Introduction to Bioinformatics for Clinical Research

VERITY Clinical Research Course | April 4, 2024

Katherine P. Liao, MD, MPH

Associate Professor of Medicine & Biomedical Informatics, Harvard Medical School
Division of Rheumatology, Inflammation, and Immunity, Brigham and Women's Hospital
Co-Director, Applied Bioinformatics Core for Clinical Research, VA Boston Healthcare System

BWH VA | U.S. Department of Veterans Affairs

U.S. Department
of Veterans Affairs

Learning objectives

- Describe the importance of defining phenotypes in clinical studies using electronic health record (EHR) data

- Recognize applications of using natural language processing in clinical EHR studies

- Describe language models and potential applications for EHR-based clinical research

Bioinformatics for Clinical Data

2009 HITECH Act

Paper charts

Manual chart review to extract data

- Limits variables and outcomes for study
- Not feasible to study large populations

Electronic health records (EHR)

High volume data from clinical care

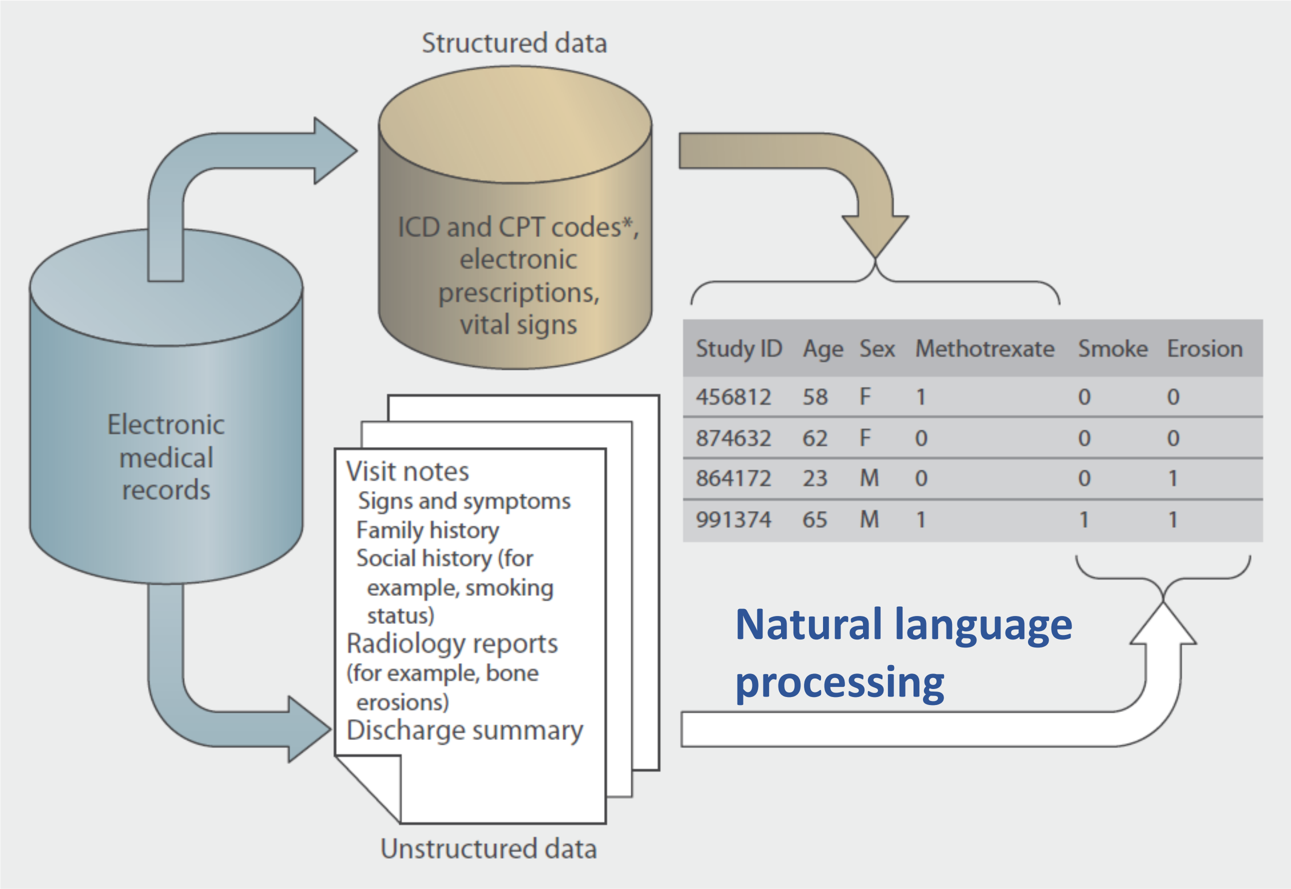
- Designed for billing, patient care
  - Suboptimal for research
- Enables studies in large populations

Project 1: Phenotyping using EHR data

Source: ACP Medicine © 2004 WebMD Inc.

BWH VA U.S. Department of Veterans Affairs

Types of EHR data



Study ID	Age	Sex	Methotrexate	Smoke	Erosion
456812	58	F	1	0	0
874632	62	F	0	0	0
864172	23	M	0	0	1
991374	65	M	1	1	1

Liao, Cai, et al., BMJ 2015

Methods for phenotyping w/ EHR data

Limitations in ICD codes for phenotyping

- Rule-based, Boolean
  - 1 RA ICD code + 1 DMARD electronic prescription, PPV 45%
- Logistic regression
  - Features/variables in a weighted
  - Probability of a condition
- Artificial Intelligence
  - Machine learning (ML)
  - Large Language Models (LLMs), *more on this later*

BWH VA | U.S. Department of Veterans Affairs

Project 1, Step 1: Who has RA in the EHR?

ID	Age	Sex	Dx code	Lab	Dis+
1563	22	M	0	-	0
2821	45	F	1	31	1
9402	75	F	1	40	1
7469	67	M	0	-	0
9468	56	M	0	56	1
5768	54	F	0	11	0
3958	81	F	1	42	1
2463	48	F	0	5	0
8465	72	F	1	6	0
3237	65	F	1	-	0

- Select structured data a priori
  - ICD
  - Medications
  - Labs
- Ceiling for performance
  - PPV ~65%

Pattern recognition

- Adding "features"
  - Improve accuracy
  - Add noise
- Difficult for "humans" to see the pattern

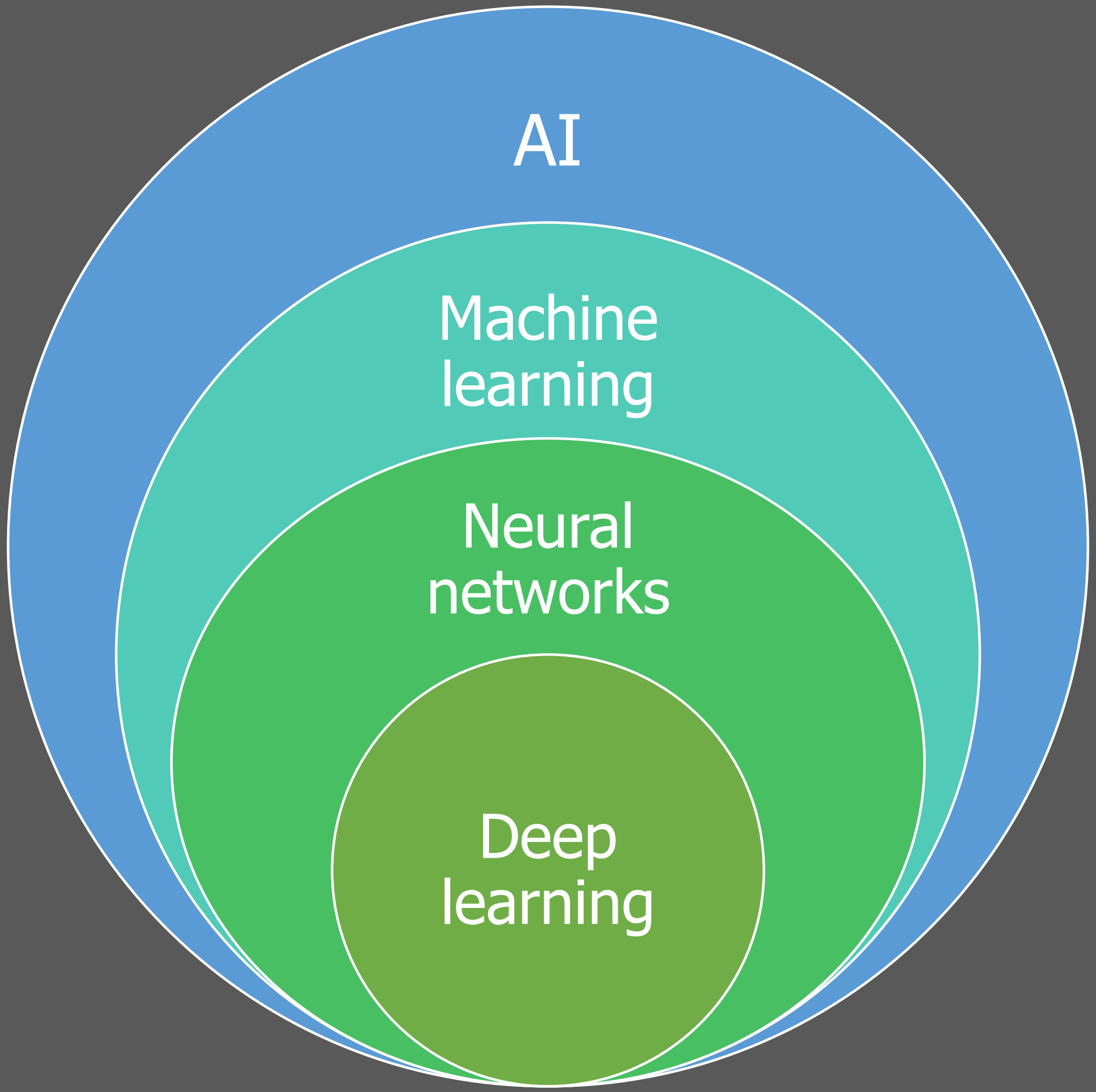
+200 subjects

ID	Age	Sex	Dx code	Lab	Dis+
1563	22	M	0	-	0
2821	45	F	1	31	1
9402	75	F	1	40	1
7469	67	M	0	-	0
9468	56	M	0	56	1
5768	54	F	0	11	0
3958	81	F	1	42	1
2463	48	F	0	5	0
8465	72	F	1	6	0
3237	65	F	1	-	0

Training set

+1000 variables

BWH VA | U.S. Department of Veterans Affairs


AI
Machine learning
Neural networks
Deep learning

Artificial Intelligence & Machine Learning

- Artificial intelligence (AI)
  - Intelligence demonstrated by machines
  - Contrast to human/natural intelligence
- Machine learning (ML) → subset of AI
  - Requires training set
  - Prediction (vs causality)
    - Does not address why or how to change outcomes
  - Learn structure from data → pattern recognition

3 broad concepts

- Feature extraction
  - Important variables
- Regularization
  - Weighting, weed out uninformative featur
- Cross-validation
  - Avoid overfitting

- Supervised vs unsupervised learning
  - Degree of input from domain experts

ID	Age	Sex	Dx code	Lab	Dis+
1563	22	M	0	-	0
2821	45	F	1	31	1
9402	75	F	1	40	1
7469	67	M	0	-	0
9468	56	M	0	56	1
5768	54	F	0	11	0
3958	81	F	1	42	1
2463	48	F	0	5	0
8465	72	F	1	6	0
3237	65	F	1	-	0

Approach to developing phenotype algorithms using EHR data

- Chart review
  - Not feasible in most cases

- Rule-based
  - Relies on human expertise to identify important features
  - Algorithm is a combination of AND, NOT, OR

- ML
  - Data driven method to select features and develop algorithm

Agarwal, et al., JAMIA 2016; Banda, et al., AMIA Jt Summ Proc 2017

Natural language processing (NLP)

