Electronic Health Records Phenotyping

Jesse Gronsbell
Jianhui Gao
Siyue Yang

Department of Statistical Sciences
University of Toronto

SSC Annual Meeting, Biostatistics Workshop
June 5, 2022

Introductions

Jesse Gronsbell

- Assistant professor in Statistical Sciences
- Cross-appointed in Family & Community Medicine
- Spent several years at Alphabet's Verily
- AI Lead at UTOPIAN

Introductions

Jianhui Gao

- First year PhD student in Statistical Sciences
- Co-supervised by Profs. Gronsbell & Sun
- Developing statistical methods for GWAS with EHR-linked biobanks

Introductions

Siyue Yang

- Third year PhD student in Statistical Sciences
- Co-supervised by Profs. Gronsbell & Sun
- Developing semi-supervised methods for evaluating algorithmic fairness

Acknowledgment

- Tianxi Cai (Harvard)
- Xu Shi (University of Michigan)
- Chuan Hong (Duke)
- Aaron Sonabend (Google)
- Molei Liu (Harvard)

Housekeeping

- All materials for this workshop can be found here
- Please feel free to ask questions throughout the workshop
- We will go through code on slides
  - Markdown files are available in the repo

Roadmap for today

- What is phenotyping and why does it matter?
- Background on electronic health record (EHR) data
- A brief history of phenotyping
- Supervised machine learning (ML) methods for phenotyping
- Weakly and semi-supervised ML methods for phenotyping
- Ongoing phenotyping research

Goals for today

- Introduce EHR research
- Understand the importance & challenge of phenotyping
- Develop intuition for statistical learning methods
- Learn to implement phenotyping methods in `R`

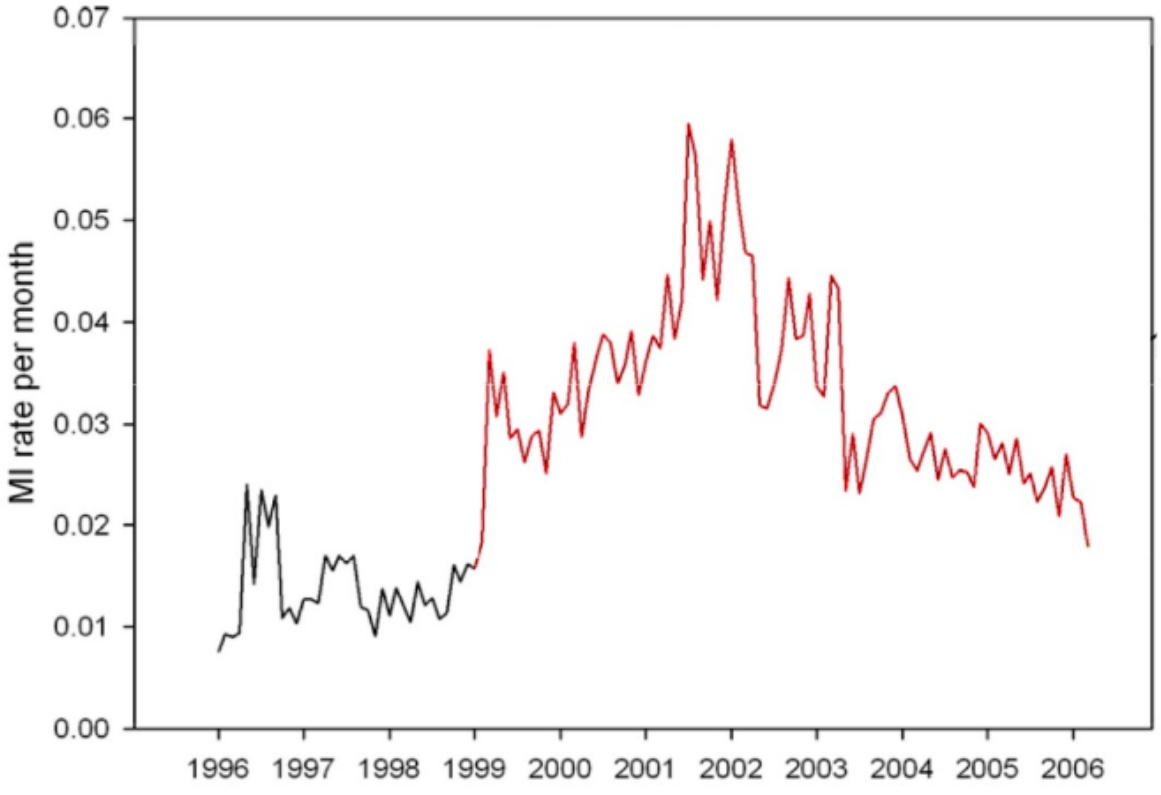
Things to keep in mind

- We'll spend a lot of time describing the problem & data
- Prediction is more of an art than a science
- Today's focus is on traditional ML methods
- The methods we discuss are applicable to other settings
- We only have a few hours

Part I
What is phenotyping and why does it matter?

Why I work with EHR data

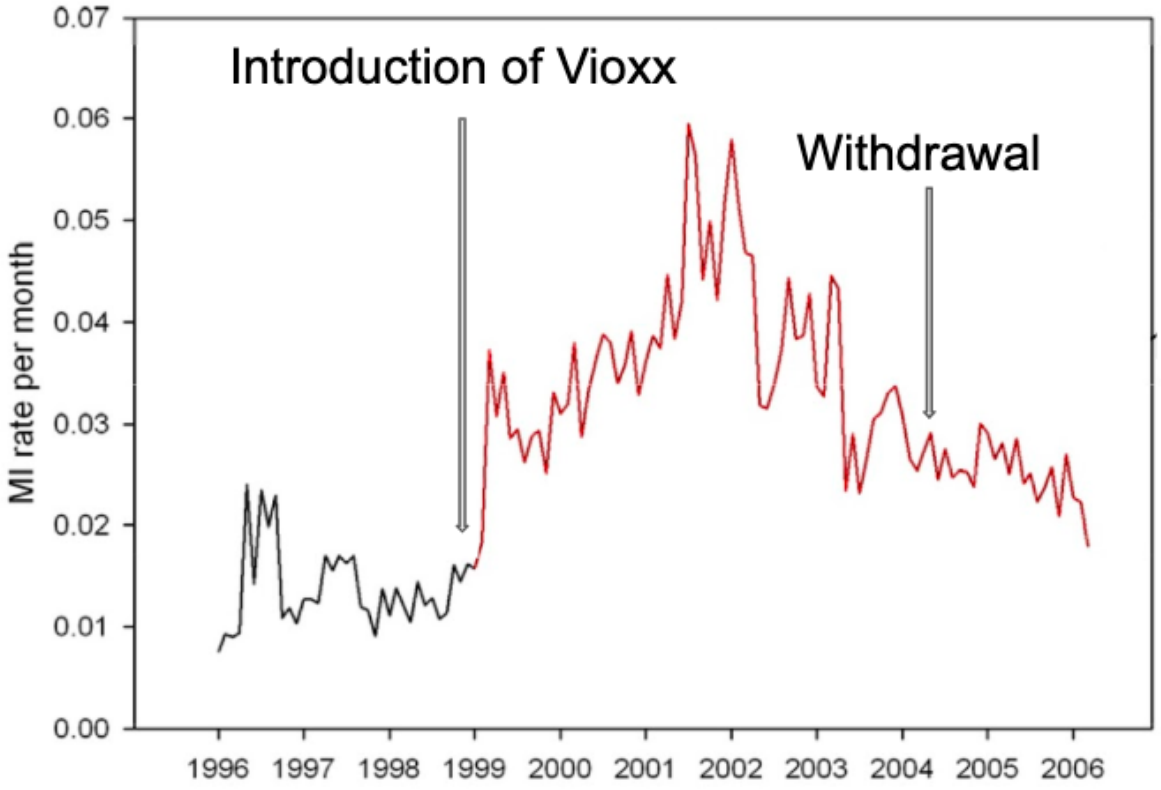
Hospitalizations due to heart attack (MI) at Partners Healthcare



Brownstein et al 2007

Why I work with EHR data

Hospitalizations due to heart attack (MI) at Partners Healthcare

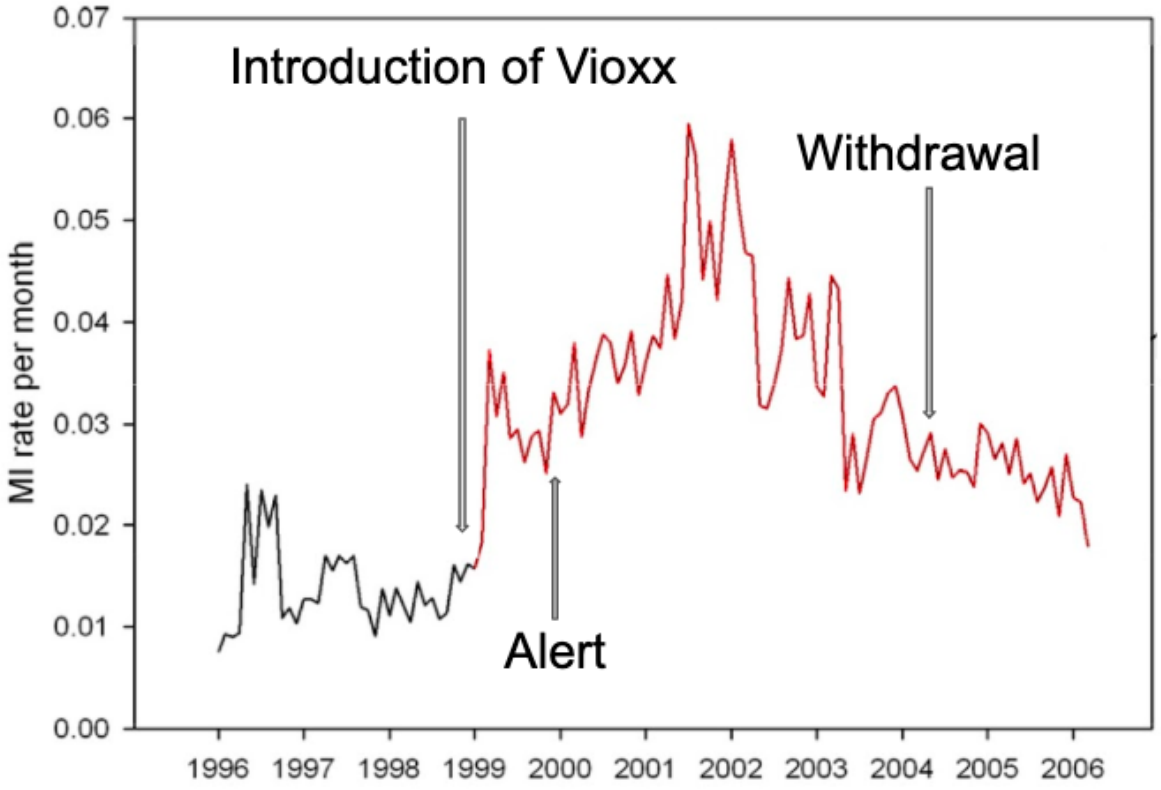


Brownstein et al 2007

5 years Vioxx was on the market $\rightarrow$ 140,000 heart attacks

Why I work with EHR data

Hospitalizations due to heart attack (MI) at Partners Healthcare



Brownstein et al 2007

monitor EHR $\rightarrow$ prevent tens of thousands of heart attacks

The potential of EHR data

"Making our data work for us."

Isaac Kohane

But this isn't so easy...

→ Vioxx example

- Main task was to identify patients with evidence of MI
- All patients hospitalized for MI with a primary or admitting diagnosis ICD-9 code 410

But this isn't so easy...

→ Vioxx example

- Main task was to identify patients with evidence of MI
- All patients hospitalized for MI with a primary or admitting diagnosis ICD-9 code 410

"In our own chart review, we found that 87% of patients assigned the code had strong confirmatory evidence of acute MI."

What is phenotyping?

The process of inferring a phenotype from the information in a patient's EHR

What is a phenotype?

Phenotype

A clinical condition, characteristic, or set of clinical features that can be determined solely from the data in EHRs

What is a phenotype?

Phenotype

A clinical condition, characteristic, or set of clinical features that can be determined solely from the data in EHRs

Also known as

- Computable phenotypes
- Case definitions
- EHR phenotype

What is a phenotype?

Phenotype

A clinical condition, characteristic, or set of clinical features that can be determined solely from the data in EHRs

e.g.

- Presence of a disease
- Time of disease onset
- Treatment response

What is a phenotype?

Phenotype

A clinical condition, characteristic, or set of clinical features that can be determined solely from the data in EHRs

e.g.

- **Presence of a disease**
- Time of disease onset
- Treatment response

Top 10 phenotypes

Phenotype of interest	Number of studies
Cancer	26
Diabetes	23
Heart failure	5
Rheumatoid arthritis	5
Cataract	4
Drug side effect	4
Pneumonia	4
Asthma	3
Peripheral arterial disease	3
Hypertension	3

Chivade et al 2013

Most literature concerns prevalent or chronic conditions

This is changing...

Phenotype of interest	Number of studies
Rheumatoid arthritis	8
Epilepsy	7
Opioid abuse	7
Type 2 diabetes mellitus	7
Alcohol abuse	6
Chronic obstructive pulmonary disease	6
Coronary artery disease	6
Ulcerative colitis	6
Congestive heart failure	5
Crohn's disease	5
Dementia	5
Homelessness	5
Tobacco abuse	5
Type 1 diabetes mellitus	5

Why does phenotyping matter?

Most common use: Cohort discovery

Identify patients with a given condition

Why does phenotyping matter?

Most common use: Cohort discovery

Identify patients with a given condition

Phenotyping is the first step of any EHR-based application

Applications of phenotyping

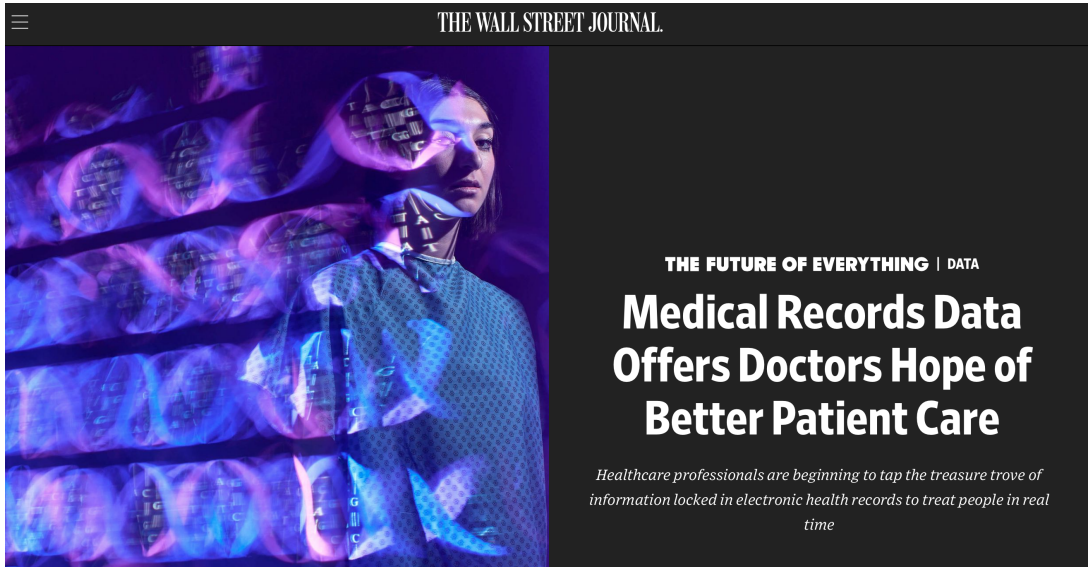
e.g. Population-based surveillance like in the Vioxx example

Applications of phenotyping

Study type	Use cases
Cross-sectional	Epidemiological research
	Hospital administration/resource allocation
	Adherence to diagnostic/treatment guidelines
	Quality measurement
Association (case-control/cohort)	Genome-wide association studies
	Pharmacovigilance
	Identifying clinical risk factors and protective factors
	Clinical decision support
	Clinical effectiveness research
	Predictive modeling
Experimental	Clinical trial recruitment
	Pragmatic trials
	Adaptive/randomized, embedded, multifactorial, adaptive platform trials

Banda et al 2018

The dream: Use EHR data to provide better care



Healthcare professionals are beginning to tap the treasure trove of information locked in electronic health records to treat people in real time

A "Green Button" for patients like mine

"a **green patients like mine button** as a tool in the EHR would both **support patient care decisions** in the absence of published evidence and, as a byproduct, quantify and **prioritize unanswered clinical questions** for EHR-enabled randomization at the point of care"

Longhurst et al 2014

A "Green Button" for patients like mine

"a **green patients like mine button** as a tool in the EHR would both **support patient care decisions** in the absence of published evidence and, as a byproduct, quantify and **prioritize unanswered clinical questions** for EHR-enabled randomization at the point of care"

Longhurst et al 2014

The Green Button is still an aspiration

Why we're here today

EHRs do not have readily available information on phenotypes

Why we're here today

EHRs do not have readily available information on phenotypes

Next: Why is this the case?

- Part II: Background on EHR data

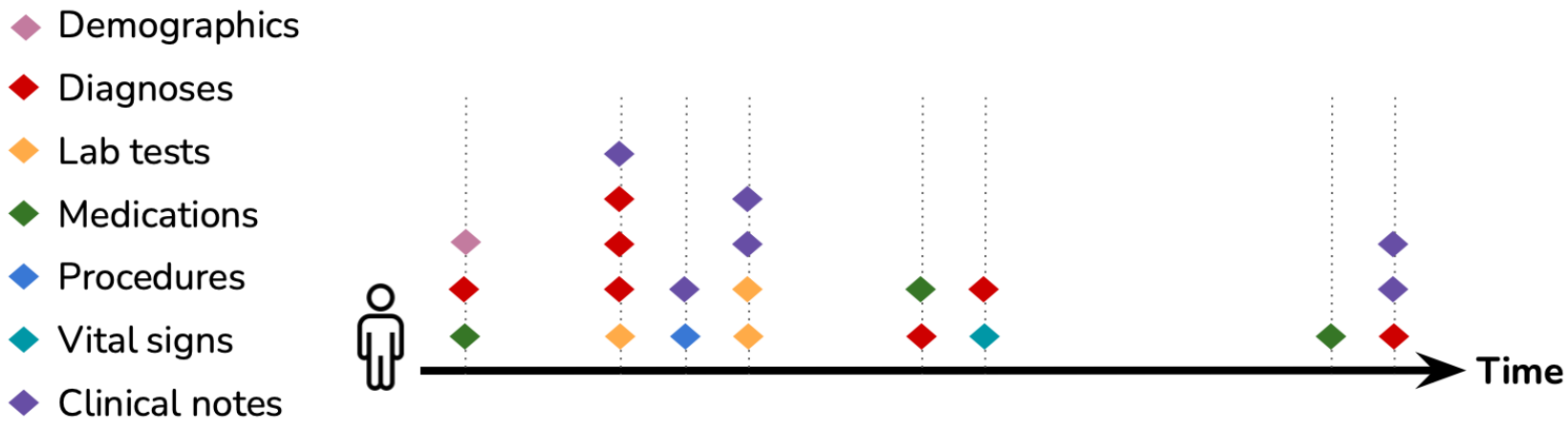
Later: How do we deal with this?

