ML) to leverage observational data to study drug effectiveness using RCTs as the gold standard

- Electronic health records (EHR)
  - Veteran Affairs (VA), ~22 million
  - Mass General Brigham (MGB), ~14 million
- RCT
  - RACAT @ VA
- Registry
  - VARA
  - BRASS



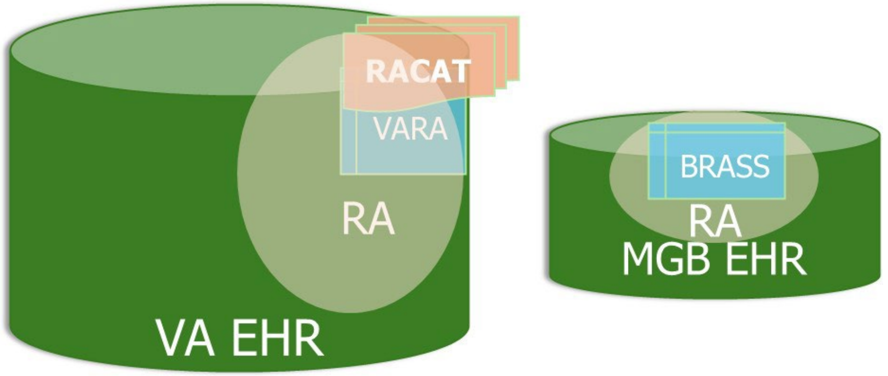
U.S. Department of Veterans Affairs

Data sources, data types, and CDEs



BRASS= Brigham Rheumatoid Arthritis Sequential Study
EHR= electronic health record
MGB= Mass General Brigham
VARA= Veteran Affairs Rheumatoid Arthritis registry
VA= Veteran Affairs

CDE related challenges



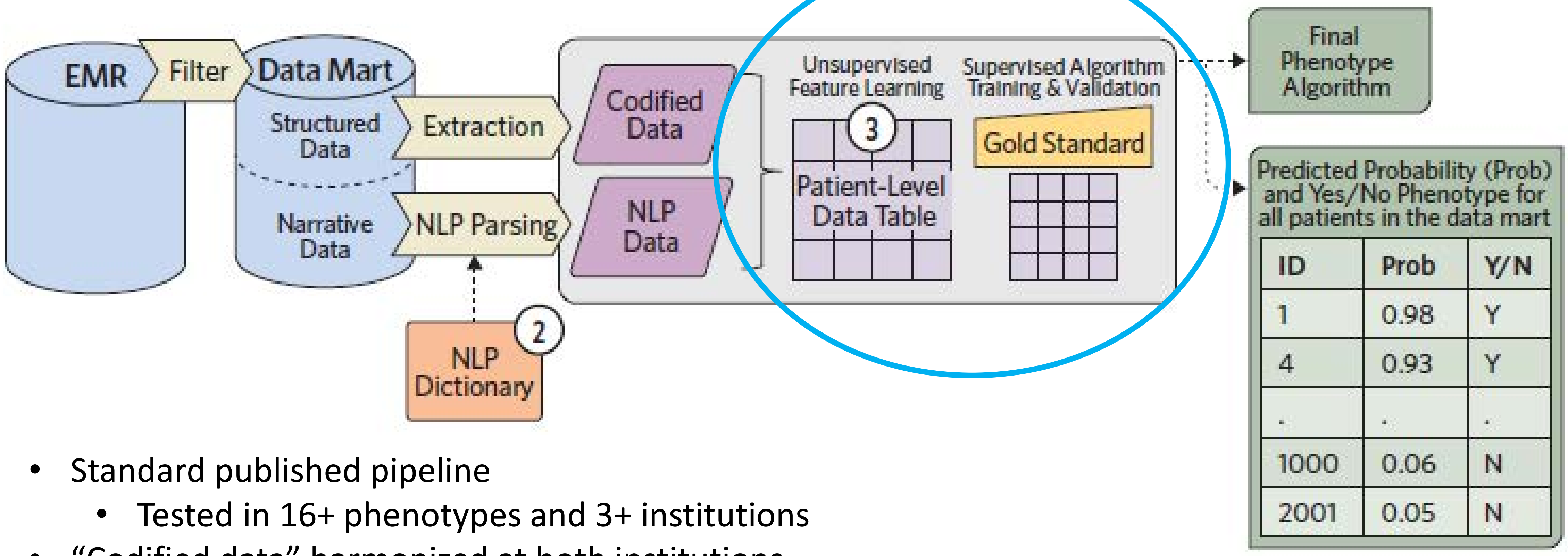
- Defining RA in two EHRs
  - Positive predictive value of ≥1 RA code 20%
  - Published ML pipeline for phenotyping
  - Mapping equivalent features in two different EHR systems
    - Medications, e.g., TNFi
- Imputing RA disease activity, no code or standard with EHR data
  - Train/co-train, port, validate
  - Harmonizing features
    - "RA" code
    - Map health system specific EHR codes to standardized
    - NLP for disease activity
      - RA-specific, e.g., synovitis