Computational method for text processing based on the rules of linguistics

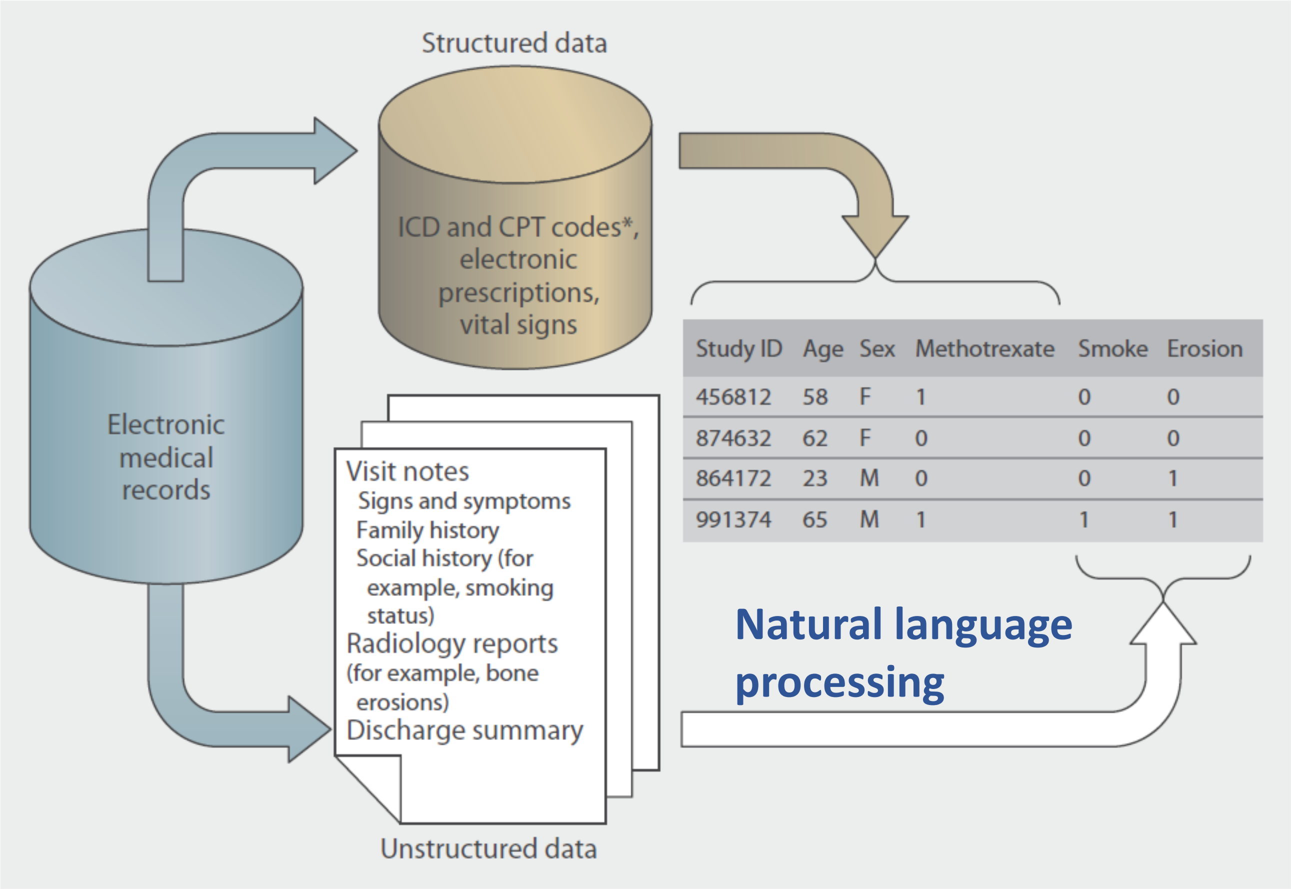
NLP

I	saw	the	girl	with	the	ophthalmoscope.
w1	w2	w3	w4	w5	w6	w7
pronoun	verb	article	noun	prep	article	noun

NLP ≠ "find" command in Word

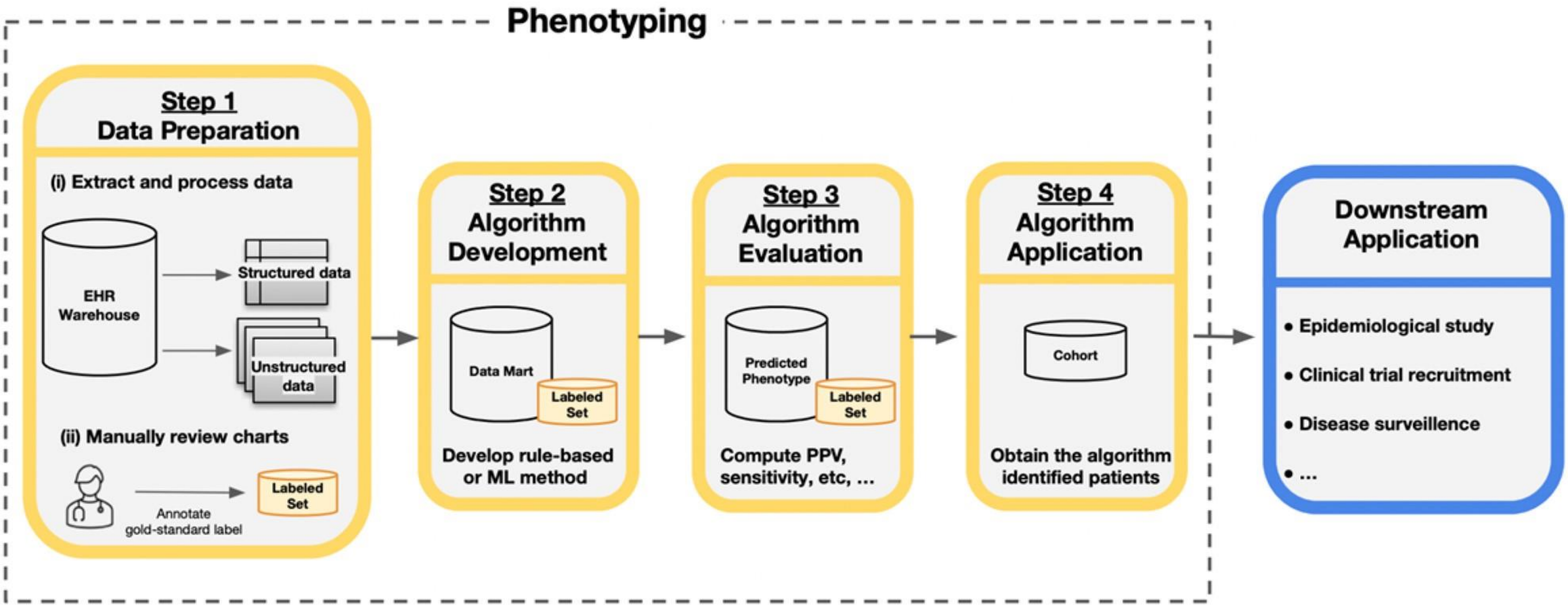
- Negation
  - The patient has no erosions in the MCPs.
- Inverted syntax
  - Colon, ascending and descending, biopsy
- Relation
  - Tamoxifen is used in the treatment of breast cancer
- Morphologic variations
  - Tobacco, 30 pack years, past smoker, +tob → smoking

Types of EHR data



Liao, Cai, et al., BMJ 2015

Overview of phenotyping with EHR data



Yang et al., JAMIA 2023

U.S. Department of Veterans Affairs

Phenotype algorithm: Weighted features in logistic regression model

$$\begin{aligned}\text{Logit (probability of PA)} &= \text{intercept} - 0.16(\text{sex})\\ &+ 0.73\log(1 + (\text{NLP PA})) + 0.88\log(1 + (\text{ICD-9 PA}))\\ &+ 0.63(\text{NLP treatment}) + \ldots\end{aligned}$$

BWH VA | U.S. Department of Veterans Affairs

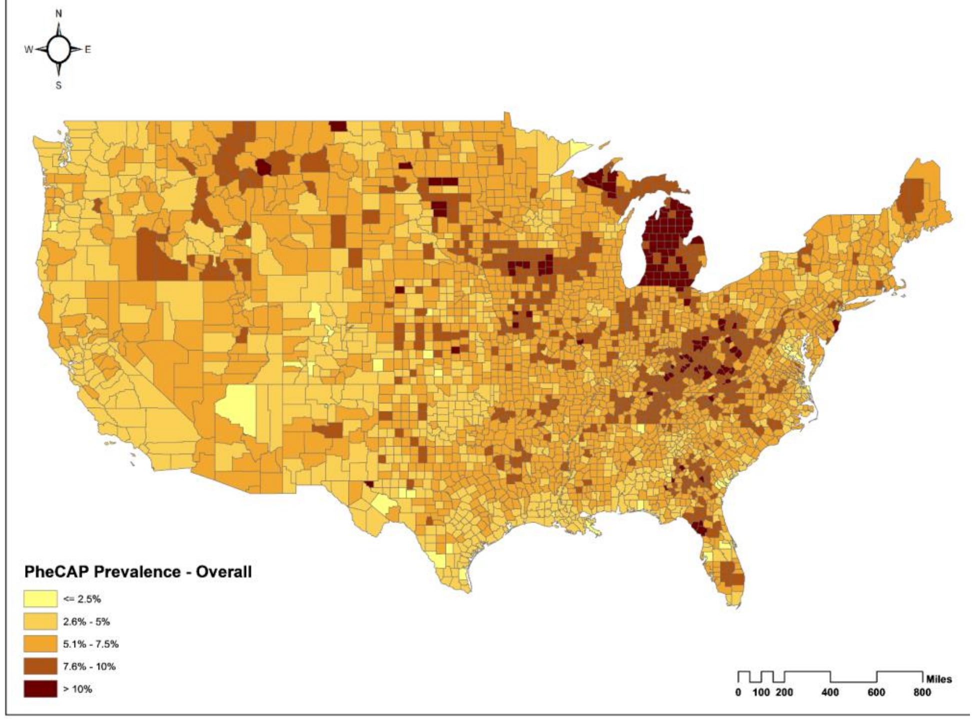
Artificial Intelligence & Machine Learning

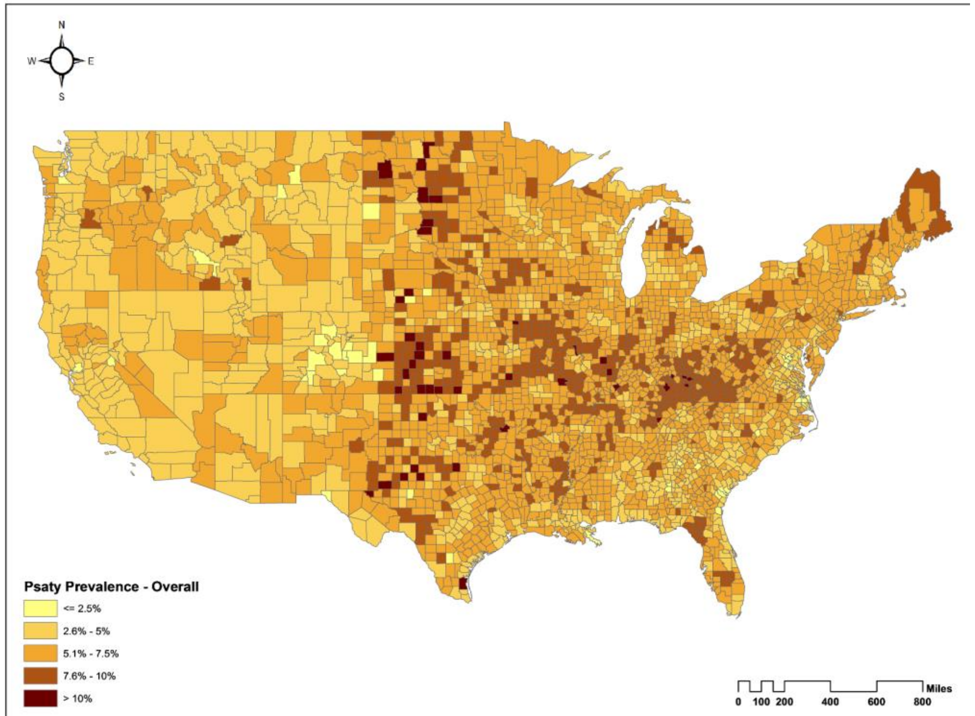
- Artificial intelligence (AI)
  - Intelligence demonstrated by machines
  - Contrast to human/natural intelligence

No "sense" to know if something is off
Example: patient with 150 ICD codes for PsA but no NLP mentions

- Machine learning (ML) → subset of AI
  - Requires training set
  - Prediction (vs causality)
    - Does not address why or how to change outcomes
  - Learn structure from data → pattern recognition

Biased training set → biased results
Example: Training set is all female but algorithm applied to 50:50 female to male pop'n





Real-life blooper

Phenotype project on myocardial infarction (MI) @ VA

Trained in one institution and validated in 3 other institutions

Top panel: ICD + NLP in algorithm

Bottom panel: ICD only

Something amiss?