- Part III: A brief history of phenotyping
- Part IV: Supervised ML methods for phenotyping
- Part V: Weakly and semi-supervised ML methods

Part II
Background on electronic health record (EHR) data

What is an Electronic Health Record (EHR)?

An electronic record of a patient's interactions with a healthcare system



An aside on terminology

- Electronic health record: Complete healthcare record maintained by the healthcare provider
- Electronic medical record: Partial healthcare record maintained by the healthcare provider
- Personal health record: Partial or complete healthcare record maintained by the patient

Encounters: Units of a patient's EHR

Encounter

An interaction with the healthcare system

e.g.

- Inpatient visit
- Outpatient visit
- Emergency department visit

Encounters: Units of a patient's EHR

At each encounter we know:

- Who the patient saw
- What happened
- When the patient was seen
- Where the patient was seen

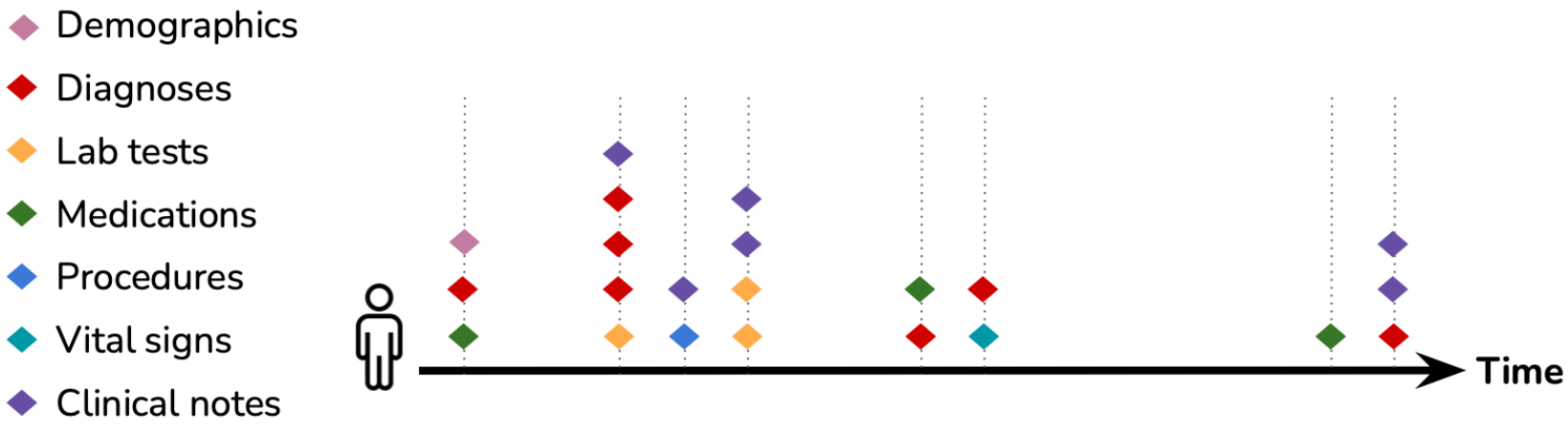
Encounters: Units of a patient's EHR

At each encounter we know:

- Who the patient saw
- What happened
- When the patient was seen
- Where the patient was seen
- We do not have explicit information on **why**
  - ⋆ e.g. Diagnoses relate to the chief complaint

Why we don't have the why

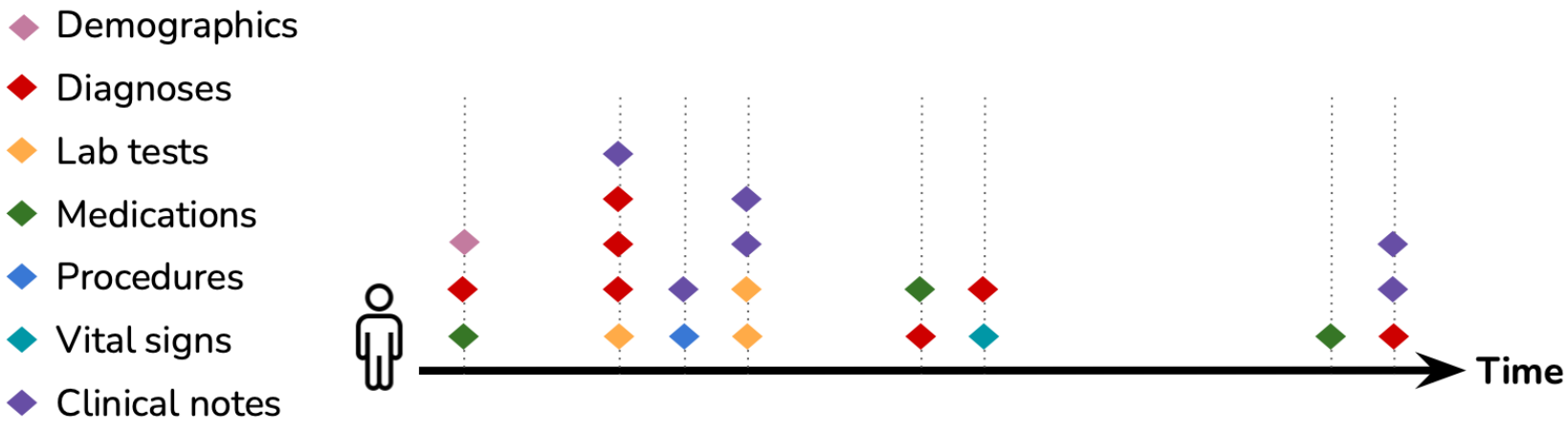
An electronic record of a patient's interactions with a healthcare system



EHR data is a byproduct of clinical care

Why we don't have the why

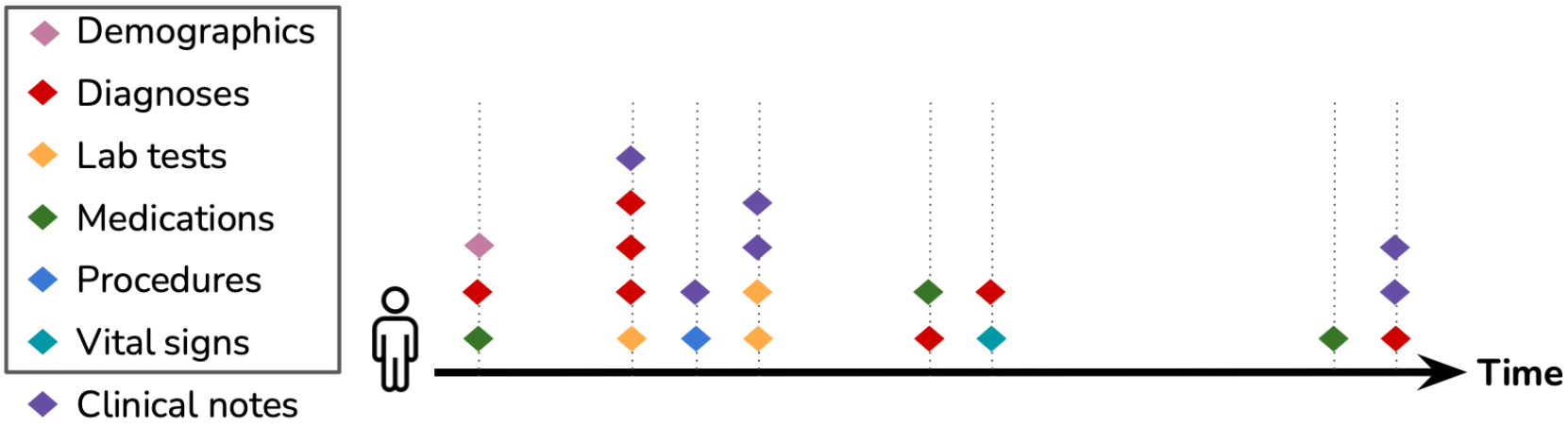
An electronic record of a patient's interactions with a healthcare system



EHR data is a byproduct of clinical care

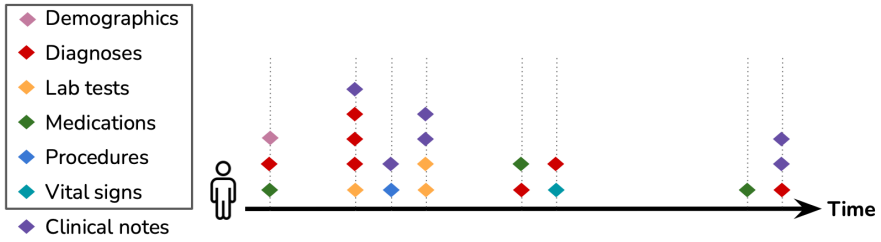
$\rightarrow$ We can understand the complexity this brings by digging into the data

The two flavors of EHR data



1. **Structured data**: Stored in a database

The two flavors of EHR data



1. Structured data: Stored in a relational database

Demographics

- Patient identification number
- Age
- Sex
- Address
- Payer
- Race and ethnicity

Structured data

Diagnoses

- Reason for an encounter

Lab tests & Vitals

- Underlying clinical state

Medications

- Treatments received

Procedures

- What was done

Structured data

Diagnoses

- Reason for an encounter
- Don't always correspond to main concern

Lab tests & Vitals

- Underlying clinical state
- Irregularly measured
- Reconciling multiple panels

Medications

- Treatments received
- Indicate a prescription, not dosage or adherence

Procedures

- What was done
- Tied to reimbursements

Structured data

Diagnoses

- Reason for an encounter
- Don't always correspond to main concern

Lab tests & Vitals

- Underlying clinical state
- Irregularly measured
- Reconciling multiple panels

Medications

- Treatments received
- Indicate a prescription, not dosage or adherence

Procedures

- What was done
- Tied to reimbursements

Lack the context we often need for research

Common terminologies for structured data

Diagnoses

- ICD-9 & 10 codes

Lab tests & Vitals

- LOINC codes

Medications

- RxNorm codes

Procedures

- CPT codes

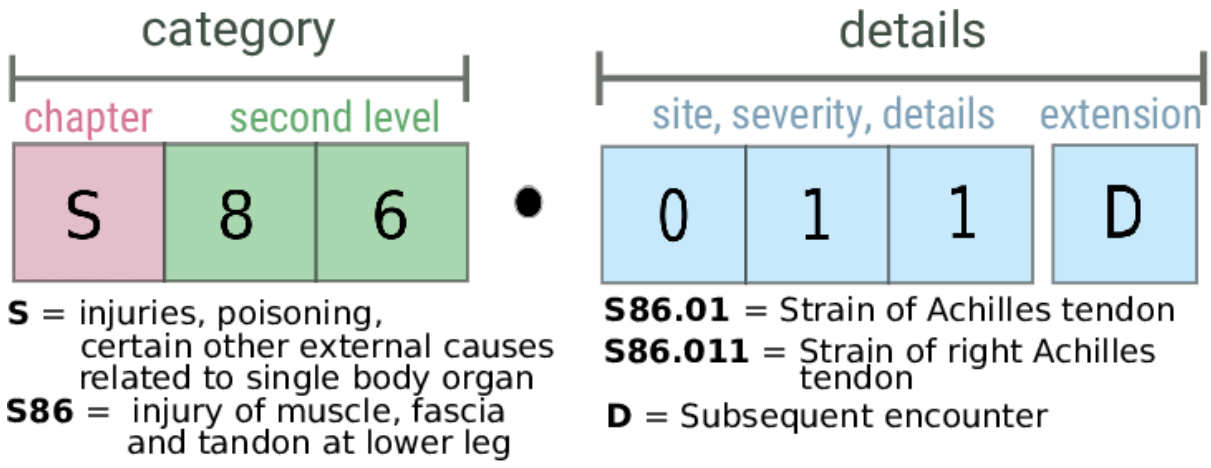
Make the data *relatively* easy to work with

International Classification of Disease (ICD) codes

- 2015: Shift from ICD-9 to ICD-10
- ICD-10 is much more granular than ICD-9
  - ICD-9: 13,000 codes vs. ICD-10: 68,000 codes

International Classification of Disease (ICD) codes

The anatomy of an ICD-10 code

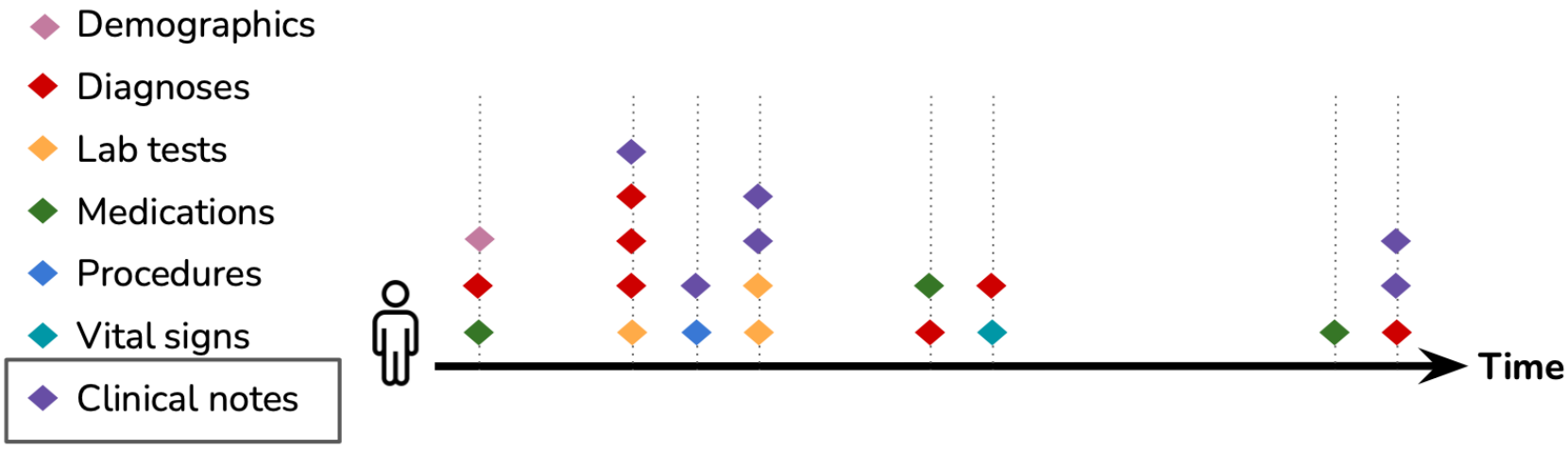


- We often roll up to the first three characters
- Always work with clinicians and informatics experts to understand codes

Other structured data elements

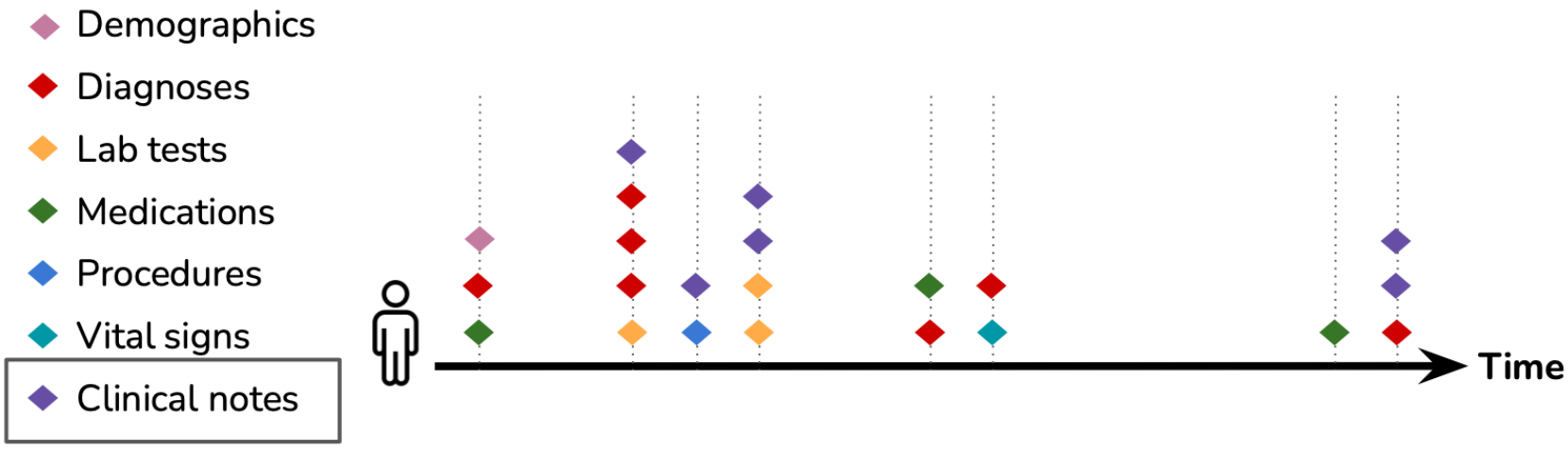
- Problem lists
- Provider data
- Smoking status, drug and alcohol use, employment status, marital status
- Patient reported outcomes

The two flavors of EHR data



2. **Unstructured data**: Stored in free-form

The two flavors of EHR data



2. **Unstructured data**: Stored in free-form

Note: Medical images are also often available, but currently not widely used in phenotyping

Types of clinical notes

- Progress notes
- Radiology reports
- Discharge summaries
- Echocardiography reports
- ...

Types of clinical notes

- Progress notes
- Radiology reports
- Discharge summaries
- Echocardiography reports
- ...