Liao et al., Arth Care & Res 2010;
Carroll et al., JAMIA 2012

BWH VA U.S. Department of Veterans Affairs

Defining RA in 2 EHRs

Semi-supervised phenotyping



- Standard published pipeline
  - Tested in 16+ phenotypes and 3+ institutions
- “Codified data” harmonized at both institutions
- NLP data mapped to Unified Medical Language System (UMLS) concepts, NIH NLM

Zhang, Cai, Yu, et al., Nat Protocol 2019

U.S. Department of Veterans Affairs

Common CDE building blocks

- "Proto-CDEs" to find equivalent features across EHRs
  - Data extracted w/ NLP → UMLS, concept unique identifiers (CUIs)
  - Structured ICD EHR data → PheWAS Code (Phecode)
    - Starting definition for "RA code"
    - Modification use Phecode as reference
  - Medications
    - RxNorm, UMLS
    - National Drug Code (NDC), US FDA
- Challenge: who was prescribed a TNFi?
  - RxNorm + National Drug Code (NDC) for all 5 TNFi's
    - Roll-up both codes to create a "TNFi" category
    - Subcategory: generic + trade name by biologic

Denny et al., Nat Biotech 2013

Algorithm to impute RA disease activity

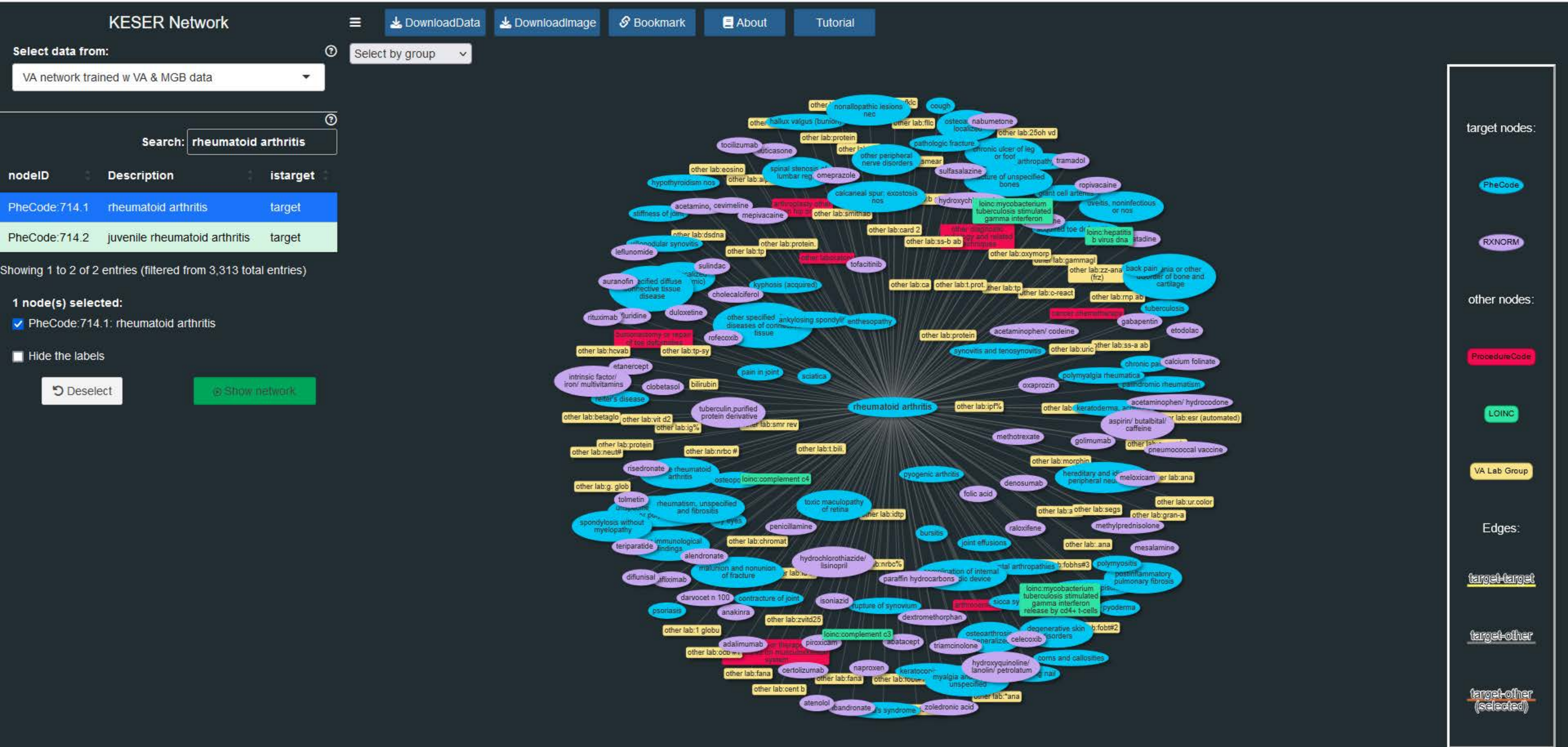
- Challenge: Identify relevant features for disease activity
  - Features available in VA and MGB EHRs
  - No codes for RA disease activity

- Create a knowledge network/graph
  - Identify entities, e.g., "phenotypes" defined by groups of ICD codes (Phecodes)
  - Quantify relationship of entities to each other, i.e., embedding vectors
  - Apply large language models (LLMs)
  - Co-trained with EHR data from MGB & VA
    - Performed in collaboration with Dept of Energy

Hong et al., NPG Digit Med 2021;
Mikolov, et al. arxiv 2013,

U.S. Department
of Veterans Affairs

Potential solution: LLM driven knowledge networks/graphs



Potential solution: LLM driven knowledge networks/graphs

KESER Network

DownloadData | DownloadImage | Bookmark | About | Tutorial

Select data from:

VA network trained w VA & MGB data

Select by group

Search: rheumatoid arthritis

nodeID	Description	istarget
PheCode:714.1	rheumatoid arthritis	target
PheCode:714.2	juvenile rheumatoid arthritis	target

Showing 1 to 2 of 2 entries (filtered from 3,313 total entries)

1 node(s) selected:

PheCode:714.1: rheumatoid arthritis

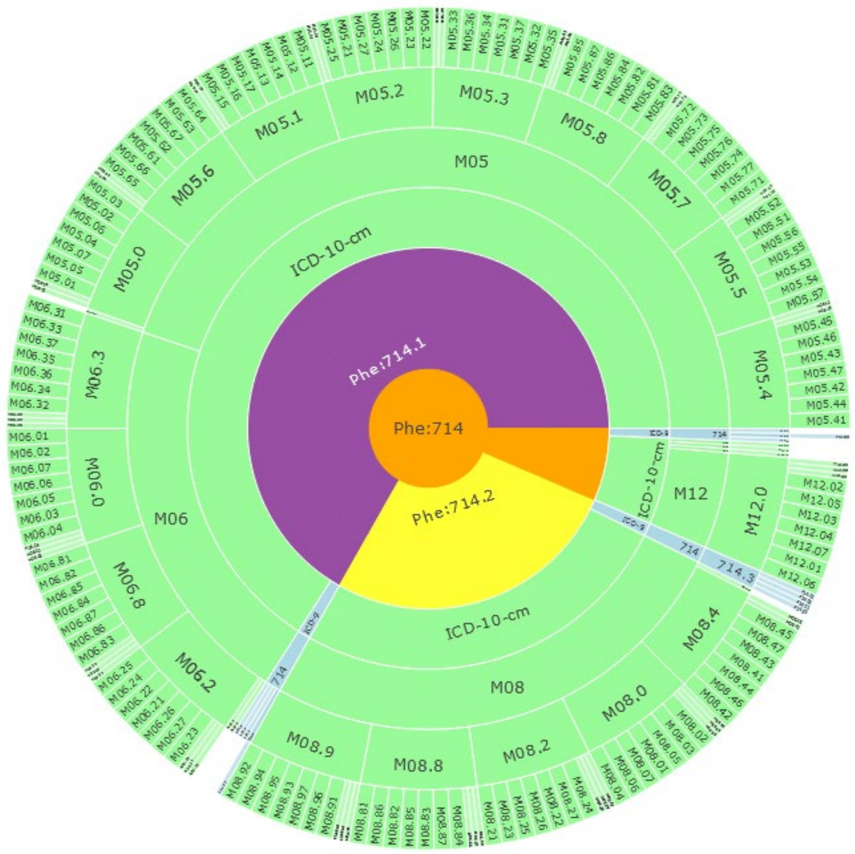
Hide the labels

Deselect

Show network

Application: **universe of codified features, “computed CDEs” for disease activity algorithm**

VA Centralized Interactive Phenomics Resource (CIPHER)



CIPHER
CENTRALIZED INTERACTIVE PHENOMICS RESOURCE

- Standardize meta-data
- Visualization of data
- Real-time comparison of “computable” phenotype definitions by components
- Tools to facilitate use of computable phenotypes
  - ICD hierarchy tool

Kirby et al., JAMIA 2016;
Honerlaw et al., JAMIA 2024 accepted

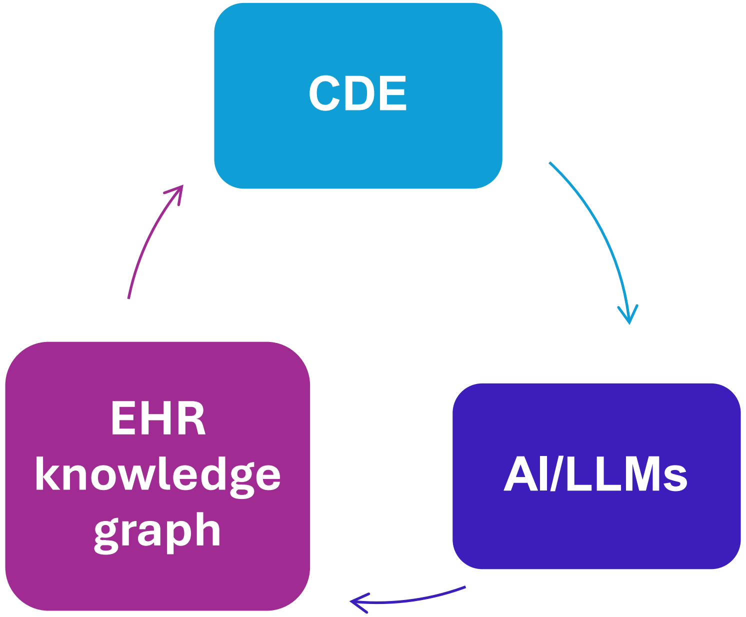
BWH VA U.S. Department of Veterans Affairs