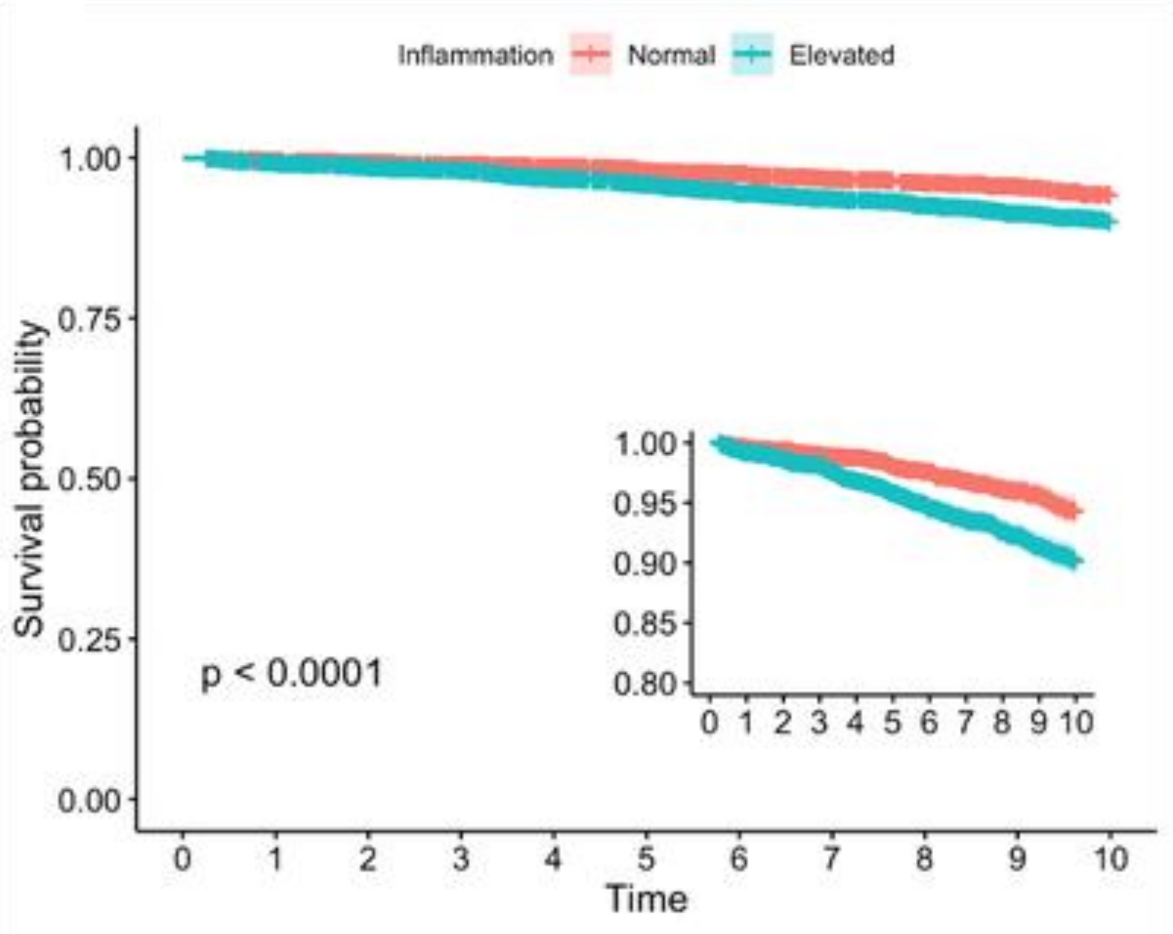
BWH VERITAS VA U.S. Department of Veterans Affairs

Project 2: EHR based cohort study

Project 2: Association between inflammation with HF subtypes in RA

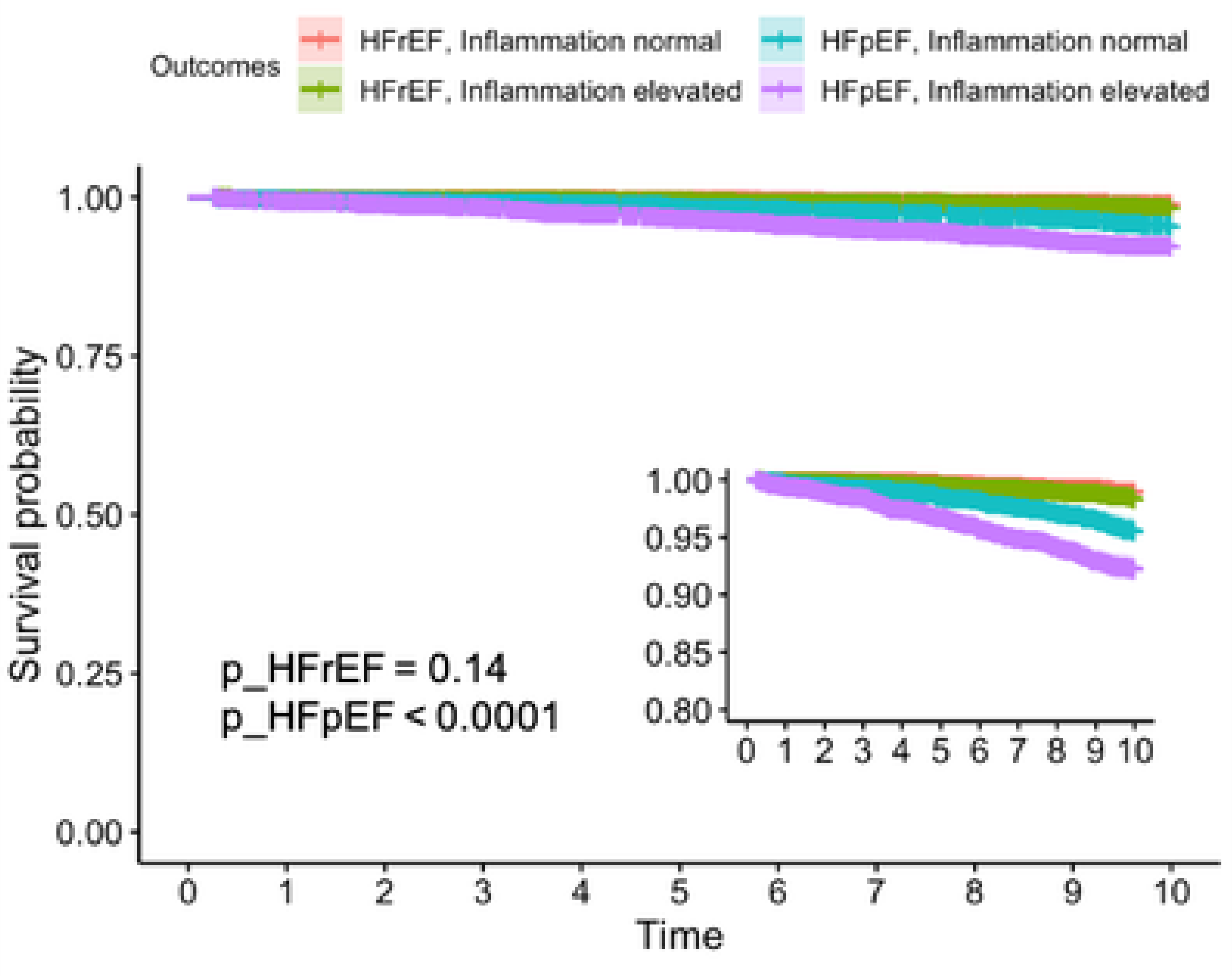
- RA patients at 1.5-2x excess risk for CVD including heart failure (HF)
- Hypothesis
  - Elevated inflammation associated with HF
    - Inflammation stronger association on HF with preserved ejection fraction (HFpEF) vs HF with reduced EF (HFrEF)
- Inflammation modifiable
- First steps for understanding prevention or treatment of HFpEF