In contrast to structured data, clinical notes require natural language processing (NLP)

NLP: Language understanding by a computer

NLP of clinical notes

- **Traditional approach**: Manually come up a term list
  - ⋆ e.g. joint pain, joint pains, painful joints, arthralgia, rheumatoid arthritis, morning stiffness, CRP, etc.

NLP: Language understanding by a computer

NLP of clinical notes

- **Traditional approach**: Manually come up a term list
  - ⋆ e.g. joint pain, joint pains, painful joints, arthralgia, rheumatoid arthritis, morning stiffness, CRP, etc.
- **Alternative approach**: Unified Medical Language System (UMLS)
  - ⋆ e.g. Concept for joint pain assigned a unique concept unique identifier (CUI), C0003862
  - ⋆ C0003862 contains 35 English synonyms such as painful joints, joint pains, joint pain, arthralgia

NLP $\neq$ text search

- Negation
  - ⋆ The patient does not have cancer.
- Family history
  - ⋆ Father had heart failure.
- Inverted syntax
  - ⋆ Colon, ascending and descending, biopsy.
- Relation
  - ⋆ Tamoxifen is used to treat breast cancer.

Basic NLP for phenotyping

1. Parse notes to identify clinical terms
2. Map the clinical terms to CUIs
3. Remove negations
4. Remove family history

Basic NLP for phenotyping

1. Parse notes to identify clinical terms
2. Map the clinical terms to CUIs
3. Remove negations
4. Remove family history

$\rightarrow$ Positive mentions of relevant CUIs about the patient

Clinical note parsing with NILE

Narrative Information Linear Extraction (NILE)

Introduction

NILE is an efficient and effective software for natural language processing (NLP) of clinical narrative texts. It uses a prefix tree algorithm for named entity recognition, and finite-state machines for semantic analysis, both of which were inspired by the natural reading behavior of humans. The design aims to directly translate linguistic and clinical knowledge to code, allowing for the development of functions to parse complex language patterns.

The software was developed by Sheng Yu and Tianxi Cai at Harvard T.H. Chan School of Public Health and Tianrun Cai at The Brigham and Women's Hospital. It is distributed free of charge for academic and non-commercial research use by the President and Fellows of Harvard College.

Clinical note parsing with NILE

The final NILE result looks like

A negative mention of C0037369 (hence not counted for downstream analysis)

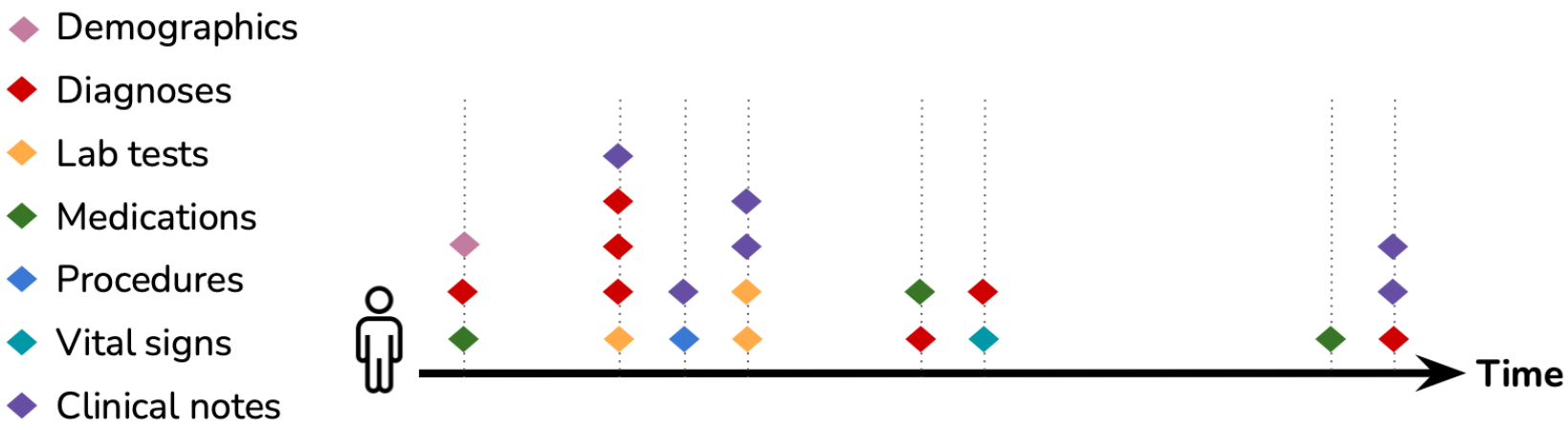
```
Patient_num|Record_ID|date|category|cuis
220|01|2067-05-03|C0262926:Y;C1522704:Y;C0037369:N;C0453996:N;C1881674:N;C002
0538:Y;C1963138:Y;C0037369:Y;C0453996:Y;C1881674:Y;C0020538:Y;C0262926:Y;C001
0054:Y;C0010068:Y;C1956346:Y;C0011849:Y;C0010054:Y;C0010068:Y;C1956346:Y;C003
8257:Y;C0013227:Y;C0070166:Y;C0004057:Y;C0004057:Y;C0017887:Y;C1271104:Y;C127
2641:Y;C0038435:Y;C0010054:Y;C0010068:Y;C1956346:Y;C1522704:Y
220|02|2068-12-05|C0262926:Y;C0028778:Y;C1279889:N;C1457868:N;C1457887:N;C002
0538:Y;C1963138:Y;C0037369:Y;C0453996:Y;C1881674:Y;C0027051:N;C0020538:Y;C026
2926:Y;C0010054:Y;C0010068:Y;C1956346:Y;C0011849:Y;C0022116:Y;C0010054:Y;C001
0068:Y;C1956346:Y;C0038257:Y;C0013227:Y;C0004057:Y;C0004057:Y;C0017887:Y;C007
0166:Y;C0262926:N;C0010054:N;C0010068:N;C1956346:N;C0262926:Y;C1271104:Y;C127
2641:Y
```

A positive mention of C0010054

Other common clinical NLP software

- cTAKES
- MetaMAP
- medspacy
- ...

Key idea



$$\text{phenotype} \approx \text{structured} + \text{unstructured data}$$

Summary

- EHR data is complex in structure
- EHRs do not have explicit information on phenotypes
- We need to aggregate the information in a patient's record to infer a phenotype

Note: Phenotyping is a team sport

- Clinicians
- Biostatisticians
- Informatics experts

Part III
A brief history of phenotyping

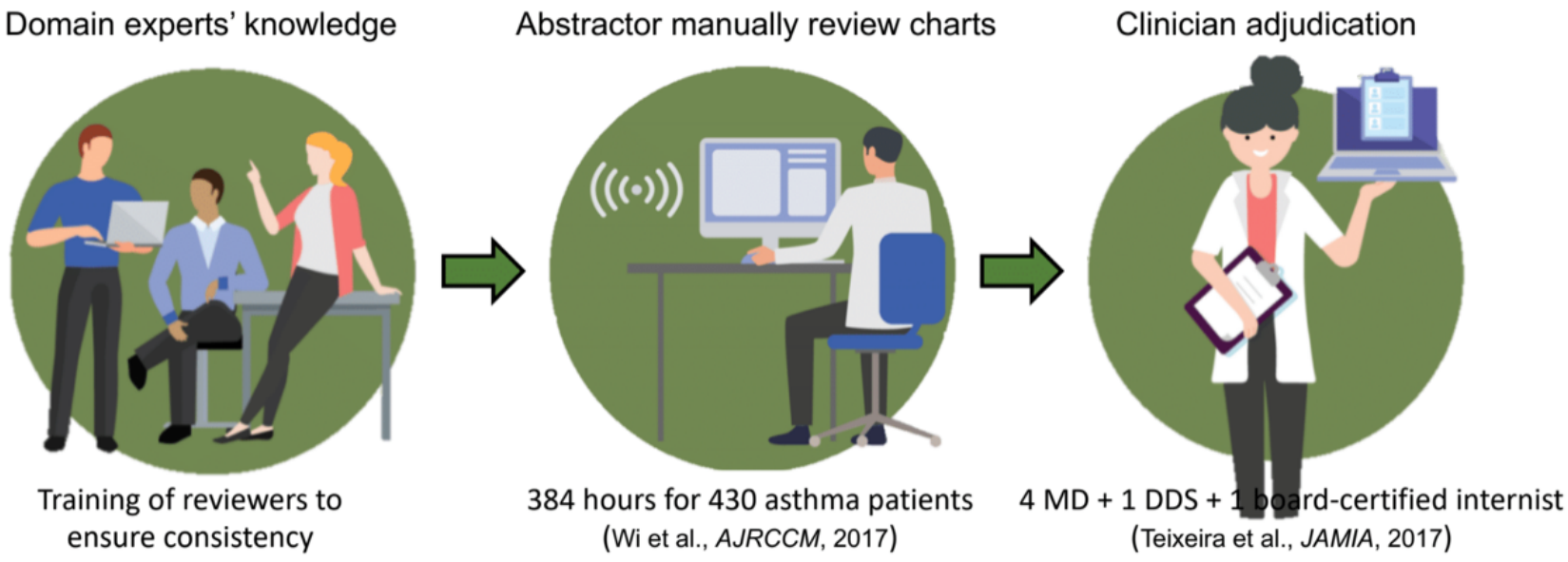
Three approaches to phenotyping

- Gold standard
  - ⋆ Manual chart review
- Phenotyping algorithms
  - ⋆ Rule-based
  - ⋆ Statistical/machine learning (ML) methods

Gold-standard: Manual chart review

"Gold standard labels"

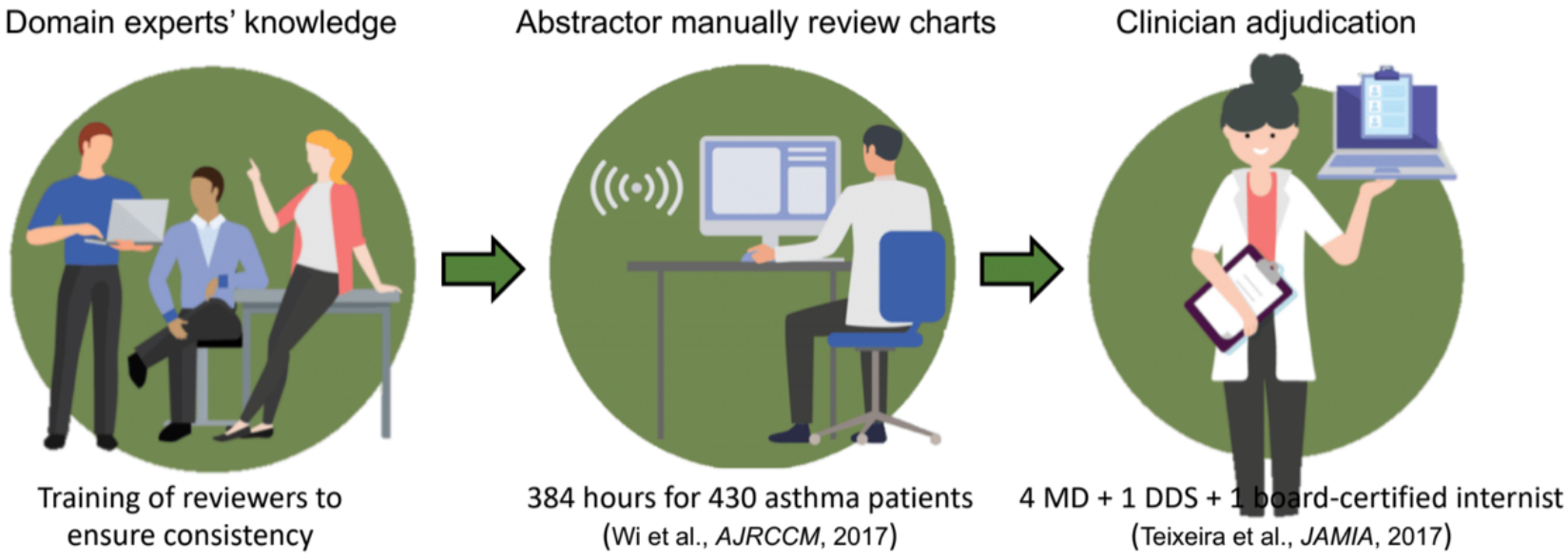
The best classification available for phenotype status derived from manual review of charts



Gold-standard: Manual chart review

"Gold standard labels"

The best classification available for phenotype status derived from manual review of charts



Time and resource intensive

Motivation for phenotyping algorithms

- Infeasibility of obtaining the gold-standard label
- Streamlining the development of disease registries
- Enabling phenotyping across healthcare settings
- Integration into the EHR system for clinical decision support

The ideal phenotyping algorithm

- **Accurate**: Can precisely identify the phenotype
- **Efficient**: Can be developed quickly
- **Portable**: Can be implemented across healthcare settings

Early phenotyping algorithms: Rule-based

- Most literature relies on "clinical decision rules"
- Based on knowledge of the phenotype and its documentation
  - ⋆ Simple (eg. height) or complex (eg. depression)
  - ⋆ May include one data element or many
  - ⋆ May include a time component
  - ⋆ May incorporate structured and unstructured data

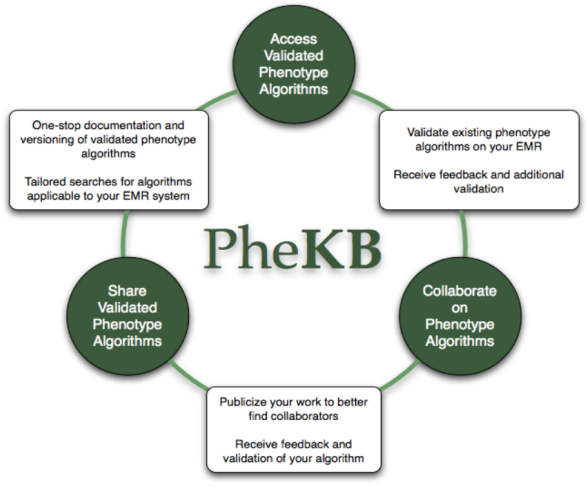
Example: PheKB of EMERGE

PheKB a knowledgebase for discovering phenotypes from electronic medical records

Home | Phenotypes | Resources | Contact Us

Click here to email us.

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

Example: PheKB of EMERGE

Title	Institution	Data Modalities and Methods Used	Owner Phenotyping Groups	View Groups	Has new content	Status	Type
Abdominal Aortic Aneurysm (AAA)	Geisinger	CPT Codes, ICD 9 Codes, Vital Signs	eMERGE Geisinger Group	eMERGE Geisinger Group, eMERGE Phenotype WG		Final	Disease or Syndrome
ACE Inhibitor (ACE-I) induced cough	Vanderbilt University	CPT Codes, ICD 9 Codes, Medications, Natural Language Processing	eMERGE Vanderbilt Group	eMERGE Phenotype WG		Final	Drug Response - adverse effect or efficacy
ADHD phenotype algorithm	CHOP	ICD 9 Codes, Medications, Natural Language Processing	eMERGE CHOP Group	eMERGE Phenotype WG		Final	Disease or Syndrome
Anxiety algorithm	CHOP	CPT Codes, ICD 10 Codes, ICD 9 Codes, Medications	eMERGE CHOP Group	eMERGE CHOP Group, eMERGE Phenotype WG		Final	Disease or Syndrome
Appendicitis	Cincinnati Children's Hospital Medical Center	CPT Codes, ICD 9 Codes, Medications, Natural Language Processing	eMERGE CCHMC/BCH Group	eMERGE Phenotype WG		Final	Disease or Syndrome
Asthma Response to Inhaled Steroids			PGPop WG	PGPop WG		Final	Drug Response - adverse effect or efficacy

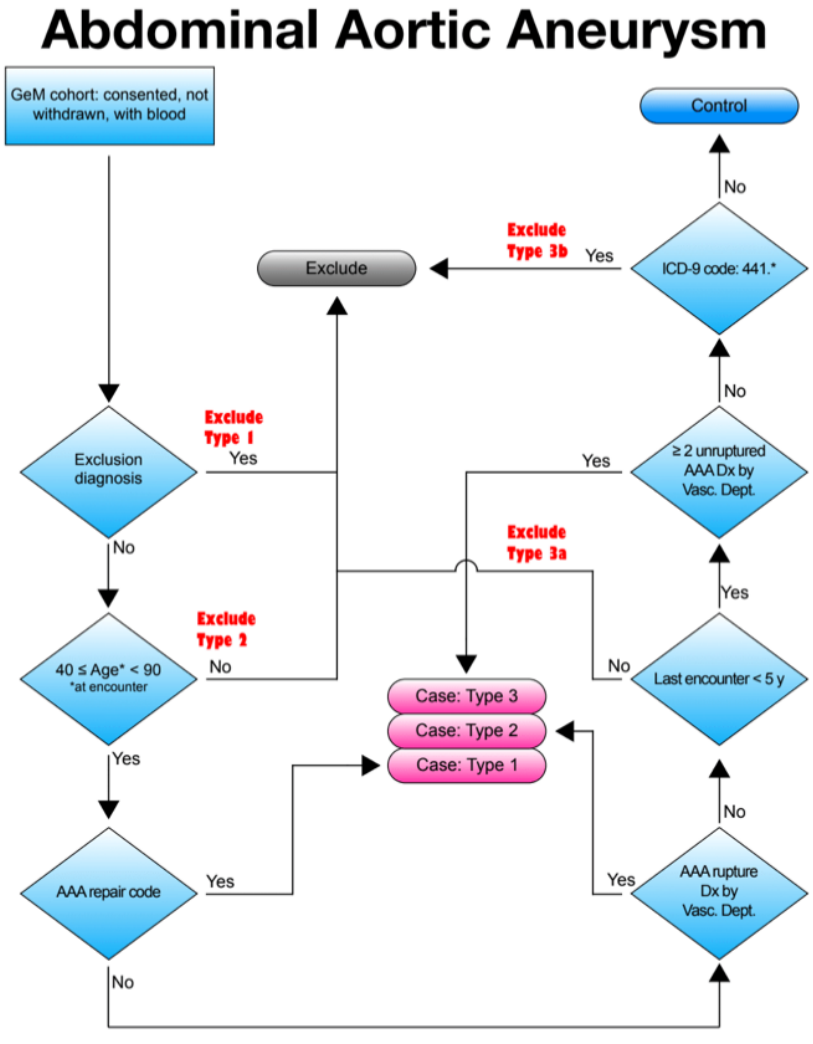
Currently contains 82 phenotyping algorithms