A clinical investigator's CDE wish list

- Resource for knowledge sources standardizing health care data
  - NIH, ONC, FDA & gov't agencies
- Consensus for core set(s) of CDEs for clinical research, "code book"
  - ICD→Phecodes
  - Groupings for LOINC, RxNorm + NDC
- Interactive, moderated platform for sharing CDEs
  - CDE easily findable, comparable to another study
  - Provenance, initial use case(s)

Future directions & thoughts

- Diversity of CDEs for clinical research
  - Meta-data for CDEs
  - Clinical trial vs large-scale EHR study
    - Flexibility on mandating CDE
    - Applicability varies



- CDE evolve w/ technologies
  - Input to ML algorithms manual ICD lists → ICD list from knowledge network ("computed CDEs")

- CDE foundation to develop and benchmark new AI methods

Thank you

TRANSLATIONAL DATA SCIENCE
CELEHS
CENTER FOR A LEARNING HEALTH SYSTEM

VERITY BIOINFORMATICS CORE TEAM		
Brigham & Women's Hospital	**Harvard Medical School**	**Mass General Brigham Research Computing**
Greg McDermott Mary Jeffway Tianrun Cai Feng Liu Yumeko Kawano Jackie Stratton Dana Weisenfeld	Tianxi Cai Vidul Panickan Clara Lea-Bonzel Mohammed Moro Sara Morini Xin Xiong Florence Bourgeois	Andrew Cagan
VA Team & CIPHER J Michael Gaziano Kelly Cho Jacqueline Honerlaw Anne Ho Lauren Costa	Alicia Chen Rahul Sangar Connor Melley Vidisha Tanukonda Monika Maripuri Ashley Galloway Dan Posner	Michael Murray Paul Monach Suma Muralidhar & VA Central Office Lead CIPHER Team members CIPHER Online Oakridge National Research Lab

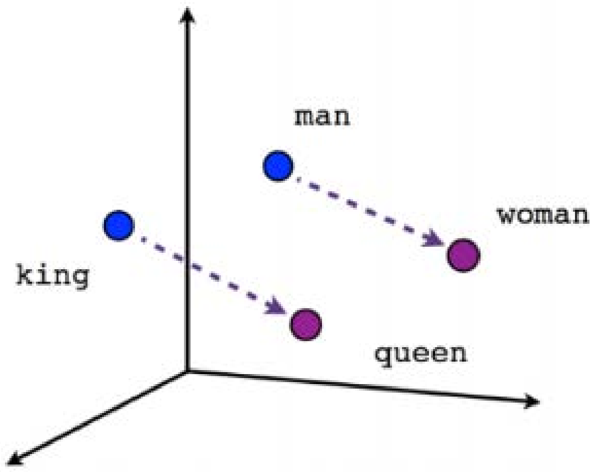
NIH
National Institutes of Health

NIAMS
P30 AR072577
R21 AR078339
R01 AR080193

FDA
U01 FD007929
mPI: Bourgeois, Cai

Creating an EHR clinical knowledge graph

using methods from language models



- Create a co-occurrence matrix
  - Relationship of all structured data to each other
    - ICD, electronic prescriptions, lab codes
  - 17 million Veterans
    - Collaboration with Dept of Energy and use of supercomputers
- Transform concept relationships to numbers
  - Create embedding vectors based on information from relationships
  - Vectors encode the "meaning" of the codes
- Quantify relationship of concepts to each with embedding vectors

Hong et al., NPG Digit Med 2021;
Mikolov, et al. arxiv 2013,
https://arxiv.org/pdf/1310.4546

BWH VA | U.S. Department of Veterans Affairs