Association between inflammation and incident HF

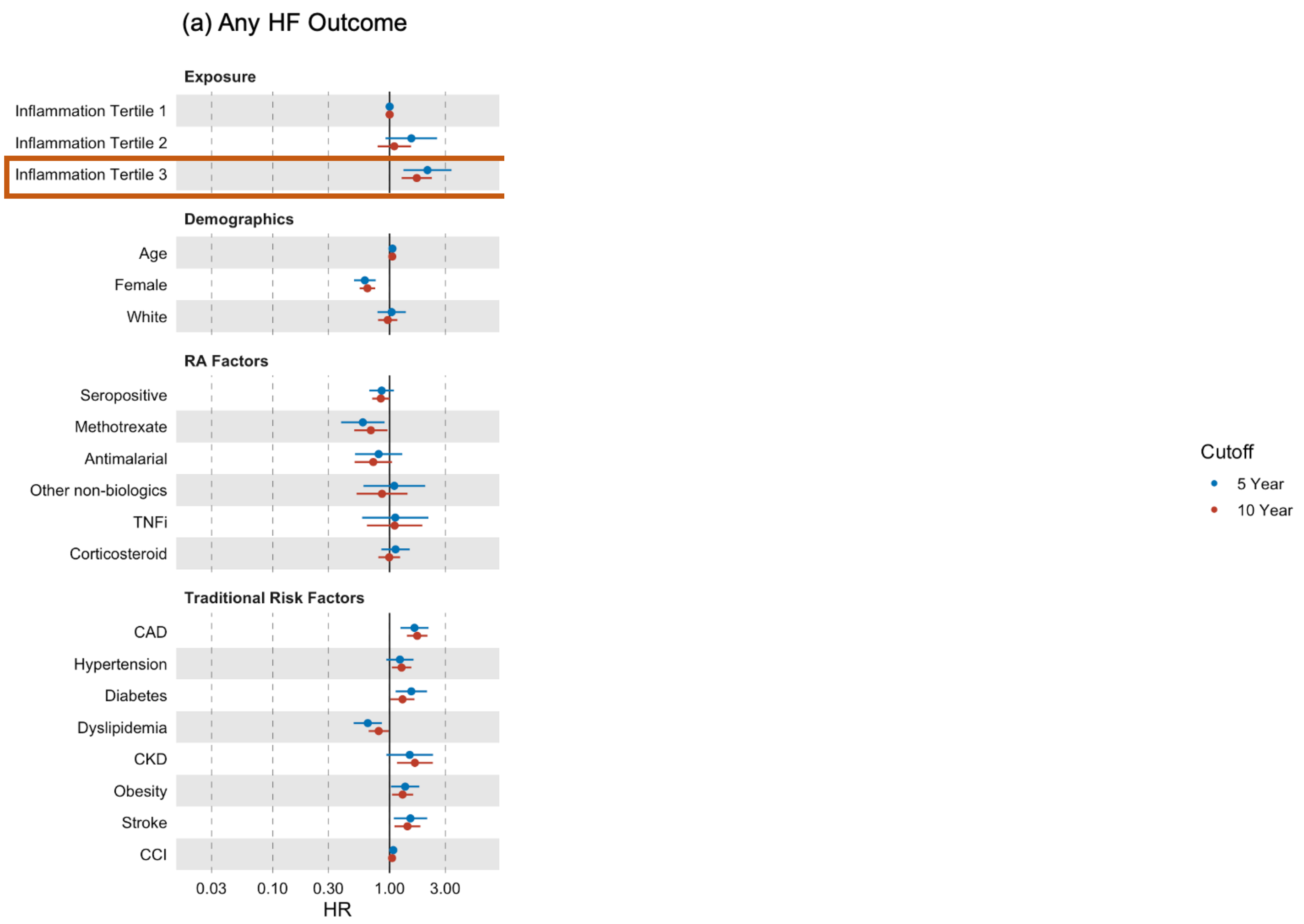


Huang et al., Arth Care & Res 2021

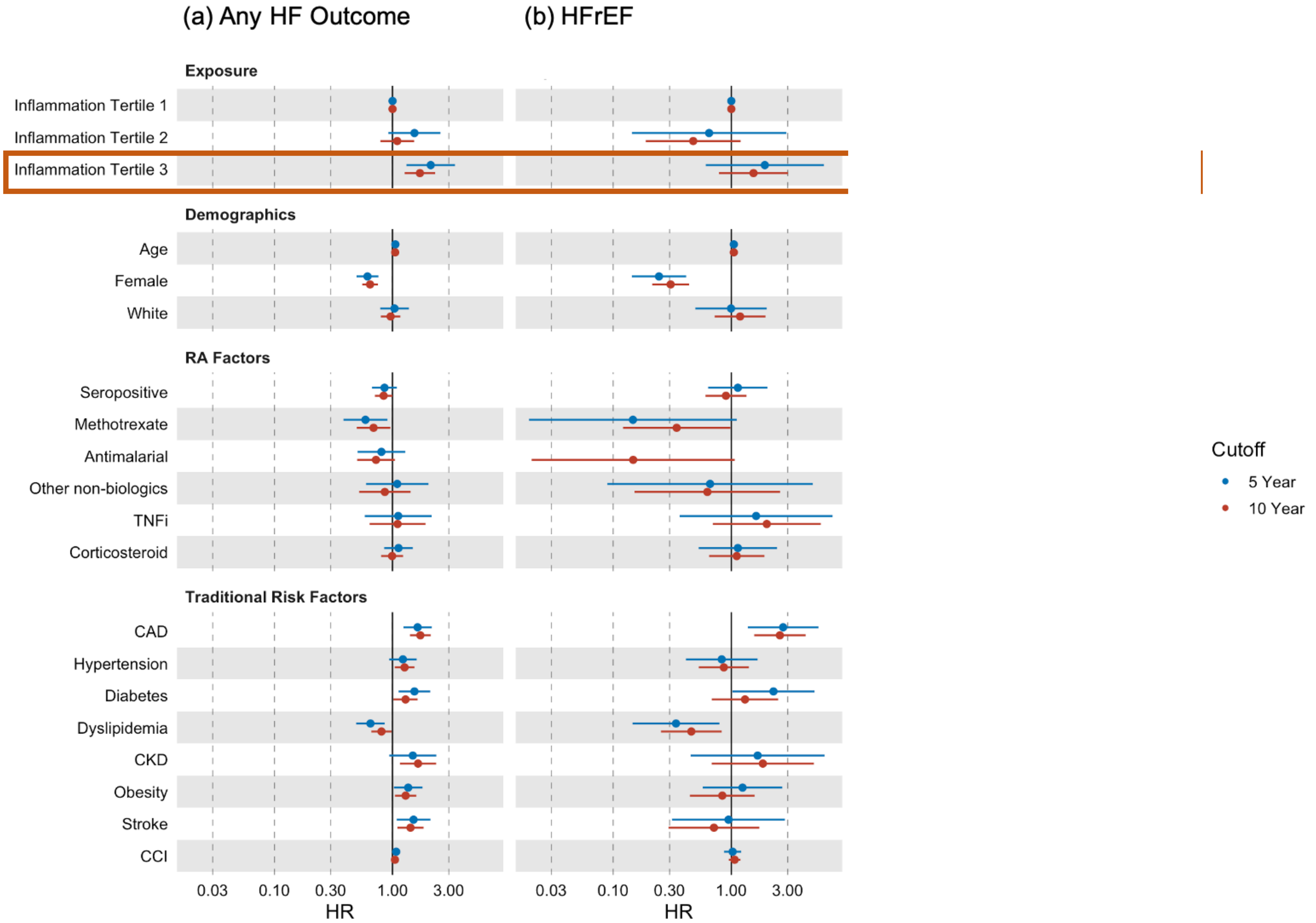
Association between inflammation and HF subtypes