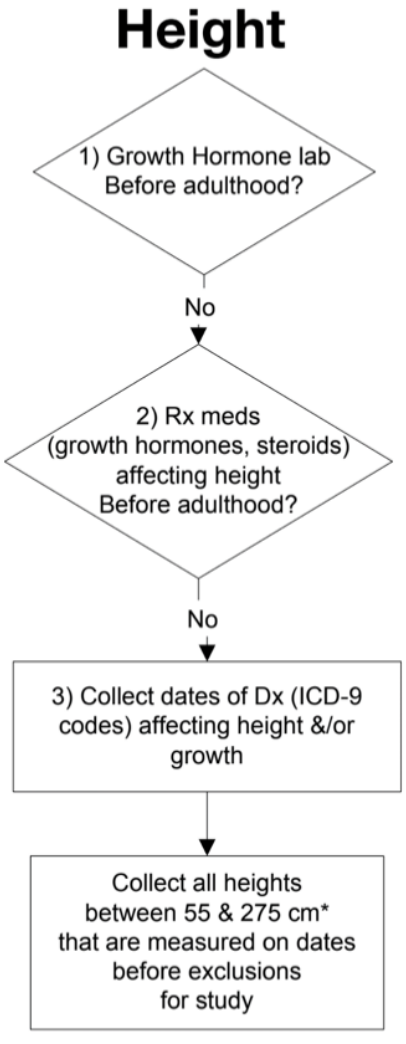
Example: CPCSSN Case definitions

CPCSSN Case Definitions

Version 2

Rules are often represented in a flow chart





Summary: Rule-based algorithms

Advantages

- Simple to use
- Physician can relate to their documentation practices

Disadvantages

- Relies on human expertise
- Hard to scale across phenotypes and institutions

Phenotyping algorithms: Machine learning (ML)

Learn the phenotyping algorithm from the data rather than human expertise

Phenotyping algorithms: Machine learning (ML)

PheCAP is a comprehensive protocol for ML-based phenotyping

nature protocols

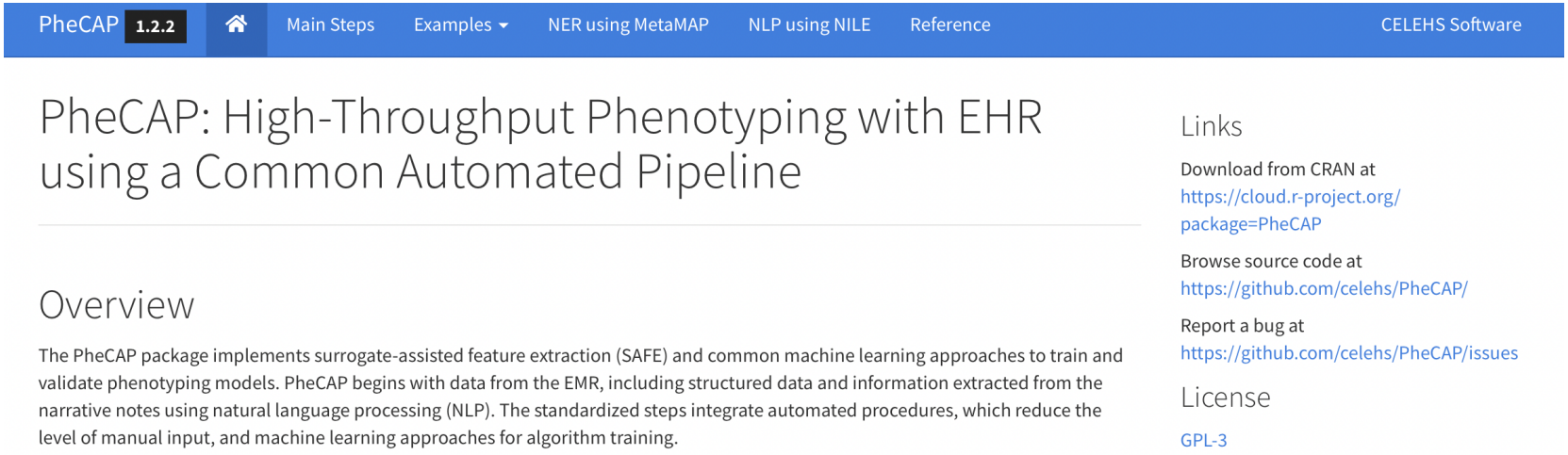
PROTOCOL

https://doi.org/10.1038/s41596-019-0227-6

High-throughput phenotyping with electronic medical record data using a common semi-supervised approach (PheCAP)

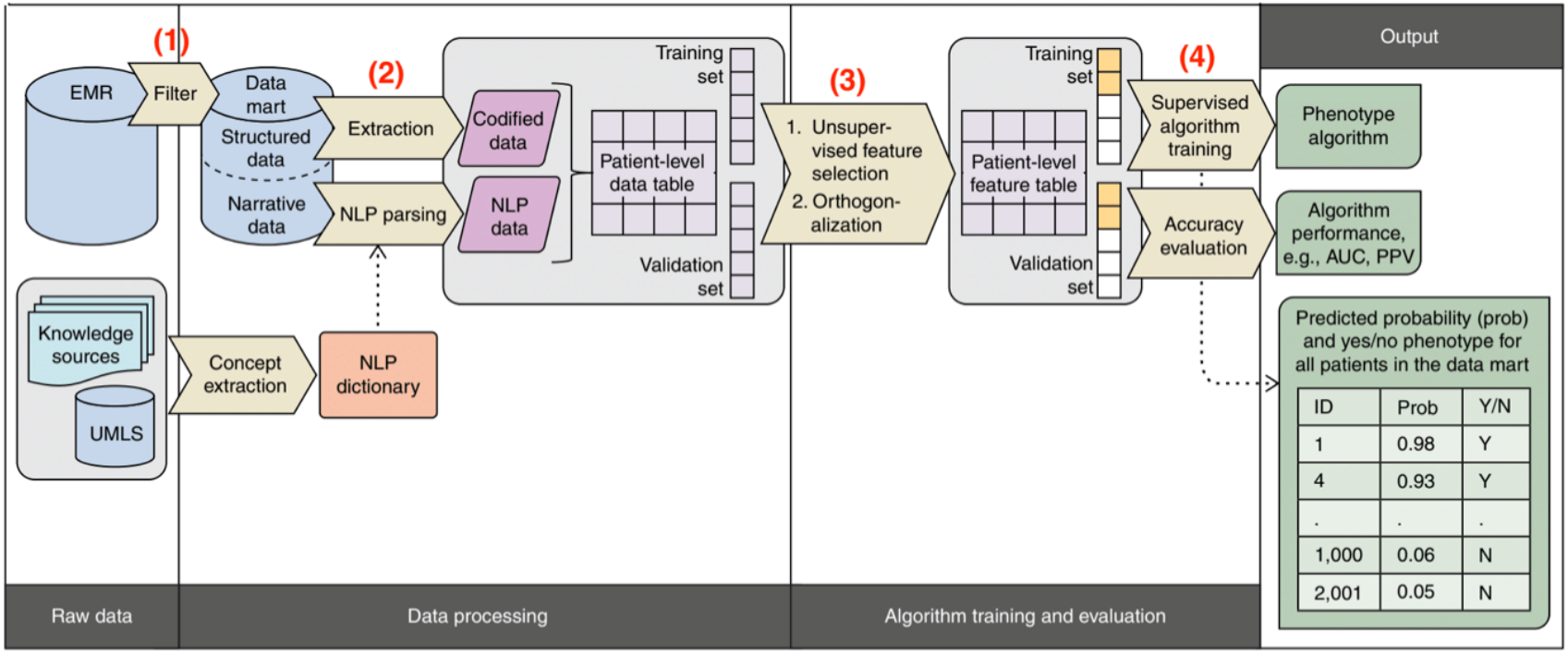
Yichi Zhang[1,18], Tianrun Cai[2,18], Sheng Yu[3,4,18], Kelly Cho[5,6], Chuan Hong[1], Jiehuan Sun[1], Jie Huang[2], Yuk-Lam Ho[5], Ashwin N. Ananthakrishnan[7], Zongqi Xia[8], Stanley Y. Shaw[9], Vivian Gainer[10], Victor Castro[10], Nicholas Link[5], Jacqueline Honerlaw[5], Sicong Huang[2], David Gagnon[5,16], Elizabeth W. Karlson[2], Robert M. Plenge[2], Peter Szolovits[11], Guergana Savova[12], Susanne Churchill[13], Christopher O'Donnell[5,14], Shawn N. Murphy[10,13,15], J. Michael Gaziano[5,6], Isaac Kohane[13], Tianxi Cai[1,13,17] and Katherine P. Liao[2,5,13,17]*

PheCAP R package



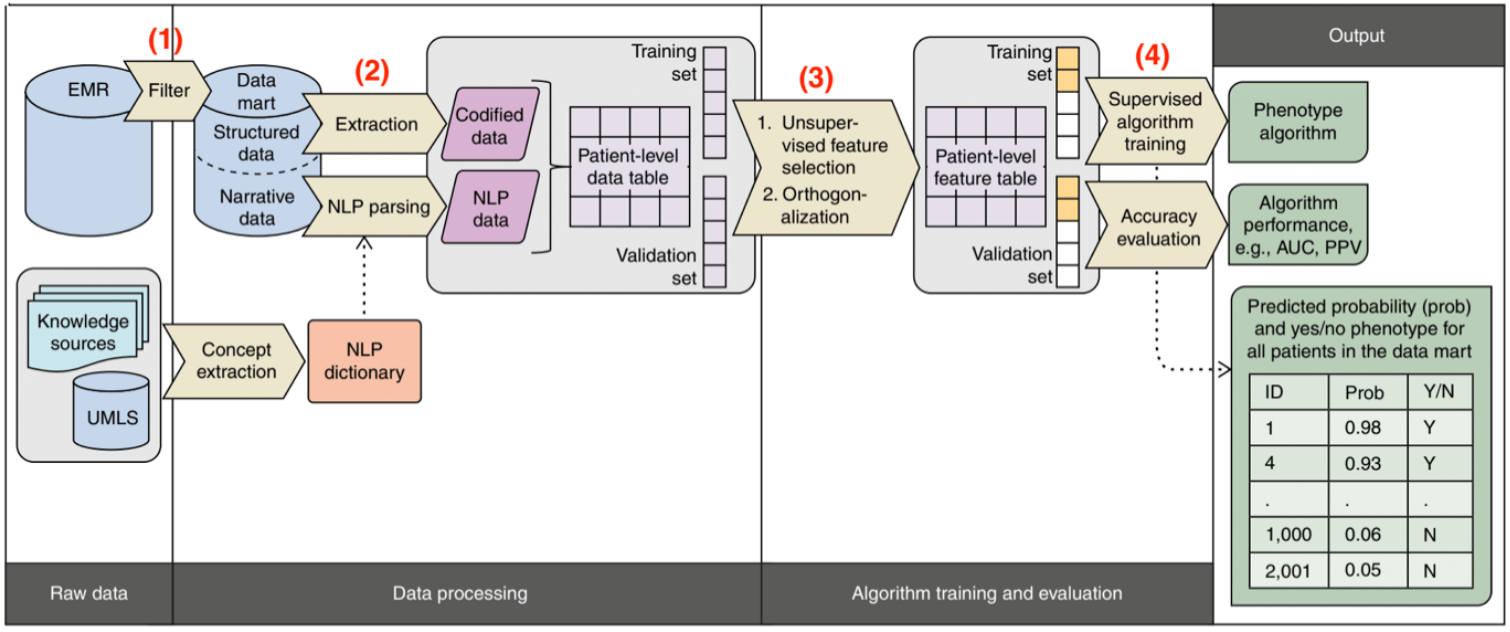
- Provides simulated and real EHR data
- Fitting functions for a variety of ML models
  - ⋆ Includes the specific PheCAP modeling approach
- Tutorial website can be found here

PheCAP pipeline



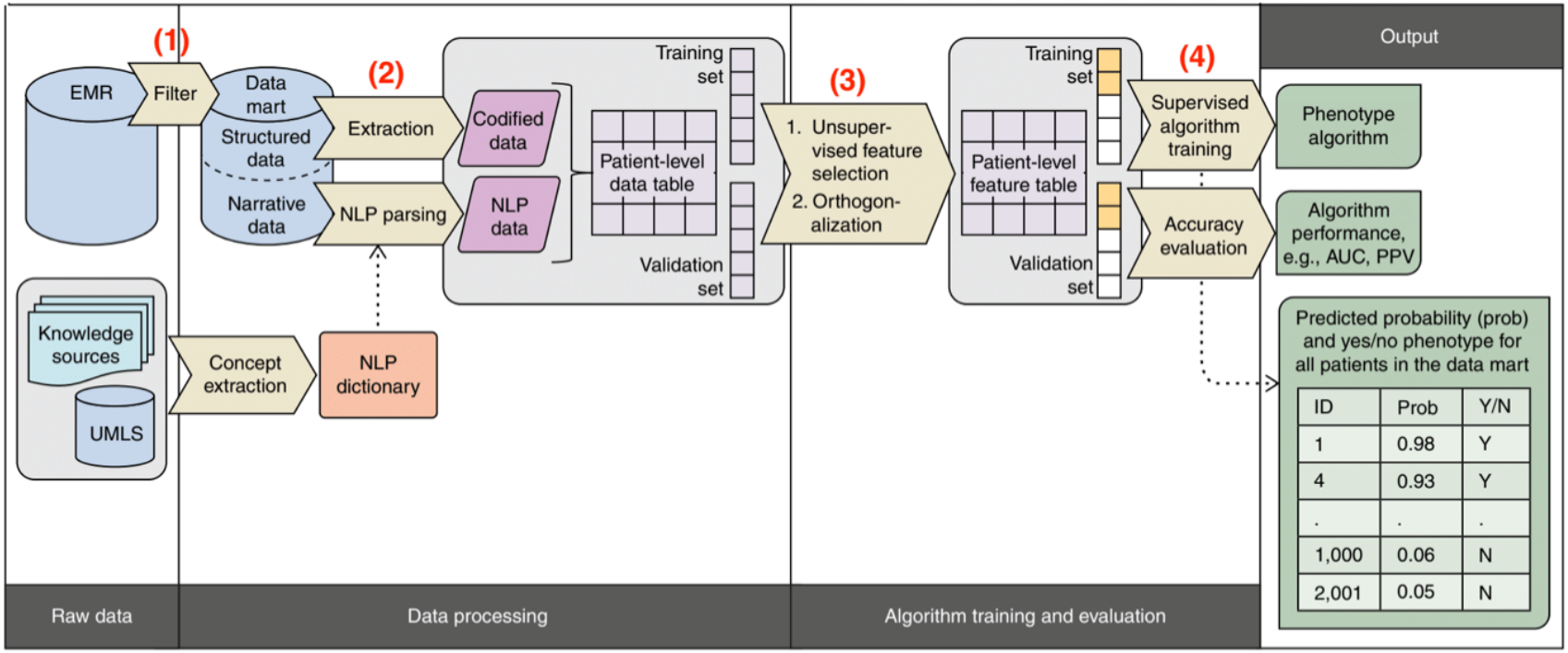
1. Filter to EHR data mart and chart review
2. Feature extraction
3. Algorithm training and evaluation
4. Output

PheCAP pipeline



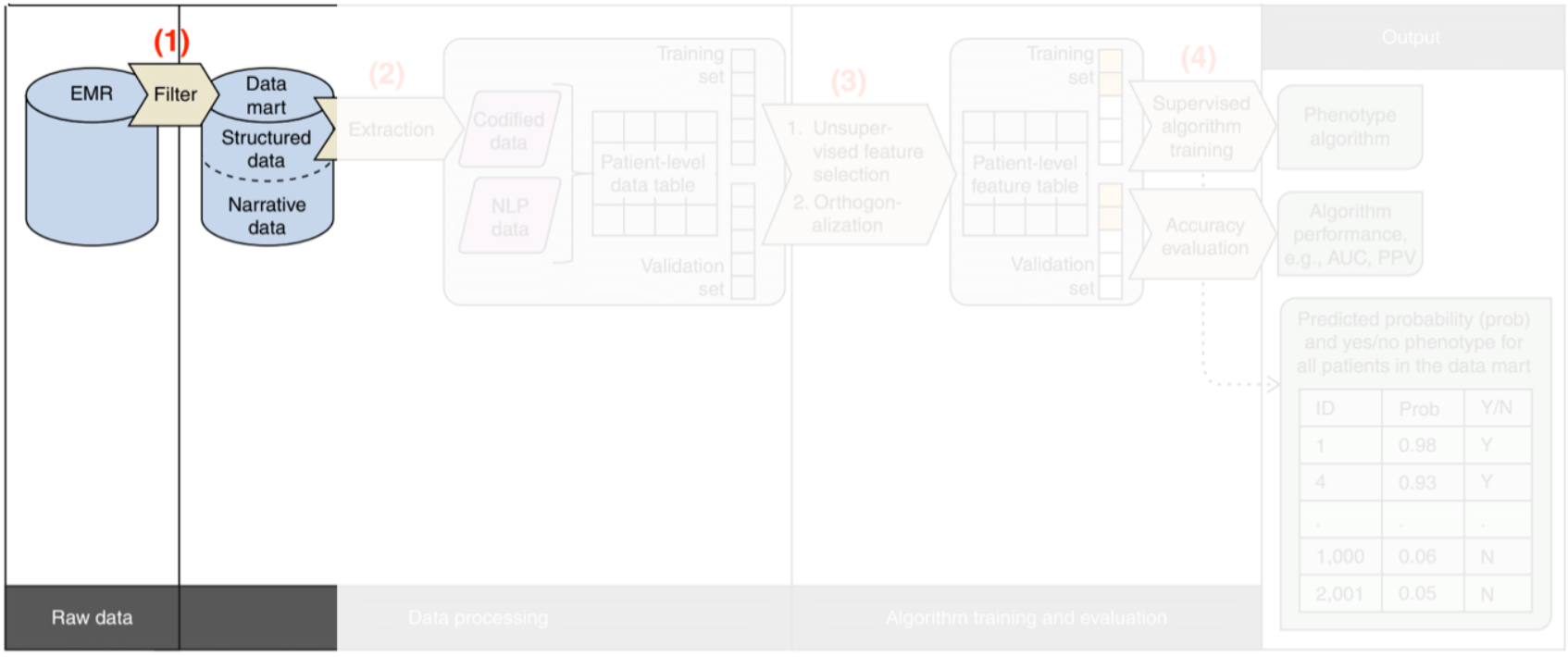
1. Filter to EHR data mart and chart review
2. Feature extraction
3. Algorithm training and evaluation
   - ⋆ Later: Semi-supervised PheCAP approach
4. Output