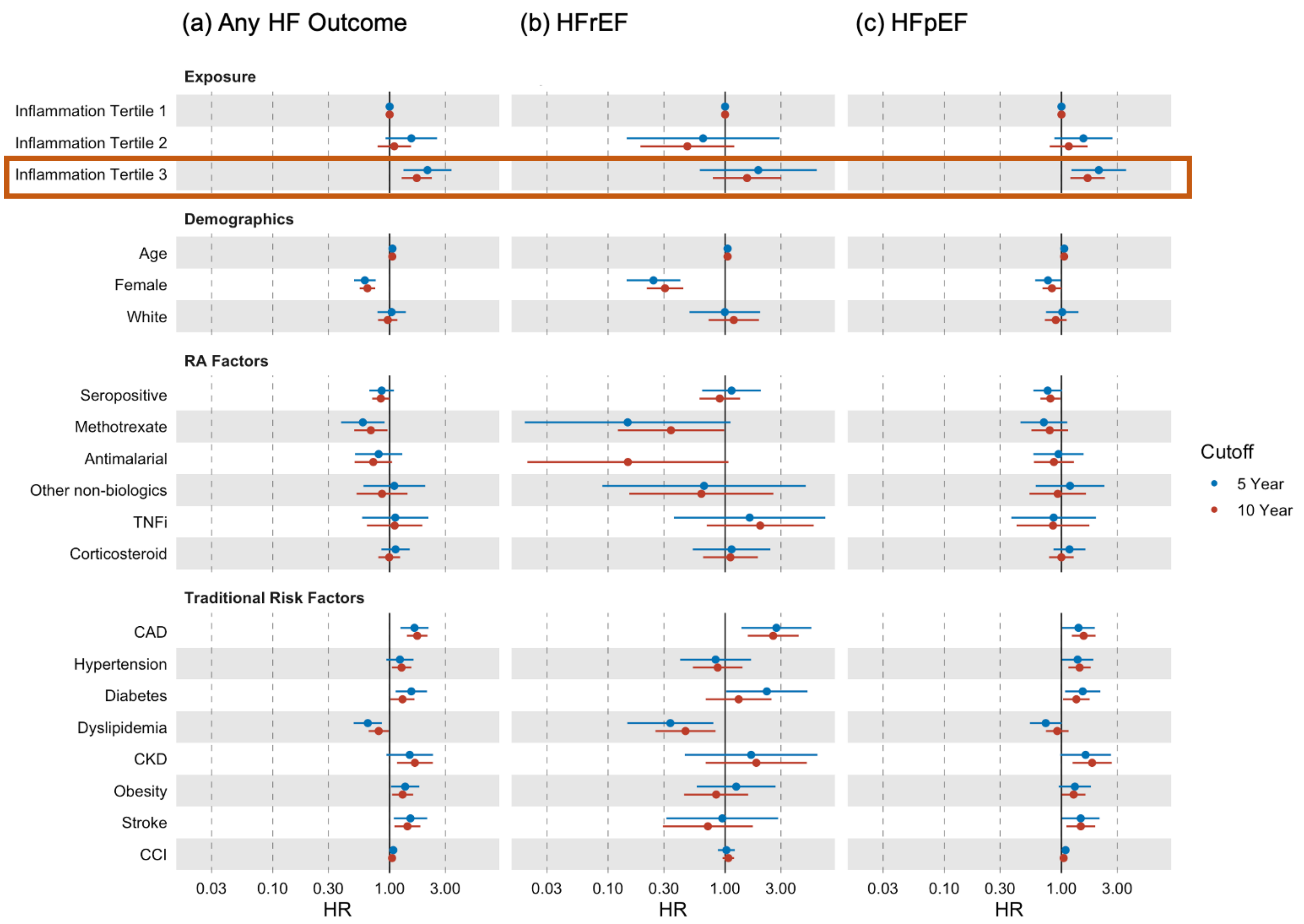
Huang et al., Arth Care & Res 2021



U.S. Department of Veterans Affairs



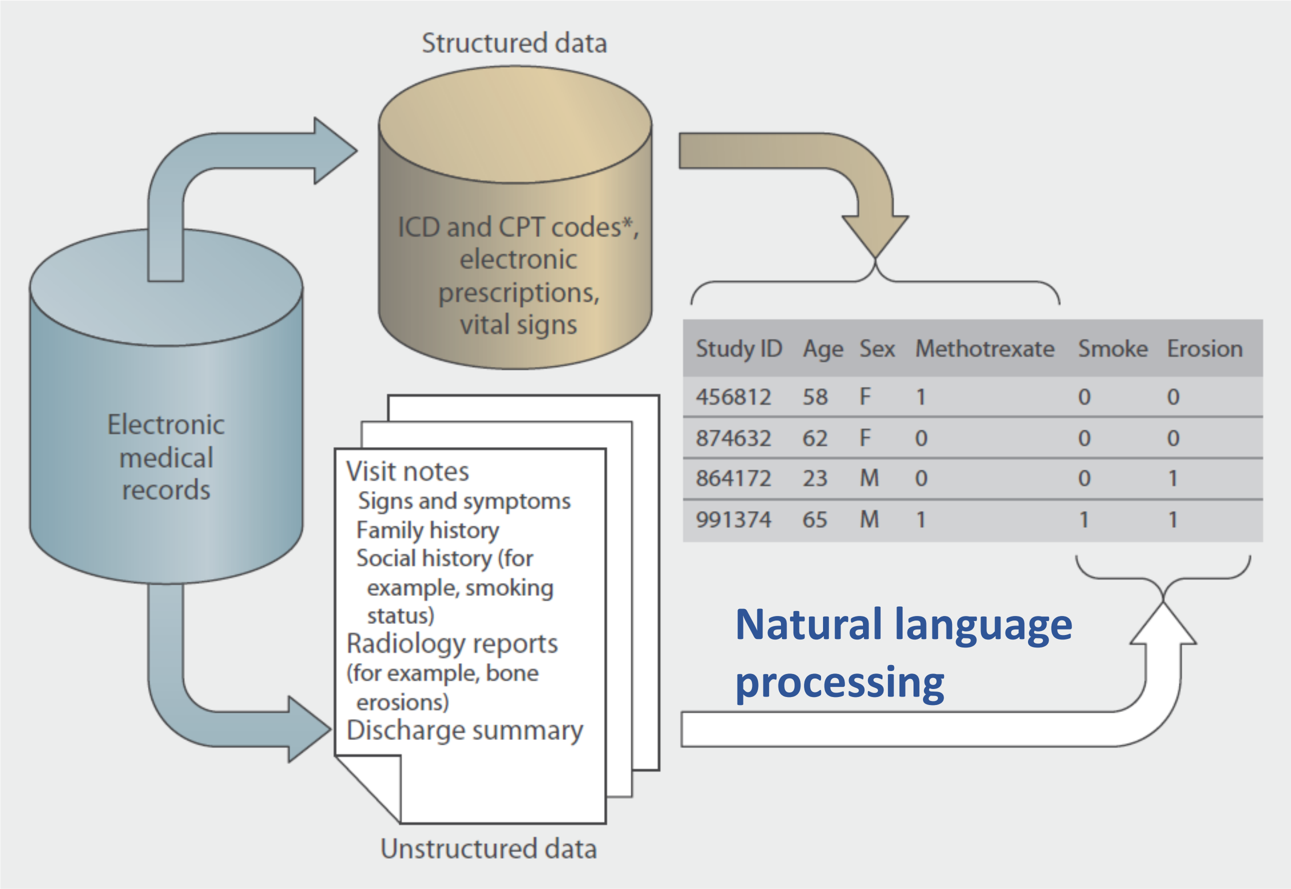
U.S. Department
of Veterans Affairs



U.S. Department of Veterans Affairs

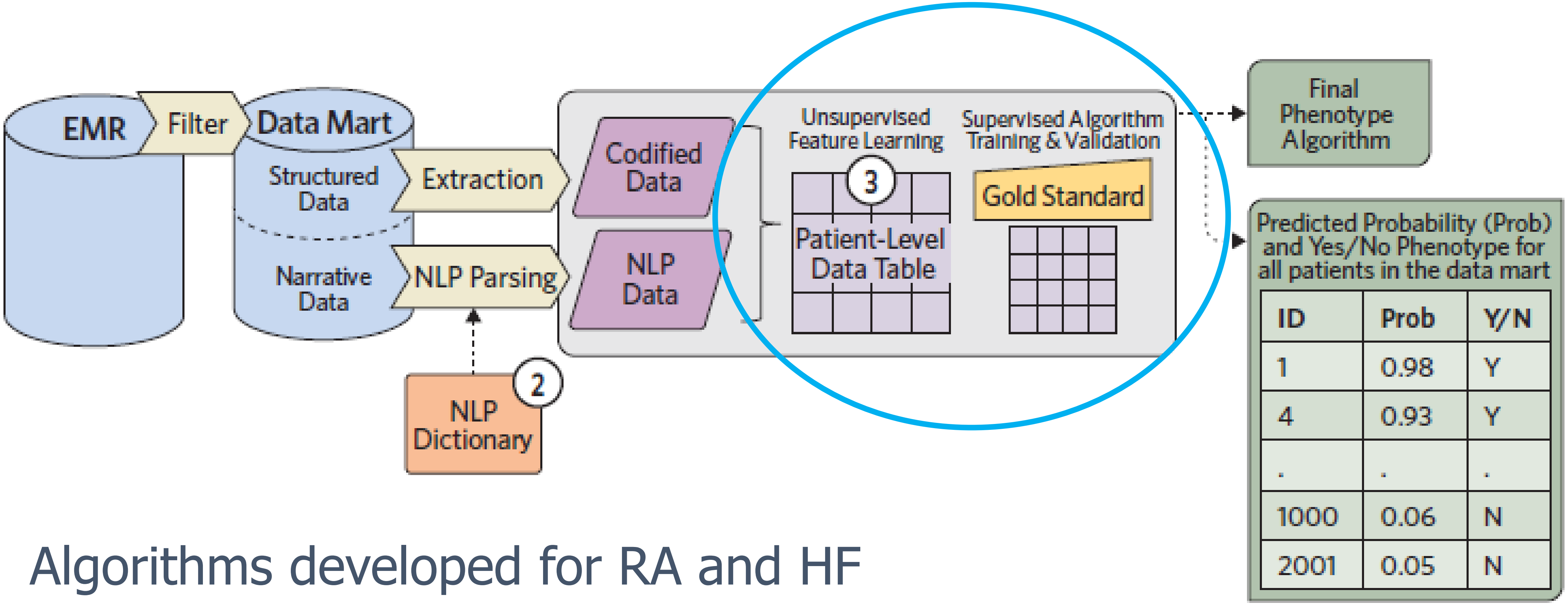
Steps prior to analysis

Types of EHR data



Liao, Cai, et al., BMJ 2015

Machine learning, NLP, and EHR
Pipeline for phenotyping (PheCAP)

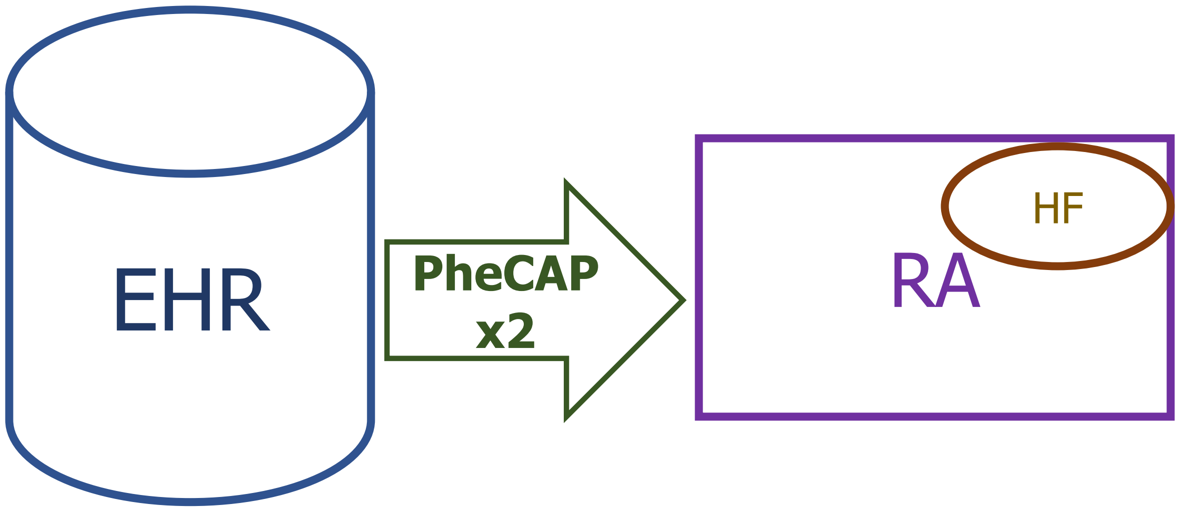


Algorithms developed for RA and HF

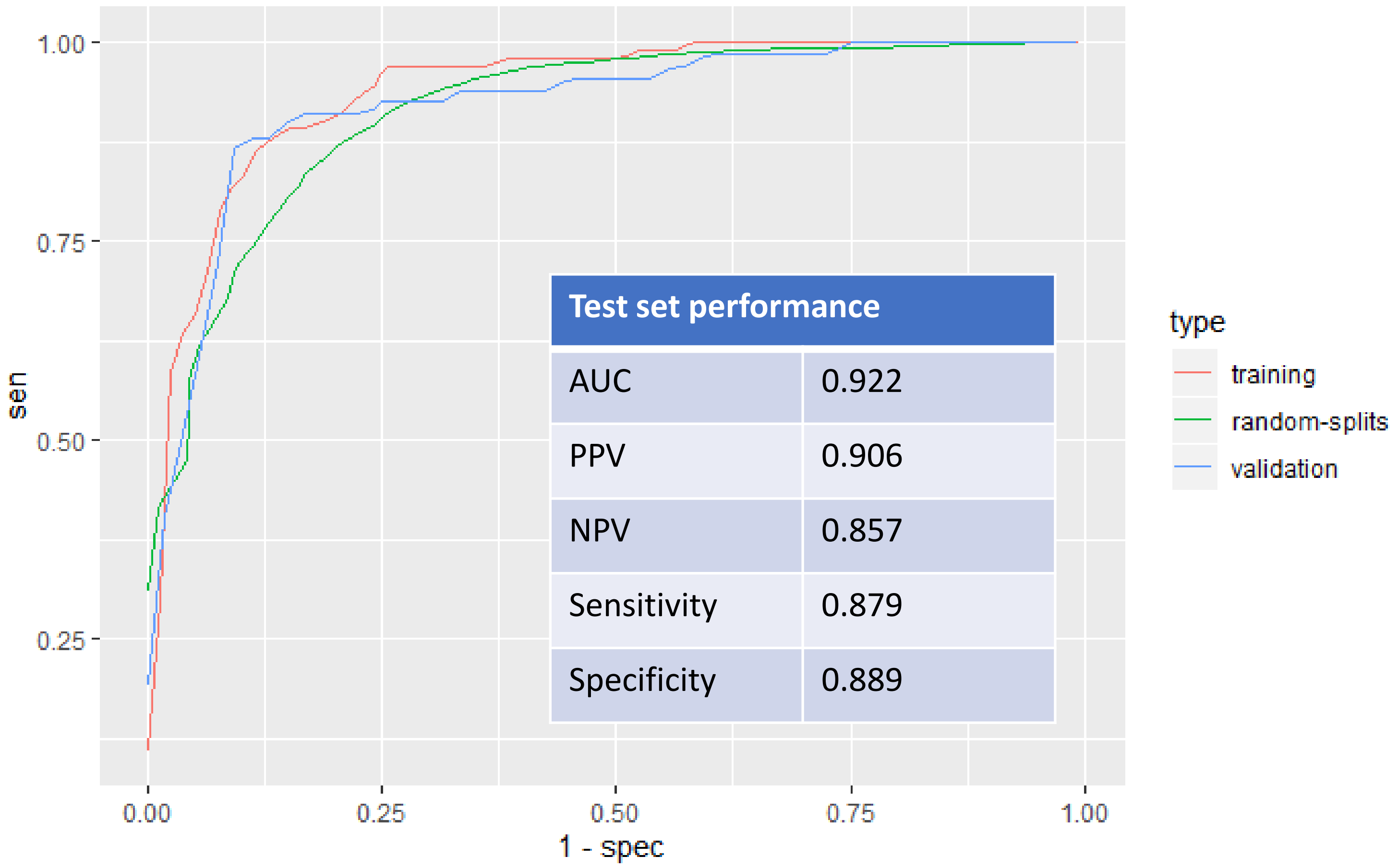
Zhang, Cai, Yu, et al., Nat Protocol 2019

BWH VA U.S. Department of Veterans Affairs

Overview

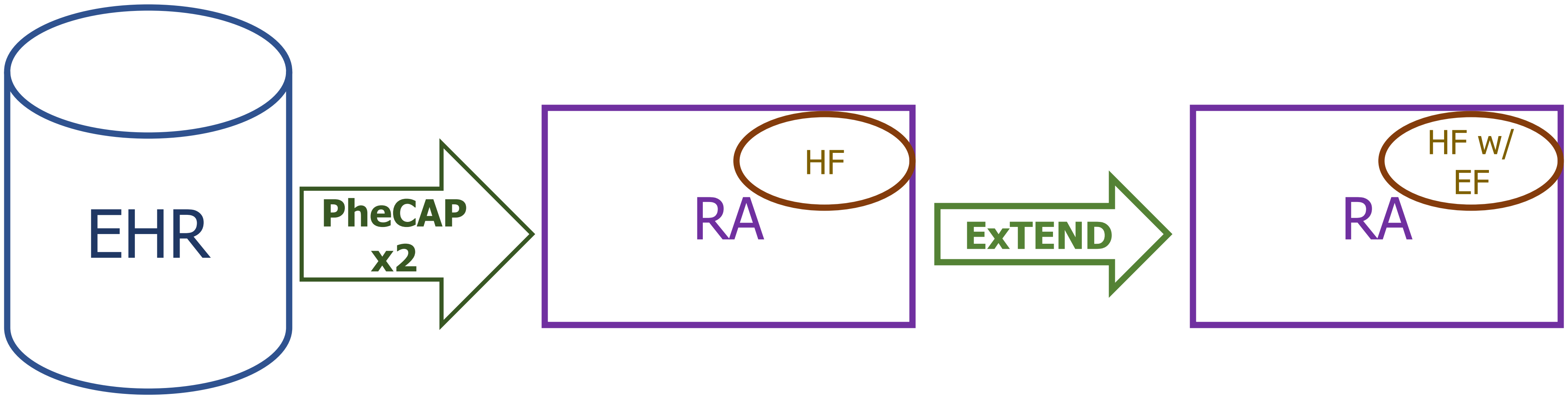


Performance of HF phenotype algorithm

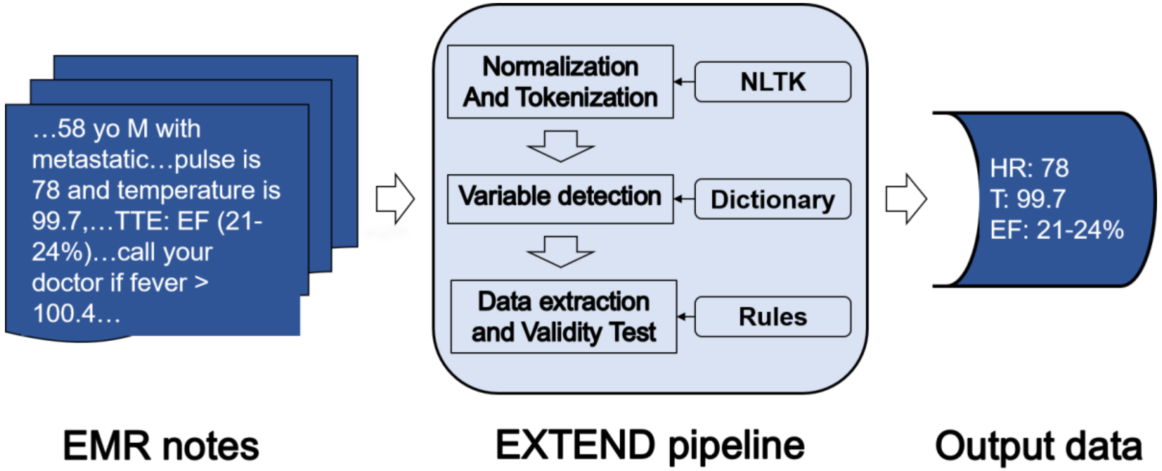


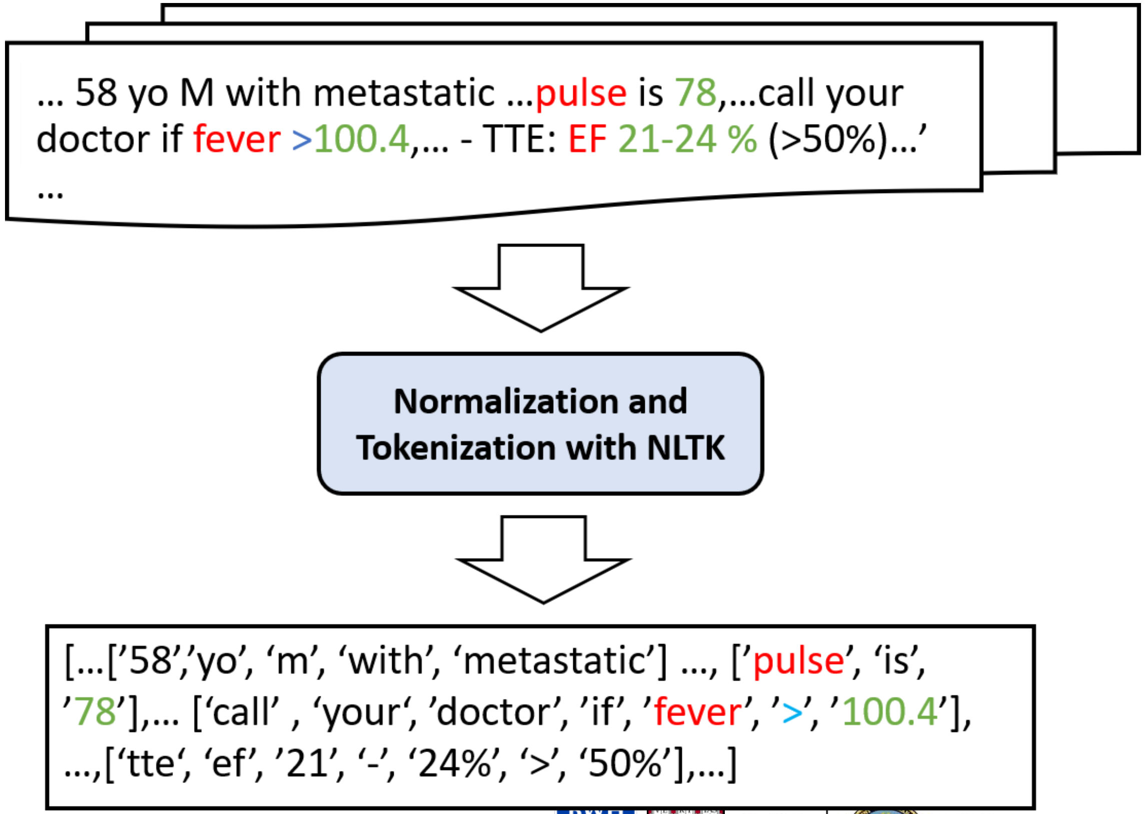
Test set performance	
AUC	0.922
PPV	0.906
NPV	0.857
Sensitivity	0.879
Specificity	0.889

Overview



NLP tool to extract numeric data from narrative notes: ExTEND

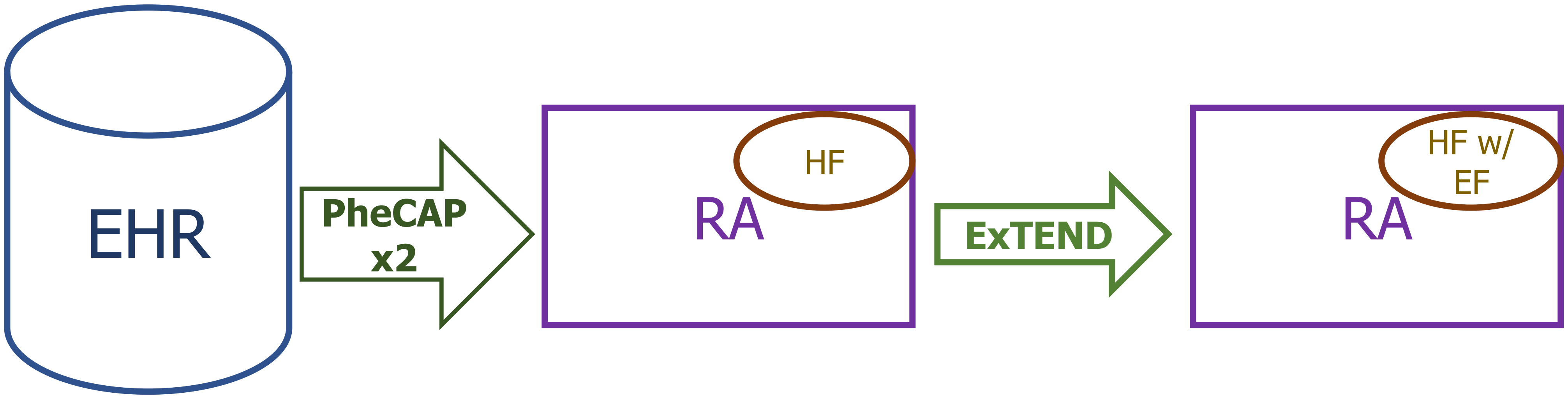




NLTK: The Natural Language Toolkit
https://www.nltk.org/

Cai et al., BMC Medical Inf and Dec Making 2019

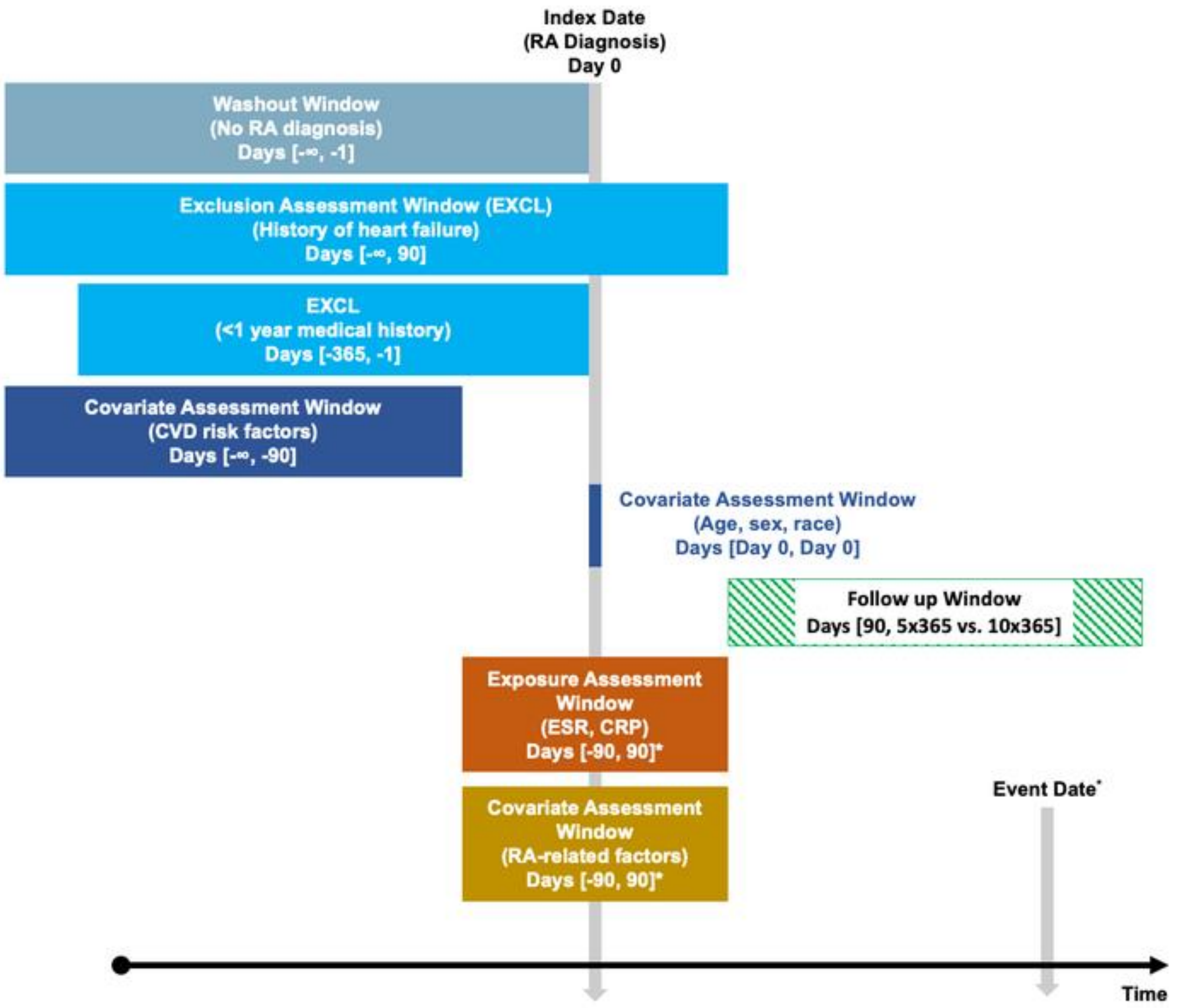
Overview



Method

- EHR based RA cohort, n=~16K
    - Incident RA
        - 1st RA ICD or NLP concept
- Elevated inflammation, extracted from EHR lab values
    - ESR >20-30mm/h
    - hsCRP>8-10mg/L
- Covariates
    - Risk factors for HF, e.g. HTN
- Outcome
    - Algorithm defined HF, PPV 90%
    - EF extracted from EHR
    - Incident HF
        - 1st HF ICD or furosemide NLP, whichever was later