Steps 1 & 2: Prepare data for ML



1. **Filter to EHR data mart and chart review**
2. **Feature extraction**
3. Algorithm training and evaluation
4. Output

1. Filter to EHR data mart and chart review



1. Filter to EHR data mart

- Goal: Identify patients with some chance of having the disease and with sufficient data
- Prevalence filter
  - ⋆ e.g. ≥ 1 ICD code for the phenotype of interest
- Information filter
  - ⋆ e.g. ≥ 2 notes with > 500 characters each
- The **EHR data mart** contains patients who pass the filters

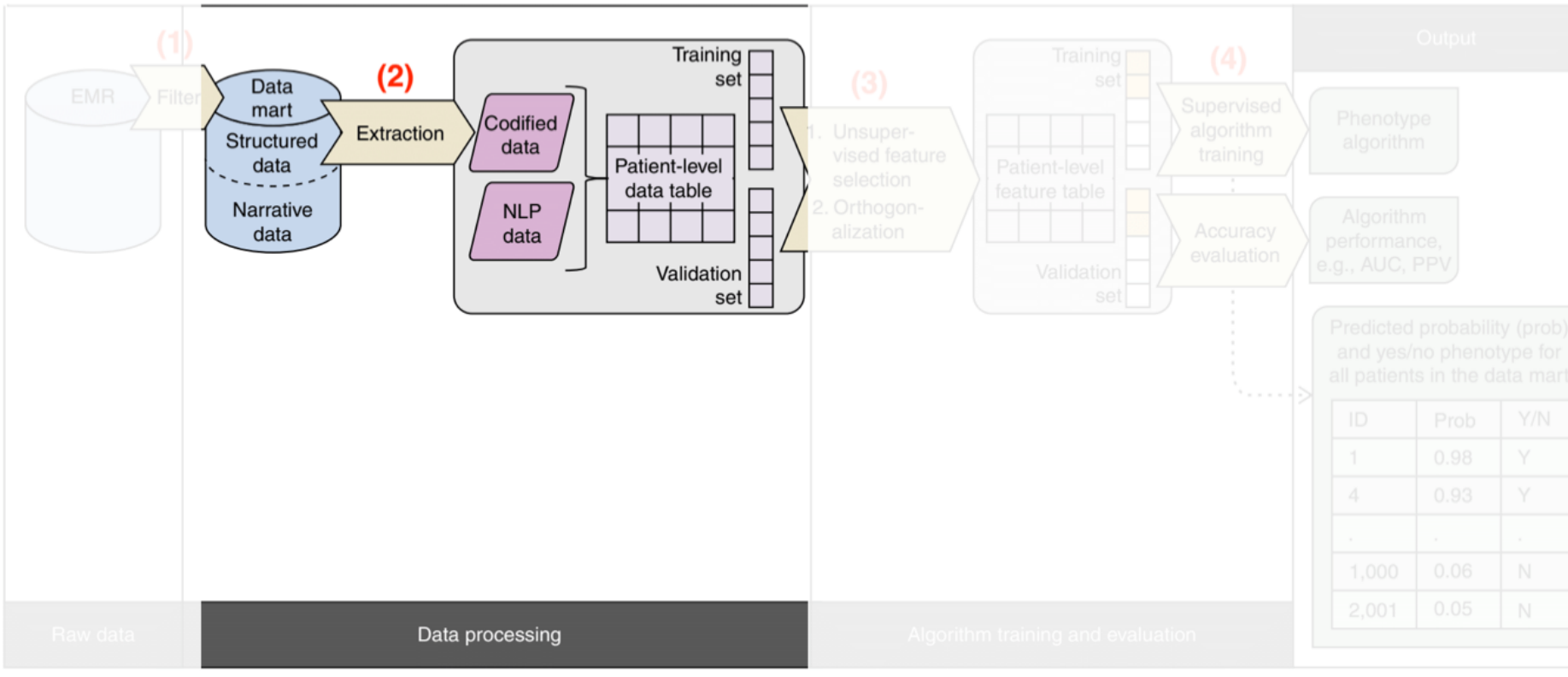
1. Chart review

- Sample patients from EHR data mart
  - ⋆ A random sample of approximately 200 subjects
  - ⋆ Often split into a training and test set
- Review charts of the sampled patients to obtain gold standard labels
  - ⋆ Develop a manual for reviewers
  - ⋆ Determine inter-rater reliability from multiple reviewers

1. Chart review

- Sample patients from EHR data mart
  - ⋆ A random sample of approximately 200 subjects
  - ⋆ Often split into a training and test set
- Review charts of the sampled patients to obtain gold standard labels
  - ⋆ Develop a manual for reviewers
  - ⋆ Determine inter-rater reliability from multiple reviewers
- VERY TIME CONSUMING
  - ⋆ Motivation for weakly & semi-supervised learning

2. Feature extraction



2. Feature extraction

- Many potential features are informative for the phenotype
- **PheCAP example**: Coronary artery disease (CAD)
  - ⋆ Structured data: ICD for myocardial infarction, CPT for cardiac catheterization, cholesterol levels, etc.
  - ⋆ Unstructured data: Progress notes, narrative information in a report about perfusion abnormalities on a cardiac stress test, etc.

2. Feature extraction

How do we determine which features to extract?

- Structured data
- Unstructured data
- Healthcare utilization feature

2. Feature extraction

How do we determine which features to extract?

- Structured data
  - ⋆ Domain experts or online resources $\rightarrow$ code counts
  - ⋆ eg. number of ICD codes for myocardial infarction
- Unstructured data
- Healthcare utilization feature

2. Feature extraction

How do we determine which features to extract?

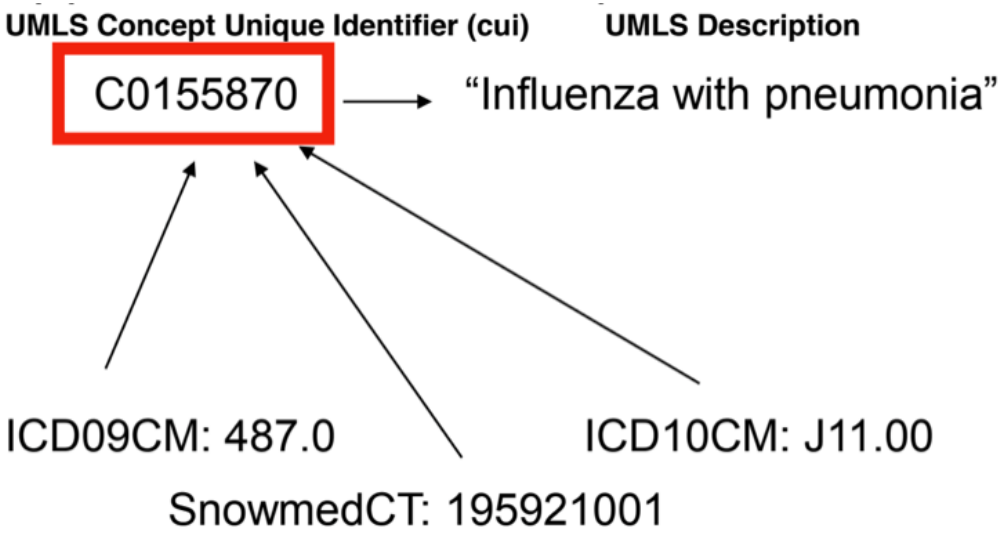
- Structured data
  - ⋆ Domain experts or online resources $\rightarrow$ code counts
  - ⋆ eg. number of ICD codes for CAD
- Unstructured data
  - ⋆ Natural language processing (NLP) $\rightarrow$ NLP counts
  - ⋆ eg. number of positive mentions of "CAD"
- Healthcare utilization feature

Extracting unstructured data

1. Generate a dictionary
   - ⋆ Identify concepts relevant to the phenotype of interest
2. Parse patients' clinical notes using NLP
   - ⋆ Identify positive mention of concepts in the dictionary
3. Obtain NLP data for algorithm development
   - ⋆ row = patient
   - ⋆ column = # positive mentions of a concept in a patient's clinical notes

Generating the NLP dictionary

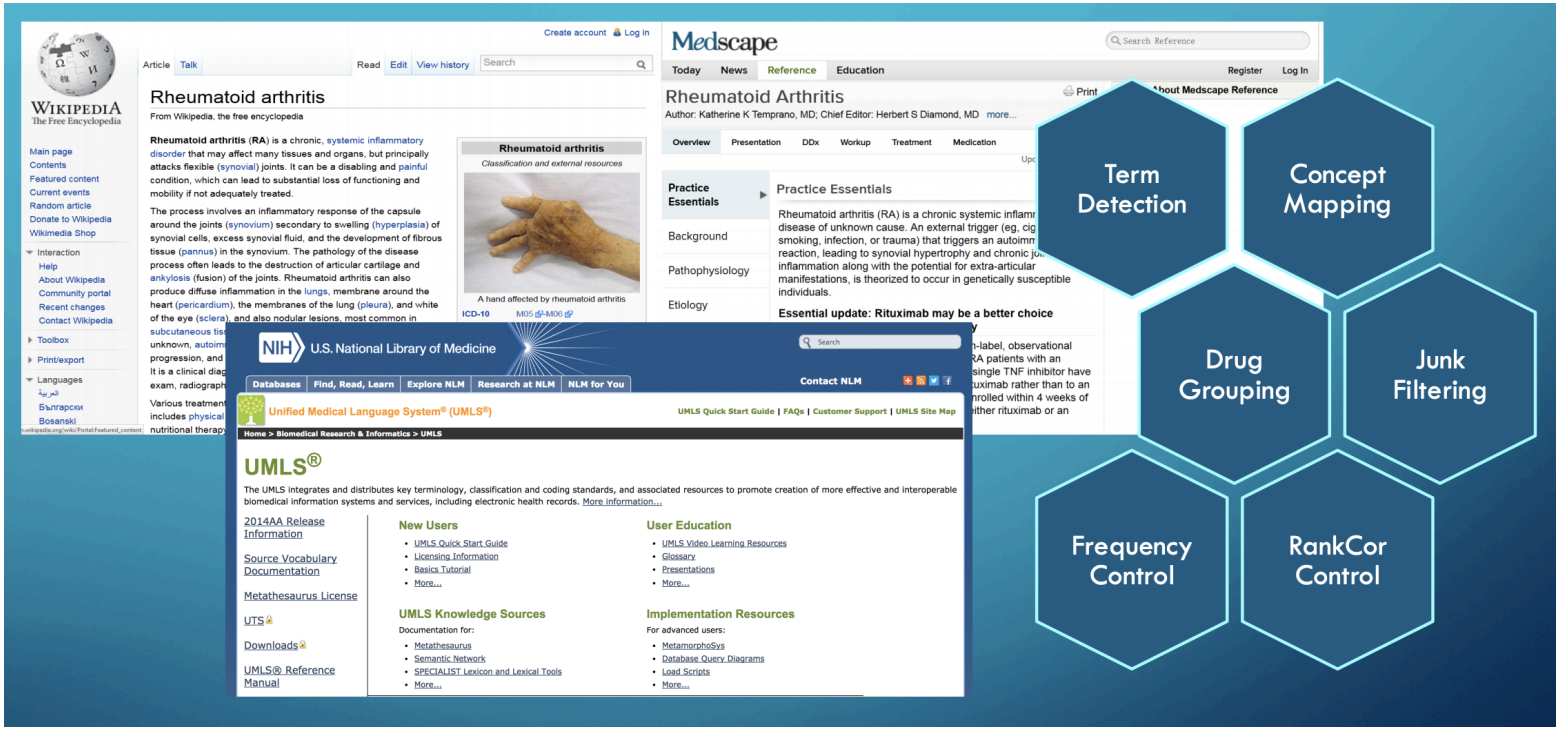
- Approach 1: Manual list
- Approach 2: Infer concepts from codes



Generating the NLP dictionary

- Approach 1: Manual list
- Approach 2: Infer concepts from codes
- Approach 3: Extract concepts from knowledge sources

Extract concepts from online knowledge sources



Yu et al 2016

Identify candidate concepts by parsing online knowledge sources such as Wikipedia, Medscape, and Merck Manuals

2. Feature extraction

How do we determine which features to extract?

- Structured data
  - ⋆ Domain experts or online resources $\rightarrow$ code counts
  - ⋆ eg. number of ICD codes for myocardial infarction
- Unstructured data
  - ⋆ Natural language processing (NLP) $\rightarrow$ NLP counts
  - ⋆ eg. number of positive mentions of "CAD"
- Healthcare utilization feature
  - ⋆ An important feature that significantly improves the prediction

Healthcare utilization feature

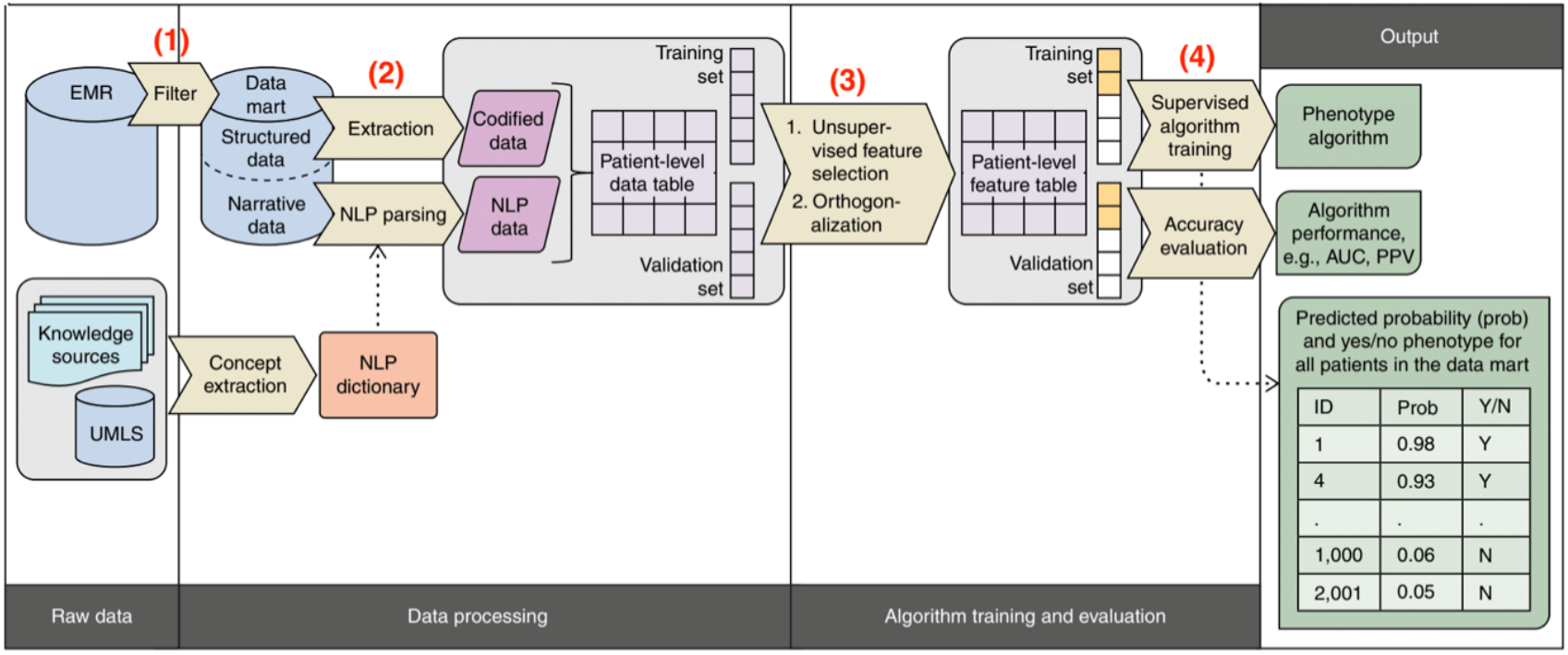
- Patients with more healthcare utilization have higher count of code/NLP counts regardless of their true underlying phenotype status
- A proxy for healthcare utilization
  - ⋆ Total number of unique billing codes, total number of visits, total number of notes, etc.