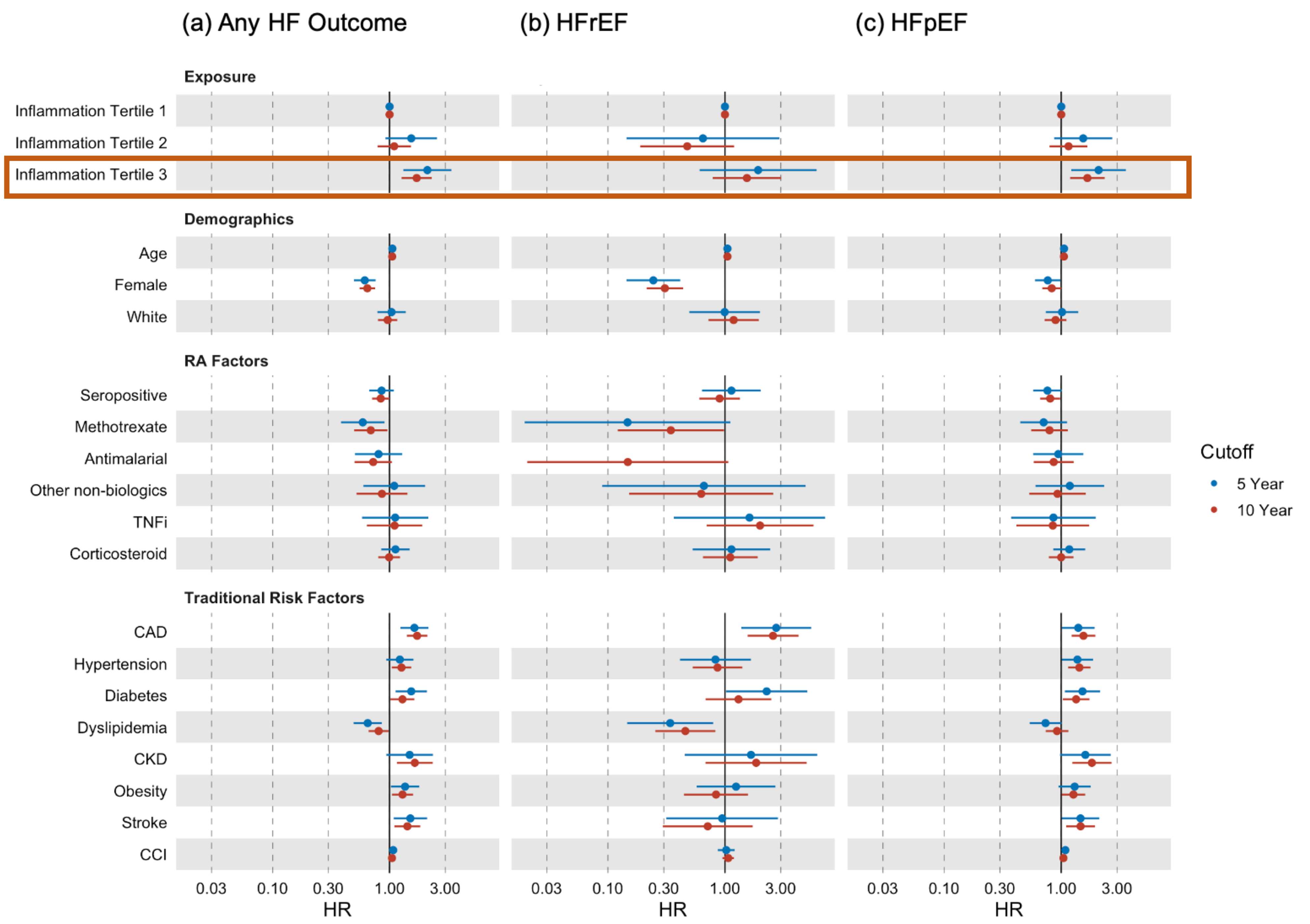
BWH VA U.S. Department of Veterans Affairs



Huang et al., Arth Care & Res 2021

Results

- N=9,087 RA subjects
  - ≥1 year of data prior to 1st RA diagnosis
    - Mean age 56, 77% female, 55% seropositive
- N=749 developed HF
  - N=561 HFpEF
  - N=127 HFrEF
  - Remaining had HFmrEF



U.S. Department
of Veterans Affairs

Conclusion

- HFpEF 4x more common in RA vs HFrEF
- Elevated inflammation a risk factor for HF, independent from traditional risk factors
  - Signal driven by HFpEF

Future directions: Language models & knowledge graphs

Potential applications

Knowledge graphs

Google

harrison ford age

80 years

July 13, 1942

Harrison Ford (**born July 13, 1942**) is an American actor. His films have grossed more than $5.4 billion in North America and more than $9.3 billion worldwide, making him the seventh-highest-grossing actor in North America.

https://en.wikipedia.org › wiki › Harrison_Ford

Harrison Ford - Wikipedia

People also search for

Clint Eastwood
92 years

Calista Flockhart
57 years

Mark Hamill
71 years

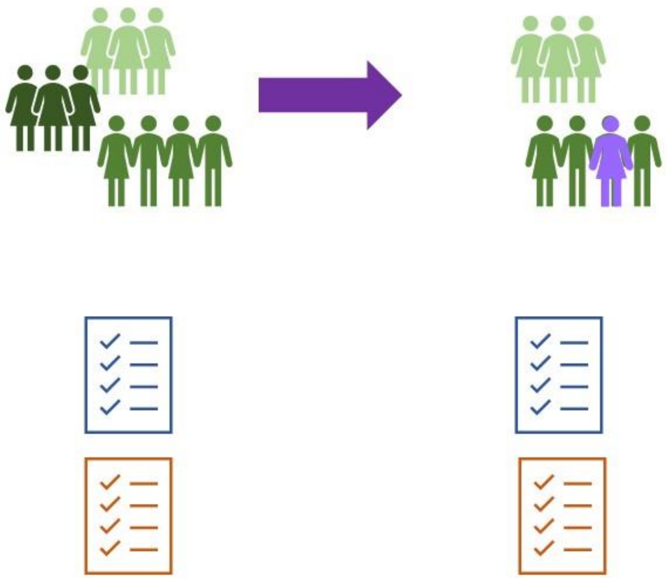
U.S. Department of Veterans Affairs

Google's knowledge graph

https://ahrefs.com/blog/google-knowledge-graph/

- Input data
  - Websites + knowledge sources
- Entity
  - Object or concept
  - Distinct identity
- Edges
  - Connect entities

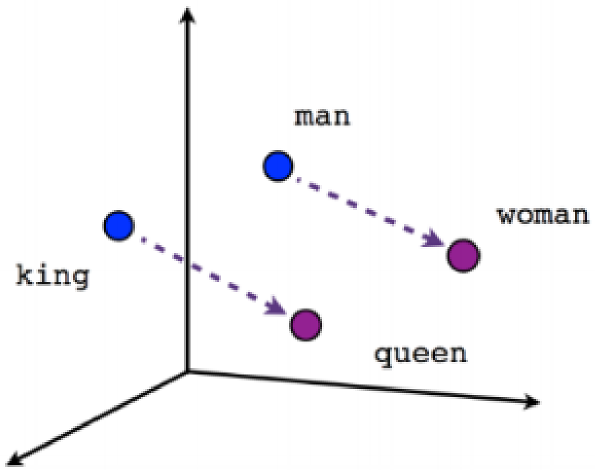
Thousands x millions of data points in EHR



- Each subject with thousands of data points
  - Structured data, e.g. ICD + data extracted w/ NLP
- Challenging to define entities
  - Phenotypes/conditions
    - Definition
  - Lab codes
    - Different across institutions, e.g., no easy way to extract complete blood count
  - Data to extract using NLP

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
;

Co-occurrence matrix

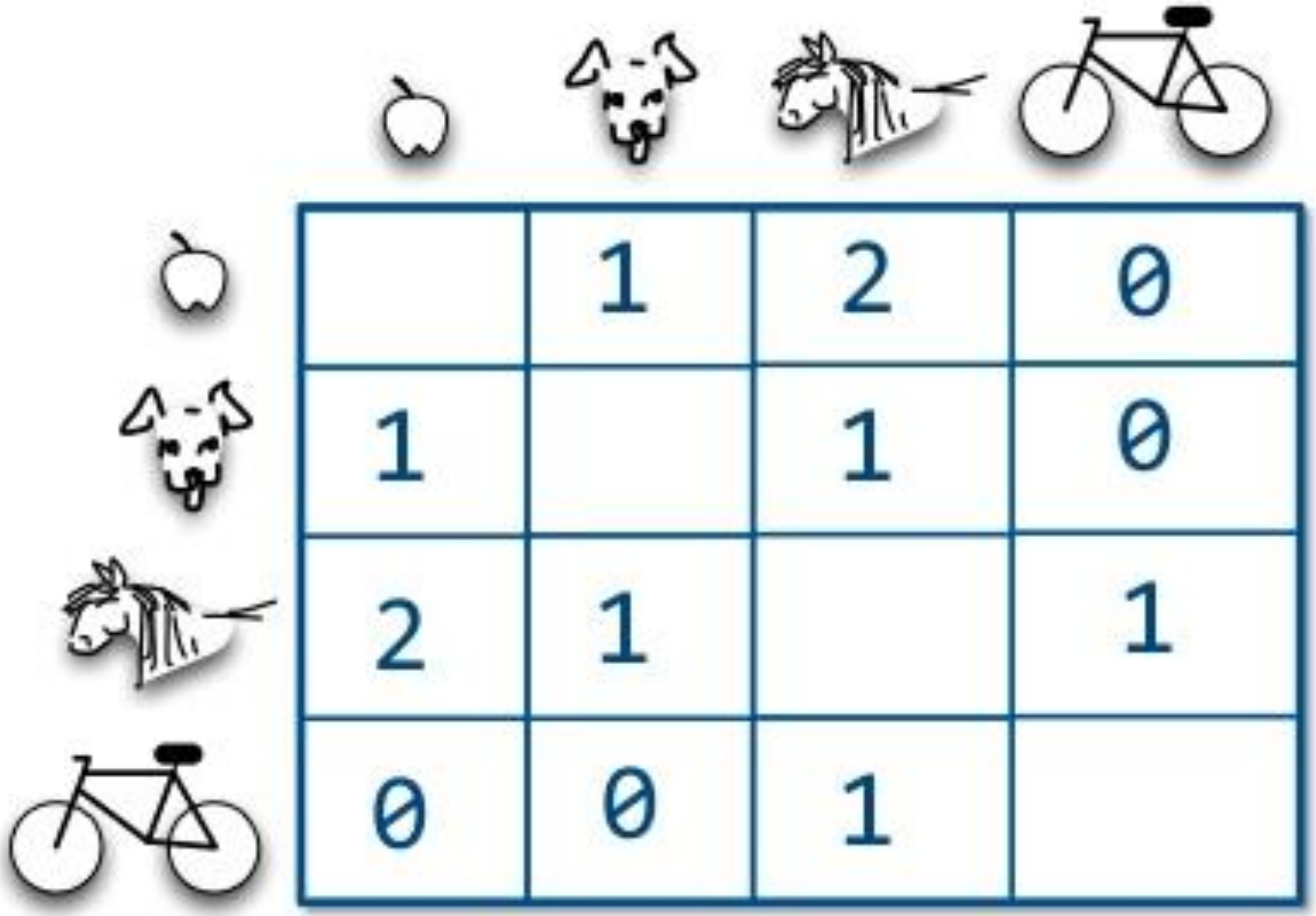
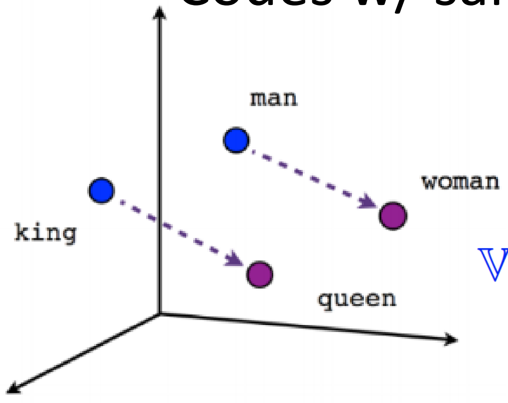


Figure courtesy of Junwei Lu

Represent EHR entities as vectors

- Convert relationships between concepts to analyzable unit, learned representation
  - Codes w/ same meaning have a similar relationship



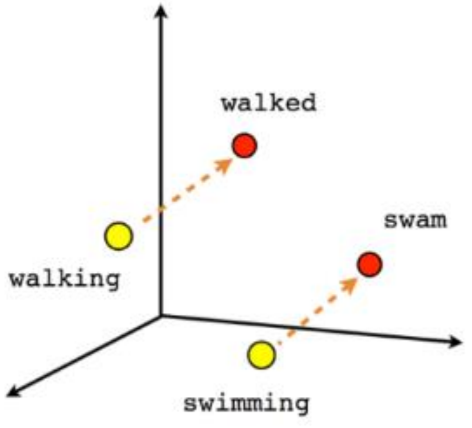
$$\mathbb{VEC}(\text{man}) - \mathbb{VEC}(\text{women}) \approx \mathbb{VEC}(\text{king}) - \mathbb{VEC}(\text{queen})$$

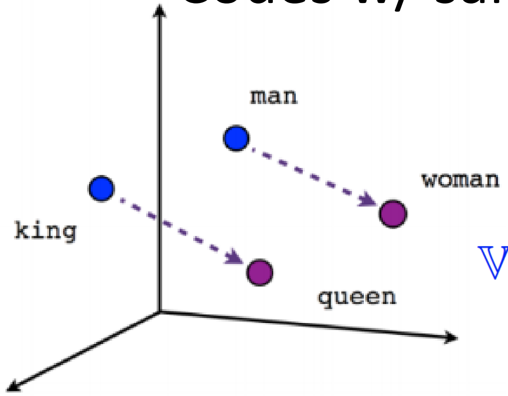
Mikolov, et al. arxiv 2013, https://arxiv.org/pdf/1310.4546

BWH VA U.S. Department of Veterans Affairs

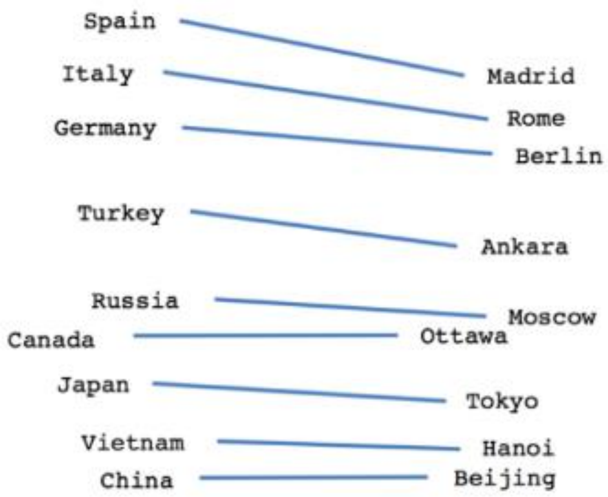
Represent EHR entities as vectors

- Convert relationships between concepts to analyzable unit, learned representation
  - Codes w/ same meaning have a similar relationship





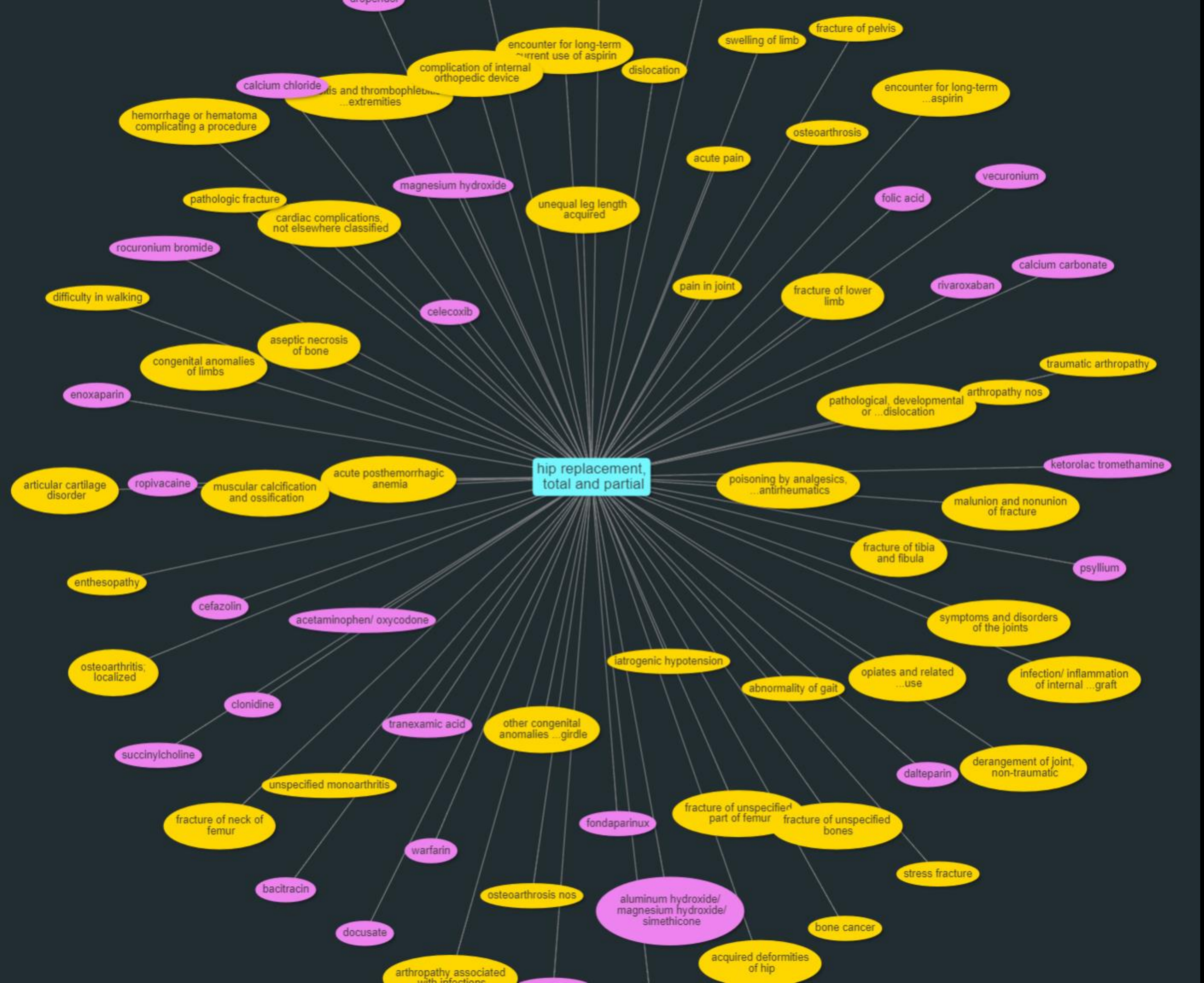
$$\mathbb{VEC}(\text{man}) - \mathbb{VEC}(\text{women}) \approx \mathbb{VEC}(\text{king}) - \mathbb{VEC}(\text{queen})$$



- Organize concepts by vectors
  - Skip-gram model
  - Neural networks

Mikolov, et al. arxiv 2013, https://arxiv.org/pdf/1310.4546

U.S. Department of Veterans Affairs



Knowledge graph

Total & partial hip replacement

Co-trained w/ EHR data from Veterans Affairs & Mass General Brigham

BWH VA U.S. Department of Veterans Affairs

Social determinants of health (SDoH)

- The conditions wherein people are born, live, learn, work, live and age
  - Estimated to be responsible for a similar number of deaths as biological & behavioral factors in the US
  - Example construct: financial insecurity

- ICD codes exist however current coding is low
  - Chart review identified 30% with financial insecurity while 4% had ICD codes in a care management program (frequent users of acute care)
- NLP improves capture of SDoH in EHR
  - "Poverty" not informative for identifying presence of financial insecurity

https://health.gov/healthypeople/priority-areas/social-determinants-health; Galea, et al., Am J Public Health 2011; Chandler et al., under review

BWH VERITAS VA U.S. Department of Veterans Affairs

Social determinants of health (SDoH)

- "Poverty" used in context of neurologic exam
  - Poverty of speech

- Future potential for language models and knowledge graphs to identify factors related to SDoH as documentation matures

https://health.gov/healthypeople/priority-areas/social-determinants-health; Galea, et al., Am J Public Health 2011; Chandler et al., in preparation

Summary

- Interdisciplinary projects- team sport!
- EHR data as alternative or complementary source for clinical research studies
  - Randomized controlled trials, prospective cohort studies, admin database
- Integration of NLP and AI into current framework for epidemiologic studies
- Future
  - Applications of language models
    - Learning from our own data
    - Studies incorporating context where data are located

TRANSLATIONAL DATA SCIENCE
CELEHS
CENTER FOR A LEARNING HEALTH SYSTEM

Thank you

VERITY BIOINFORMATICS CORE TEAM		
BWH	**Harvard Medical School**	**MGB Research Computing**
Greg McDermott Mary Jeffway Tianrun Cai Feng Liu Jackie Stratton Kumiko Schnock Yumeko Kawano Dana Weisenfeld	Tianxi Cai Vidul Panickan Clara Lea-Bonzel Mohammed Moro Sara Morini Xin Xiong Junwei Liu	Andrew Cagan

NIAMS
P30 AR072577
R21 AR078339
R01 AR080193

VA Boston

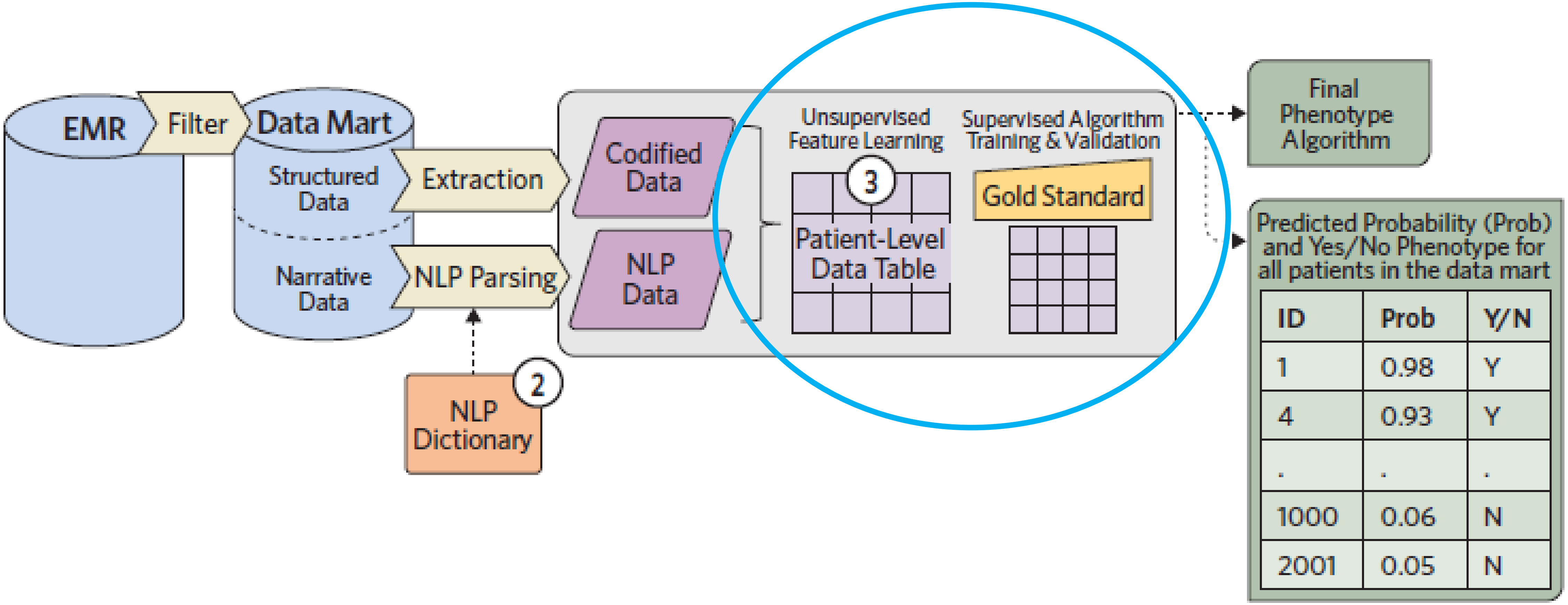
J Michael Gaziano
Kelly Cho
Jacqueline Honerlaw
Anne Ho
Lauren Costa

Alicia Chen
Rahul Sangar
Connor Melley
Vidisha Tanukonda
Monika Maripuri
Ashley Galloway

Michael Murray
Paul Monach
Dan Posner

VA
U.S. Department of Veterans Affairs
Veterans Health Administration
Office of Research & Development

Machine learning, NLP, and EHR
Pipeline for phenotyping



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