Healthcare utilization feature

The code & NLP counts are normalized by the healthcare utilization feature via:

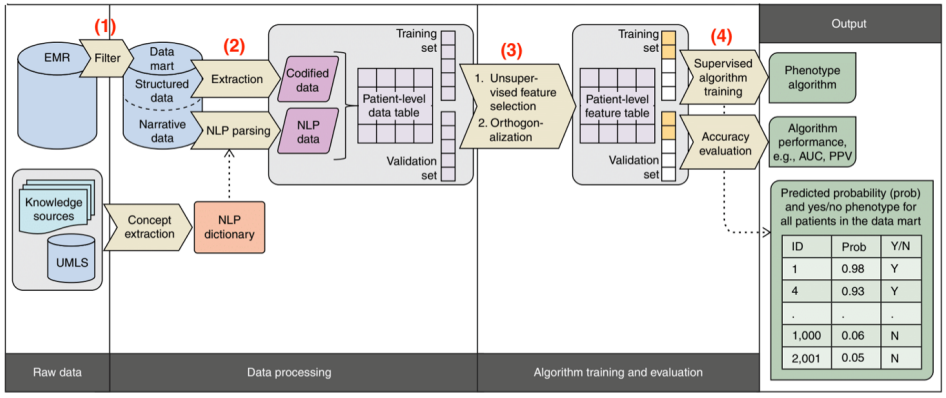
- Direct adjustment in the ML model
- Normalized frequency counts
  - ⋆ e.g. # CAD ICD codes/total # of ICD codes
- Orthogonalization
  - ⋆ Residual from the regression of a feature against the healthcare utilization
  - ⋆ Residual only provides information about the phenotype above and beyond healthcare utilization

Summary of Steps 1 & 2: Prepare data for ML



1. **Filter to EHR data mart and chart review**
2. **Feature extraction**
3. Algorithm training and evaluation
4. Output

Summary of Steps 1 & 2: Prepare data for ML



1. **Filter to EHR data mart and chart review**
   - ⋆ Obtain candidate patients and a subset of gold standard labels
2. **Feature extraction**
   - ⋆ Obtain code counts, NLP counts, and healthcare utilization feature

CAD Example: Data structure

Patient ID	CAD ICD	MI ICD	CAD NLP	Angina NLP	...	# Visits	CAD
00001	20	0	45	10	...	121	1
00002	2	5	2	0	...	5	1
00003	3	0	0	0	...	11	0
00004	10	0	11	0	...	41	0
00005	1	0	3	1	...	5	1
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
9996	4	6	1	1	...	15	NA
9997	1	0	0	0	...	19	NA
9998	41	0	23	4	...	65	NA
9999	42	8	56	22	...	11	NA
10000	3	0	1	1	...	30	NA

A few notes

- The data comes from Mass General Brigham (formerly Partners Healthcare)
- Variable names are not provided in the CAD data
- There are instructions on the website for generating the NLP Dictionary
- There are links to more datasets if you'd like to run NLP
- Another openly available dataset is MIMIC-IV

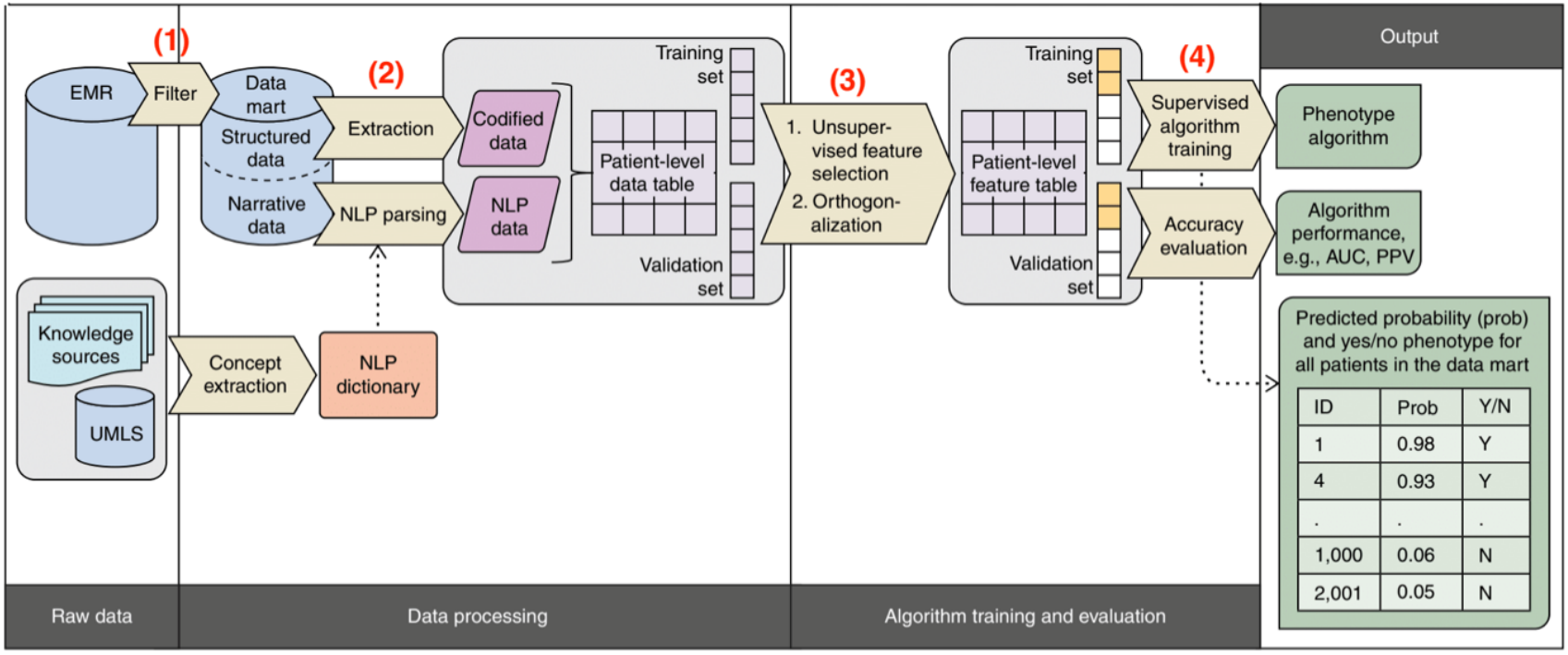
Module 1: Let's take a look at PheCAP!

Part IV
Supervised machine learning (ML) for phenotyping

Key idea of ML-based phenotyping

Learn the phenotyping algorithm from the data
rather than human expertise

PheCAP pipeline



1. Filter to EHR data mart and chart review
2. Feature extraction
3. **Algorithm training and evaluation**
4. Output

Flavors of ML

Supervised learning

- Learn a relationship between the phenotype (Y) and the features ($\boldsymbol{X}$)
- Requires **labeled data**: $\{(y_1, \boldsymbol{x}_1^T)^T, \ldots, (y_n, \boldsymbol{x}_n^T)^T\}$

Unsupervised learning

- Understand relationships among the features ($\boldsymbol{X}$)
- Requires **unlabeled data**: $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$

Focus of this section

Supervised learning

- Learn a relationship between the phenotype (Y) and the features ($\boldsymbol{X}$)
- Requires **labeled data**: $\{(y_1, \boldsymbol{x}_1^T)^T, \ldots, (y_n, \boldsymbol{x}_n^T)^T\}$

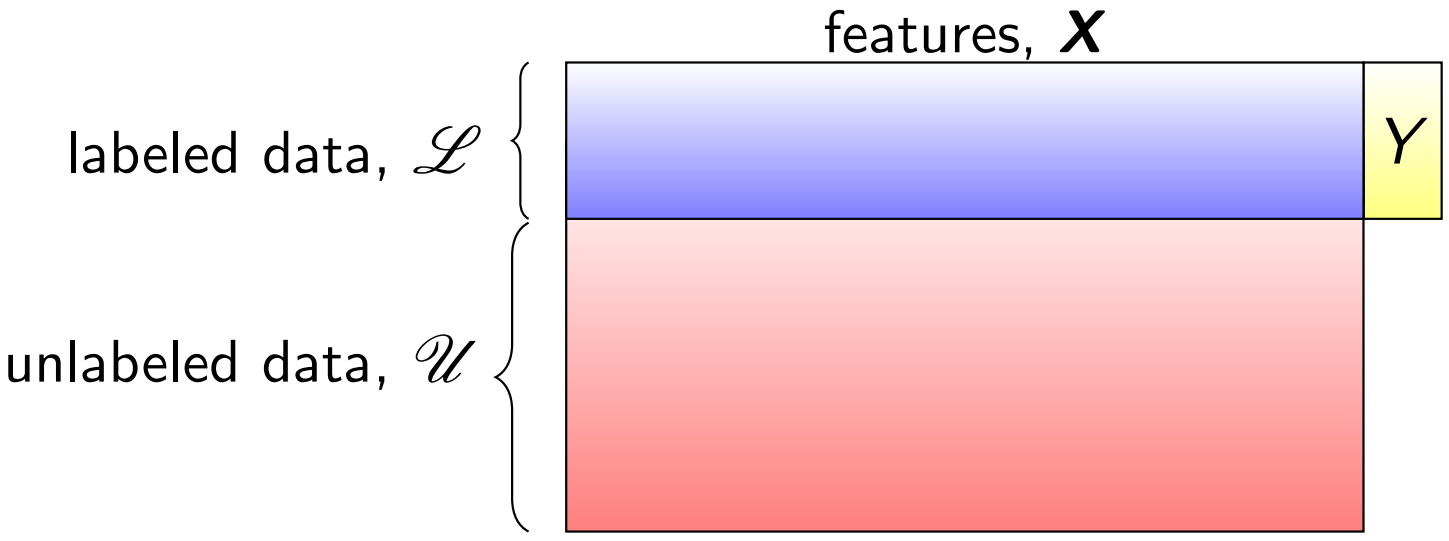
Focus of this section

Supervised learning

- Learn a relationship between the phenotype (Y) and the features ($\boldsymbol{X}$)
- Requires **labeled data**: $\{(y_1, \boldsymbol{x}_1^T)^T, \ldots, (y_n, \boldsymbol{x}_n^T)^T\}$

$\rightarrow$ We will get to weakly and semi-supervised learning later!

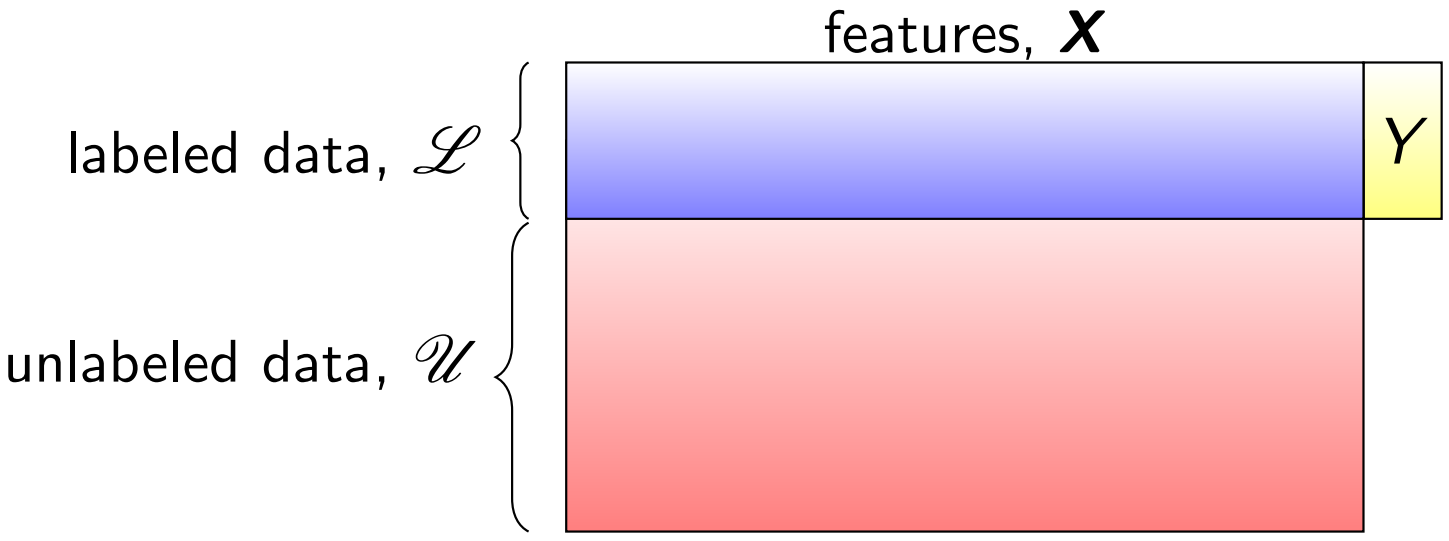
Formalizing our problem set-up



- Y: The binary phenotype from chart review
- $\boldsymbol{X}$: The $p \times 1$ feature vector

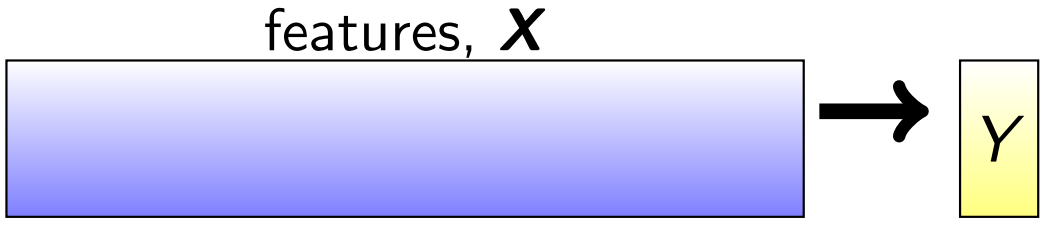
- Labeled data: $\mathscr{L} = \{(y_i, \boldsymbol{x}_i^\top)^\top\}_{i=1}^{n}$
- Unlabeled data: $\mathscr{U} = \{\boldsymbol{x}_i\}_{i=n+1}^{n+N}$

Formalizing the problem set-up



Use $\mathscr{L}$ to train a supervised ML model to predict Y in $\mathscr{U}$

Goal of supervised learning



Learn a function that maps X to Y

How do we learn the function?

- Logistic regression
- Support vector machine (SVM)
- Random forest
- Generalized additive model
- Gradient boosting machine
- Neural network
- ...

How do we learn the function?

- **Logistic regression**
- **Support vector machine** (SVM)
- **Random forest**
- Generalized additive model
- Gradient boosting machine
- Neural network
- ...

Our friend logistic regression

$$P(Y = 1|\boldsymbol{X}) = g(\alpha_0 + \boldsymbol{\beta}_0^\mathsf{T} X) \quad \text{where } g(x) = \frac{1}{1+e^{-x}}$$

- Outputs probabilities of Y given $X = x$
- The final classifications are derived from the probabilities
  - ⋆ i.e $\hat{Y} = I\{g(\hat{\alpha} + \hat{\boldsymbol{\beta}}^\mathsf{T} x) \geq c\}$
- Parameters estimated via maximum likelihood, i.e.

$$(\hat{\alpha}, \hat{\boldsymbol{\beta}}) = \arg\min_{\alpha,\beta} \sum_{i=1}^{n} \ell(y_i, \alpha + \boldsymbol{\beta}^\mathsf{T} x_i)$$

where $\ell(y, \eta) = -y\eta + \log(1 + e^\eta)$ is the negative log-likelihood function

In praise of logistic regression

- Interpretable
- Simple to fit
- Never the correct model, but performs well
- Easy to extend to more complex settings

But wait...

- For the CAD data, $p > n$
- We cannot fit a standard logistic regression in this case
- Regularization must be added

Ridge regression

- Fits the model subject to the constraint

$$\sum_{j=1}^{p} \beta_j^2 \leq t$$

- Or, equivalently, by adding the penalty to the likelihood function

$$\sum_{i=1}^{n} \ell(y_i, \alpha + \boldsymbol{\beta}^{\mathsf{T}} x_i) + \lambda \sum_{j=1}^{p} \beta_j^2$$

- Shrinks coefficients toward zero

Lasso regression

- Fits the model subject to the constraint

$$\sum_{j=1}^{p} |\beta_j| \leq t$$

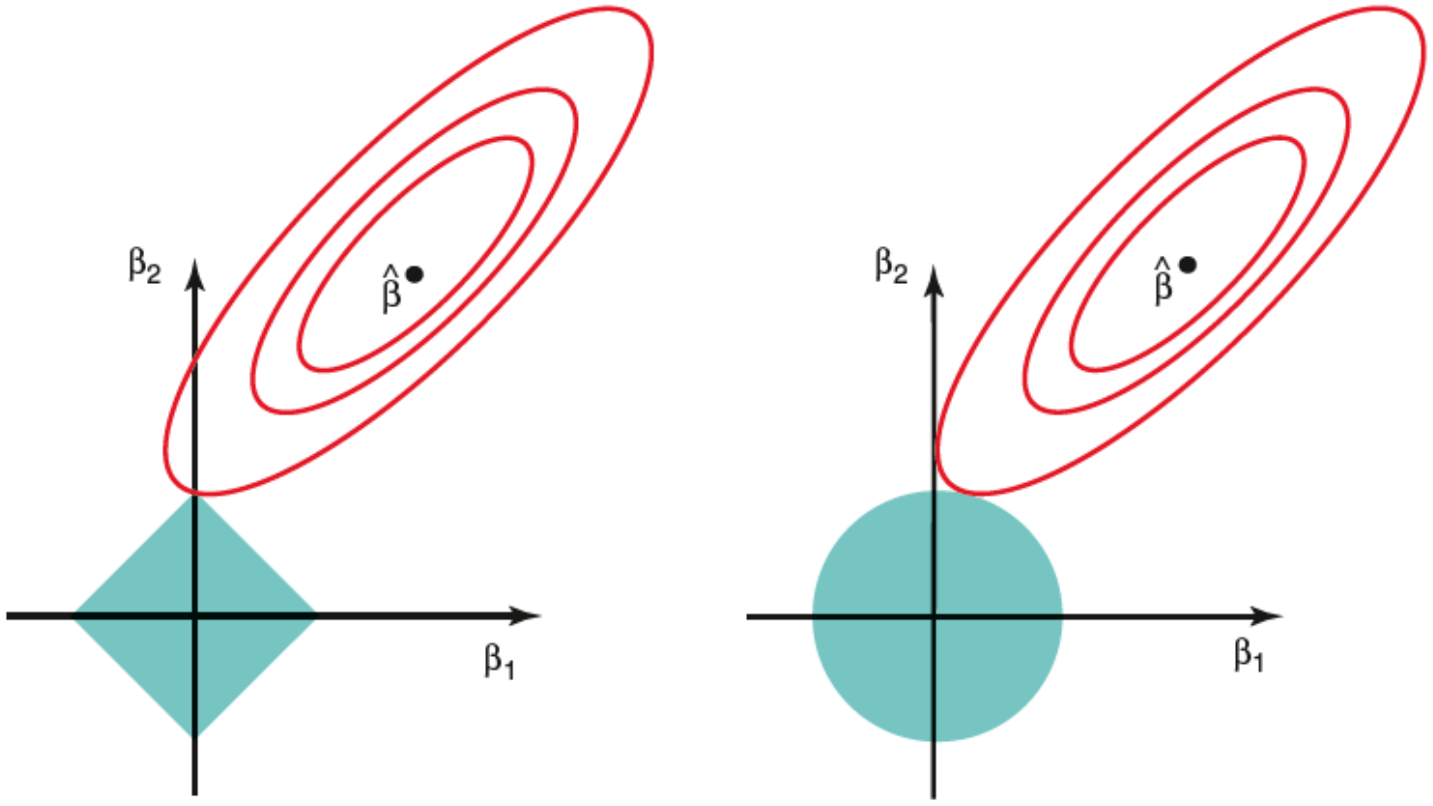
- Or, equivalently, by adding the penalty to the likelihood function

$$\sum_{i=1}^{n} \ell(y_i, \alpha + \boldsymbol{\beta}^{\mathsf{T}} x_i) + \lambda \sum_{j=1}^{p} |\beta_j|$$

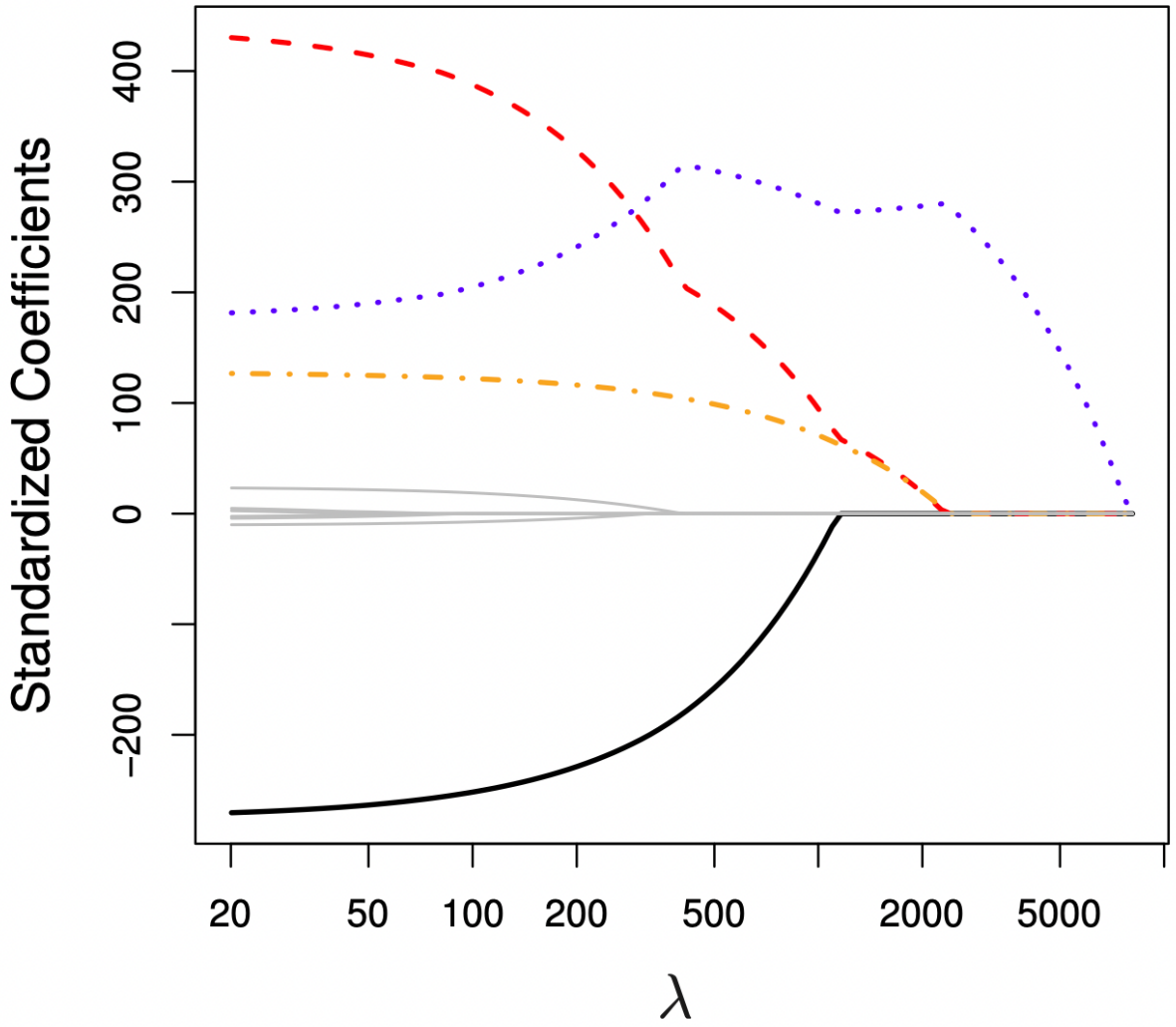
- Performs both shrinkage and variable selection

- A go-to for phenotyping

Lasso vs. Ridge



Lasso coefficient path



How to select λ?

Cross validation is a common choice

- Choose a grid of λ values
- Compute the cross-validation error for each value of λ
  - In `glmnet` the default is the deviance
- Select the value of λ for which the cross-validation error is smallest
- Refit the model using all of the available observations and the selected value of λ

Adaptive Lasso

- Lasso tends to overselects relevant features
- The adaptive lasso (ALASSO) addresses this issue through an augmentation of the penalty function as

$$\lambda \sum_{j=1}^{p} |\beta_j| / |\hat{\beta}_j^{init}|$$

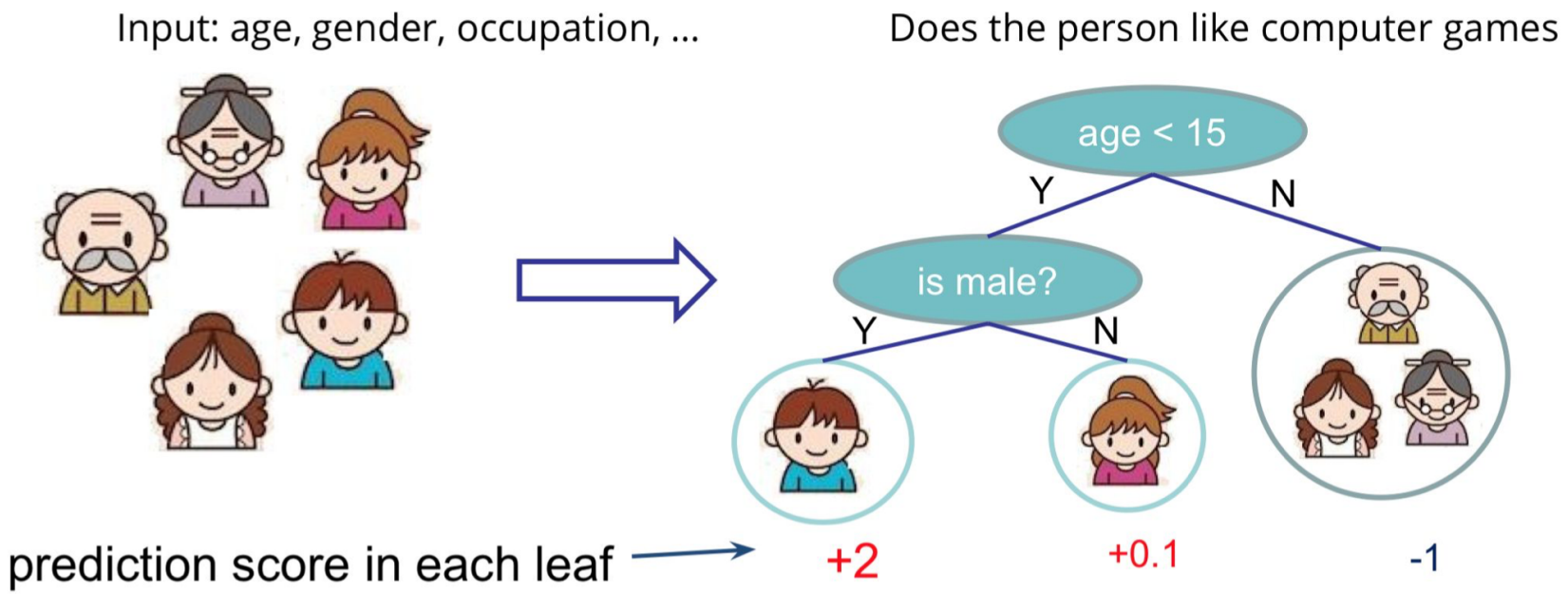
  where $\hat{\beta}_j^{init}$ is an initial estimate from an unpenalized or ridge regression

- An oracle procedure
  - ⋆ Identifies the right subset model
  - ⋆ Has the optimal estimation rate

Random forest

- Random forest is a decision tree-based method
- A decision tree learns a sequence of yes/no questions about the features to determine the phenotype status

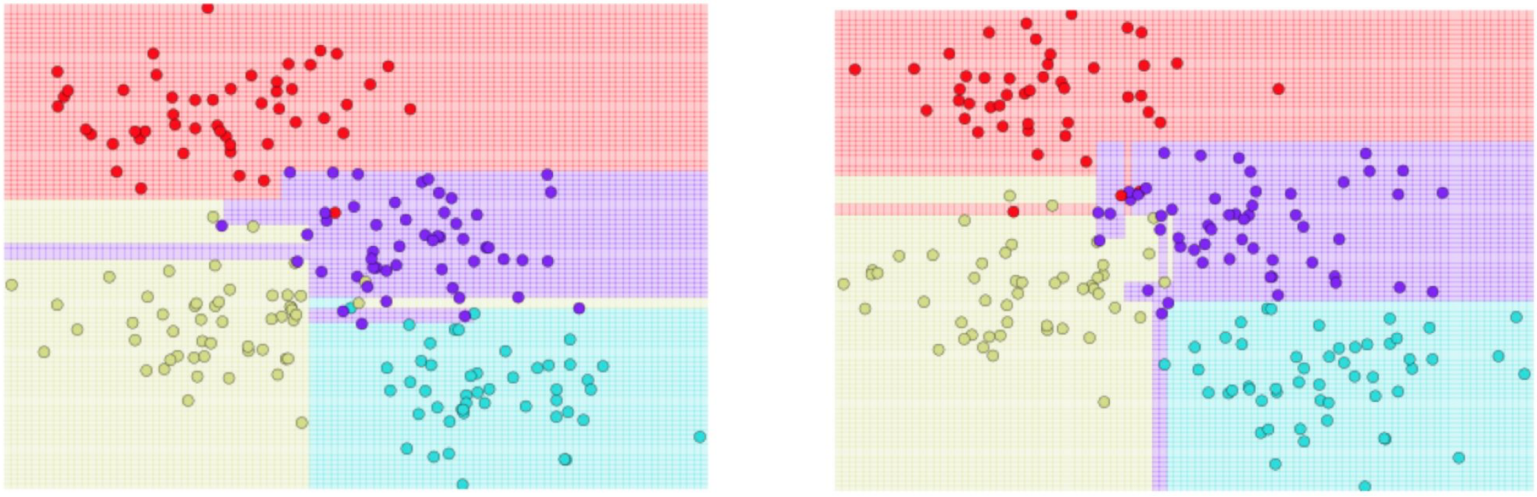
Decision tree example



CSC2515 slides

Decision tree example

Decision boundaries for two datasets



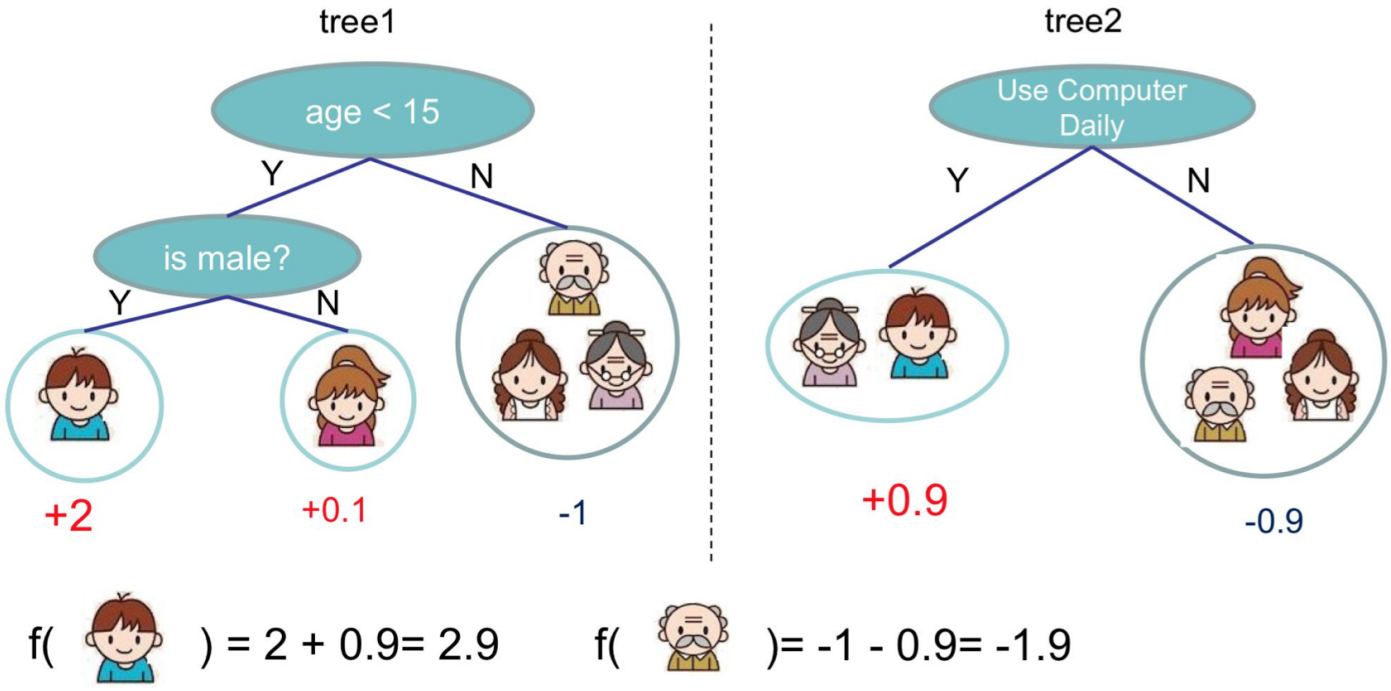
Often results in overfitting

Random forest

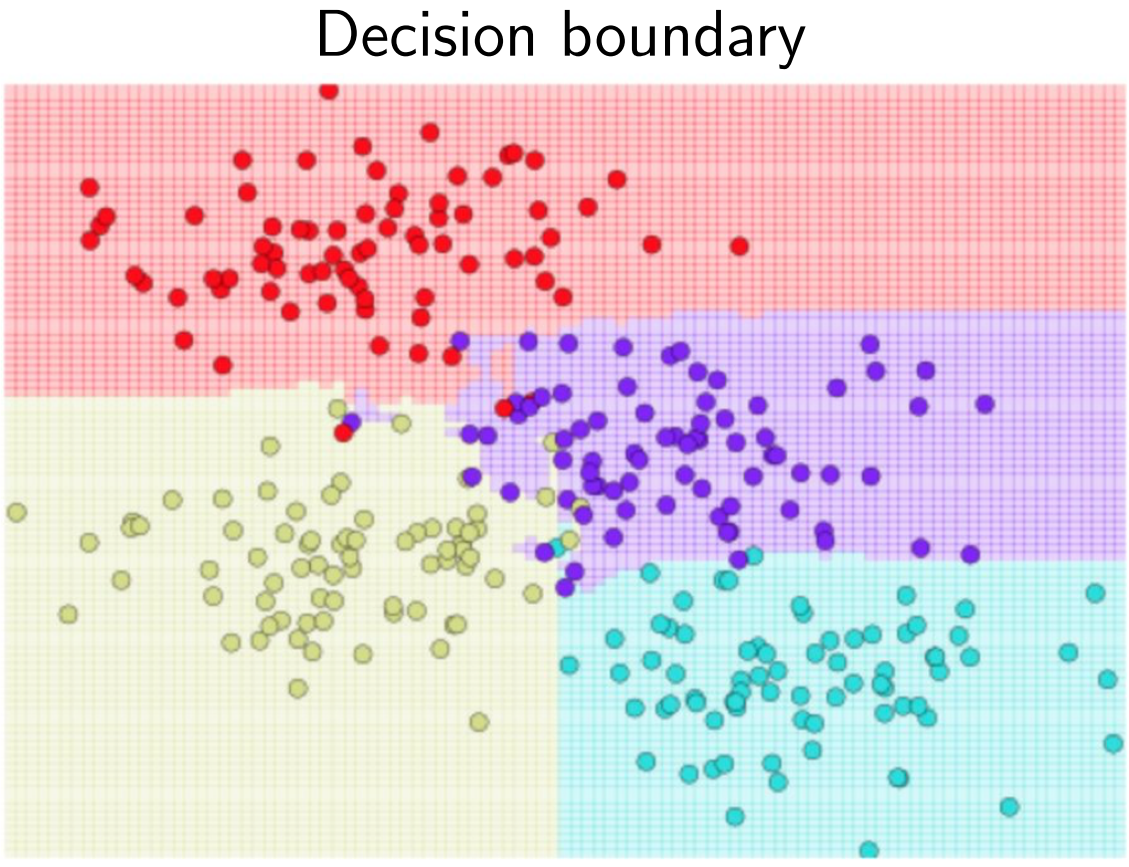
- Random forest is a decision tree-based method
- A random forest utilizes bagging to prevent overfitting
  - ⋆ Learns decision trees on splits of data called bags
  - ⋆ Different subset of variables are used at each tree split
  - ⋆ The final classification is based on all the trees

Random forest example

Average multiple decision trees

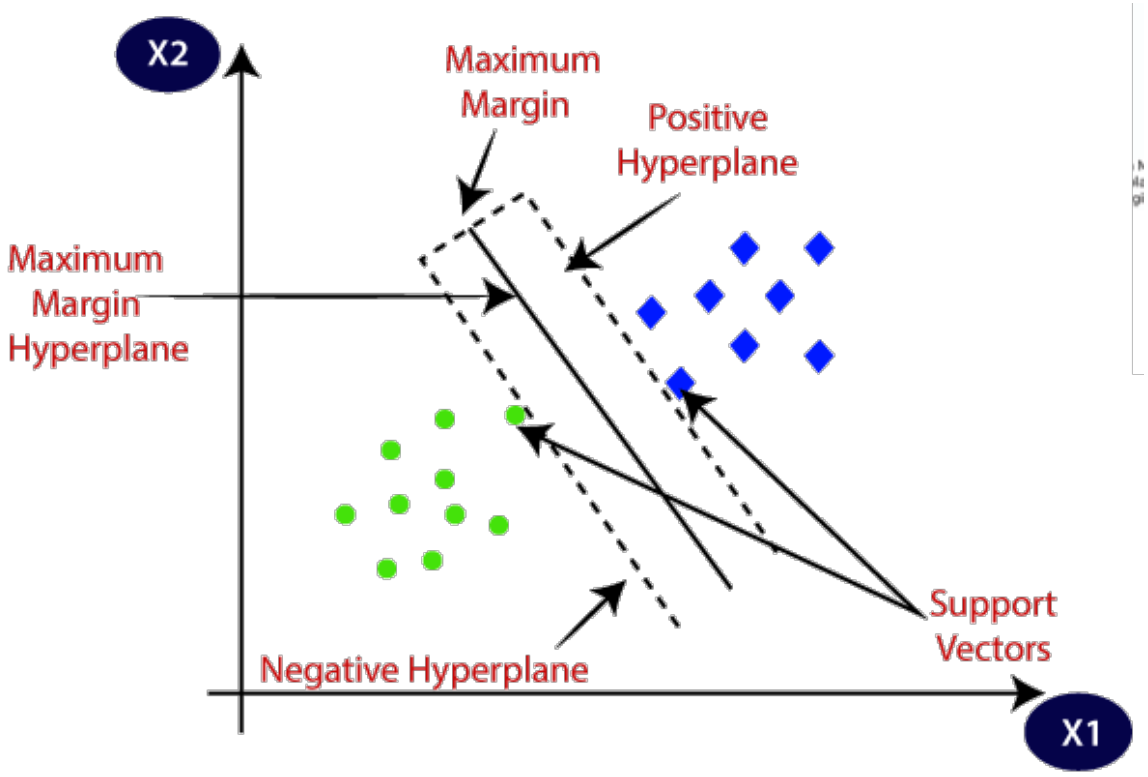


Random forest example

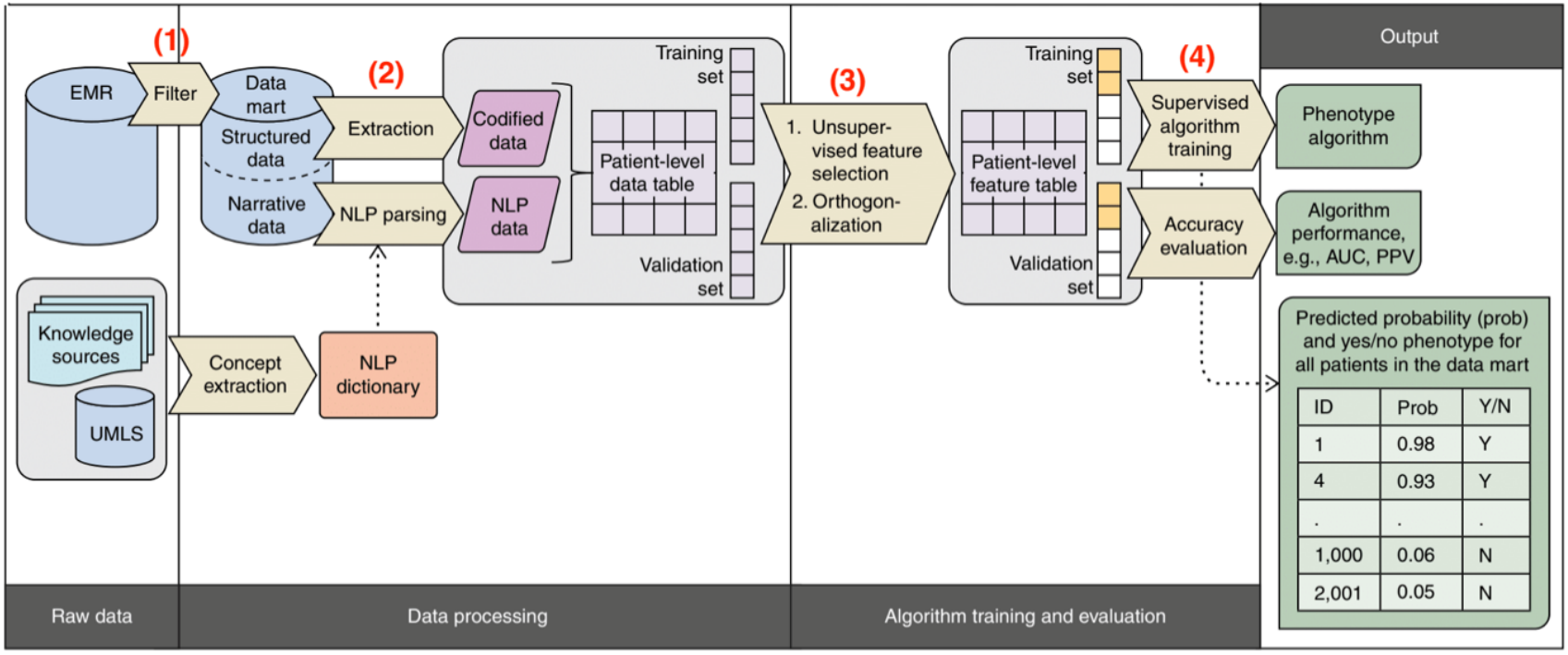


Support vector machine (SVM)

- Learns a hyperplane with the max margin between classes



PheCAP pipeline



1. Filter to EHR data mart and chart review
2. Feature extraction
3. **Algorithm training and evaluation**
4. Output

How do we evaluate our models?

- Absolute accuracy
- ROC analysis
- Calibration
- ...

How do we evaluate our models?

- Absolute accuracy
- **ROC analysis**
- Calibration
- ...

Defining prediction metrics

	Observed Positive	Observed Negative
Predicted Positive	True Positives (TP)	False Positives (FP)
Predicted Negative	False Negatives (FN)	True Negatives (TN)

Let P be all observed positives and N all the negatives.

- **Accuracy**: $(\mathrm{TP} + \mathrm{TN})/(\mathrm{P} + \mathrm{N})$

Defining prediction metrics

	Observed Positive	Observed Negative
Predicted Positive	True Positives (TP)	False Positives (FP)
Predicted Negative	False Negatives (FN)	True Negatives (TN)

- **Sensitivity**: $\text{TPR} = \text{TP}/\text{P} = P(\hat{Y} = 1 | Y = 1)$
  - ⋆ Also called **Recall**
- **Specificity**: $\text{TNR} = \text{TN}/\text{N} = P(\hat{Y} = 0 | Y = 0)$
  - ⋆ $\text{TNR} = 1 - \text{FPR} = 1 - \text{FP}/\text{N}$

Defining prediction metrics

	Observed Positive	Observed Negative
Predicted Positive	True Positives (TP)	False Positives (FP)
Predicted Negative	False Negatives (FN)	True Negatives (TN)

- **Positive predictive value** or **Precision**
  $PPV = \frac{TP}{TP+FP} = P(Y = 1|\hat{Y} = 1)$
- **Negative predictive value**
  $NPV = \frac{TN}{TN+FN} = P(Y = 0|\hat{Y} = 0)$

Defining prediction metrics

	Observed Positive	Observed Negative
Predicted Positive	True Positives (TP)	False Positives (FP)
Predicted Negative	False Negatives (FN)	True Negatives (TN)

- **F1**: 2 (PPV*TPR)/(PPV+TPR) = 2TP/(2TP+FP+FN)

Phenotyping model evaluation

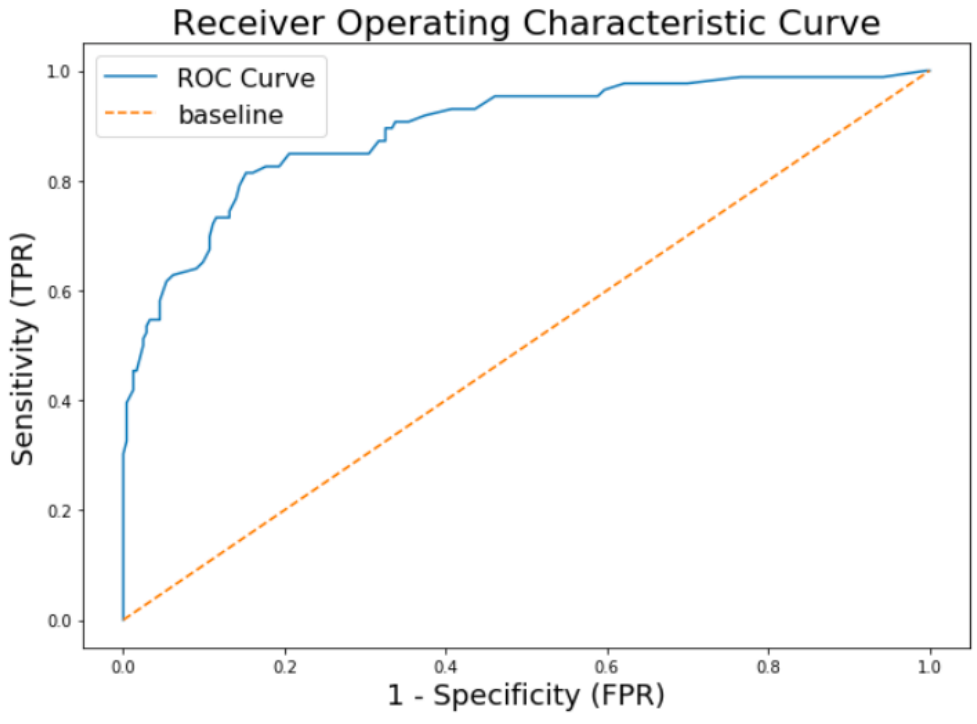
- Often aim to balance the TPR and PPV
- High TPR: Maximize capture of patients
  - ⋆ e.g. Enrollment into a clinical trial
- High PPV: Maximize accuracy of the capture
  - ⋆ e.g. GWAS conducted with EHR-linked biobank data

Receiver Operating Characteristic (ROC) Curve

- First used during World War II to analyze radar signals following the attack on Pearl Harbor 1941
- Plots the TPR (sensitivity/recall) vs. FPR (1 - specificity)
  - ⋆ Visual of the trade off between TPR and FPR across different thresholds for classification
- The most commonly used classifier evaluation criterion!

Computing the ROC curve

- Order probabilities from highest to lowest
- Start from the highest probability and draw a threshold at each point, each time checking TPR and FPR



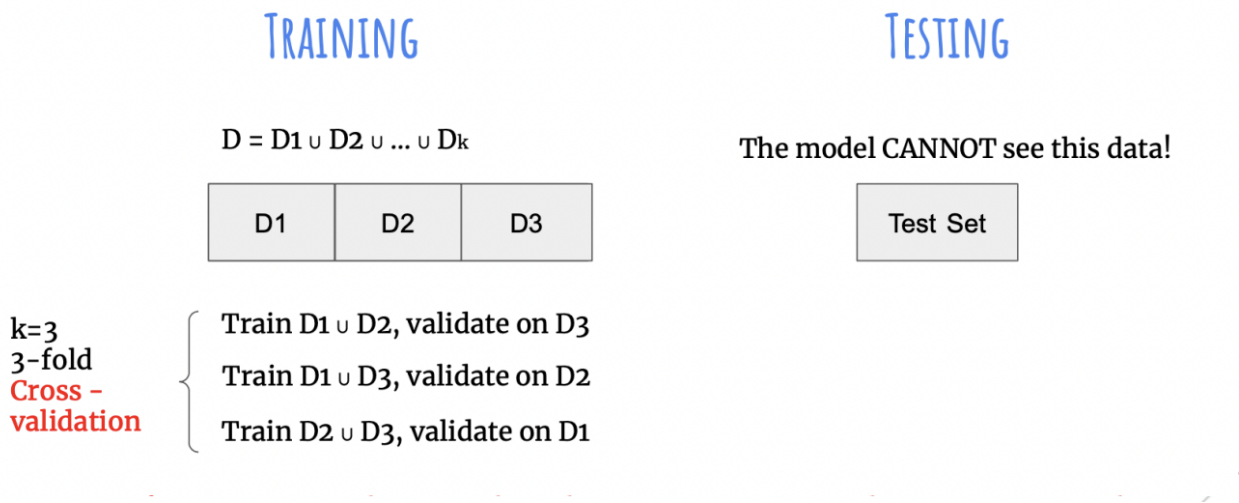
Animation

Train-test split

- Enables accurate evaluation of the prediction metrics
- Cross-validation is an alternative approach
  - ⋆ Favorable when the amount of labeled data is very small

Train-test split

- Enables accurate evaluation of the prediction metrics
- Cross-validation is an alternative approach
  - ⋆ Favorable when the amount of labeled data is very small

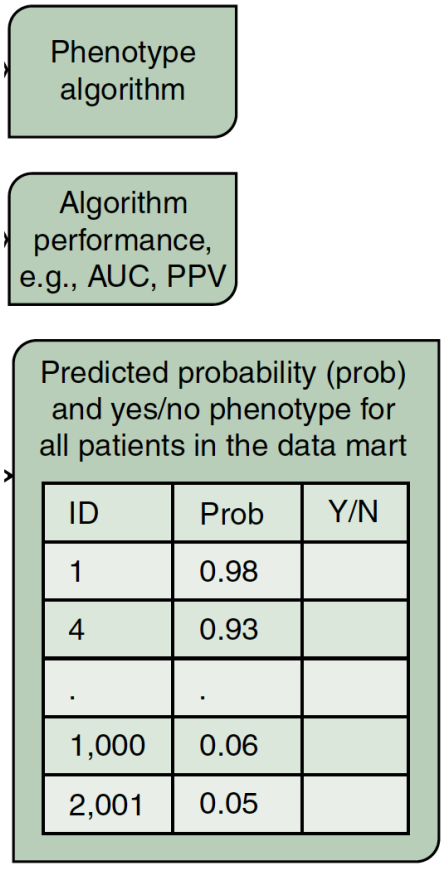


Aside: Inference for prediction metrics

- Confidence intervals should be reported with estimates of prediction performance
- Standard errors can be obtained with bootstrapping

Output

- After this process we obtain:
  - ⋆ The algorithm
  - ⋆ The prediction metrics
  - ⋆ Probability of the phenotype for each patient
- We threshold the probability to obtain the final phenotype classification

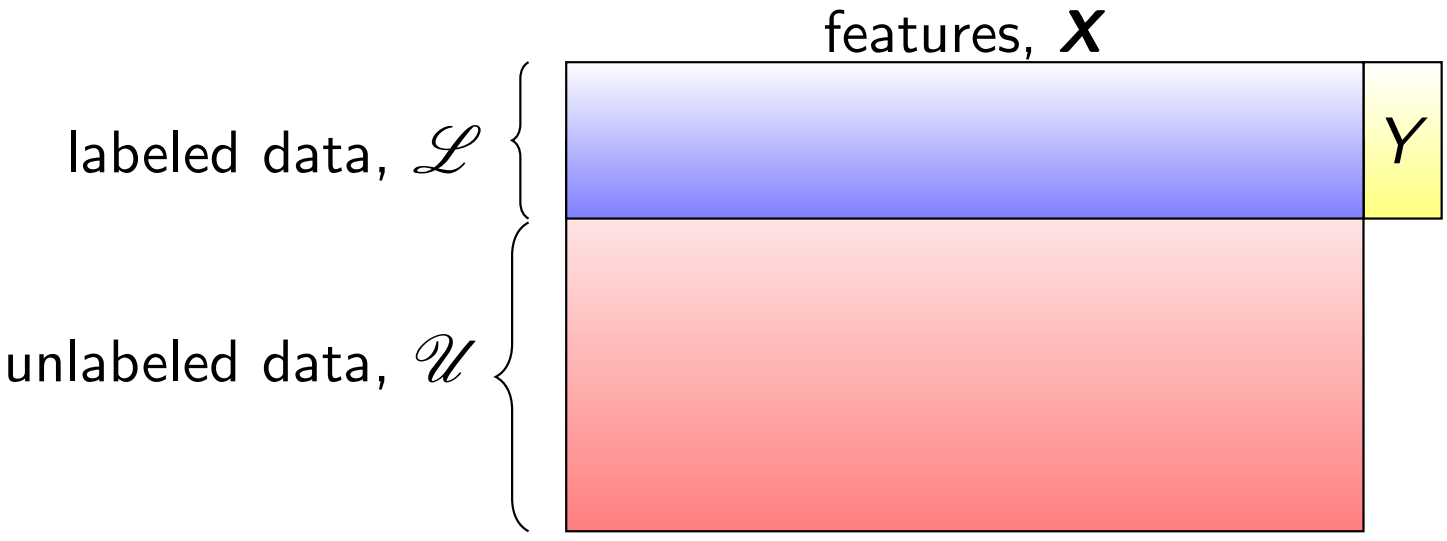


ID	Prob	Y/N
1	0.98	
4	0.93	
.	.	
1,000	0.06	
2,001	0.05	

Module 2: Supervised learning with PheCAP

Part V
Weakly and semi-supervised ML for phenotyping

Recap: Supervised learning



Use $\mathscr{L}$ to train a supervised ML model to predict Y in $\mathscr{U}$

Question of interest

Can we make use the unlabeled data to reduce the amount of labeled data we need?

Question of interest

Can we make use the unlabeled data to reduce the amount of labeled data we need?

$\rightarrow$ The goal of semi-supervised learning

PheCAP backbone: Surrogate Assisted Feature Selection (SAFE)

Challenge

- Y is only available for a small subset of data
- Can only afford to build a model with a small subset of features

Can we use the unlabeled data for feature selection?

Background on SAFE

Goal Remove noise from the candidate feature set

Observation Some features S are more predictive of Y than other features

- $S =$ "surrogate variable"
- e.g. Main ICD or NLP count, labs

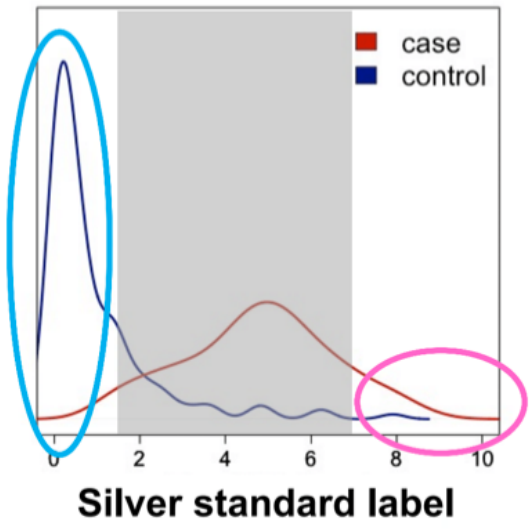
Approach Select features predictive of Y with those predictive of S

Intuition The extremes of S are informative of Y

- Many ICD codes $\rightarrow$ likely case
- Few ICD codes $\rightarrow$ likely control

Overview of SAFE

1. Sample "cases" ($S = 1$) and "controls" ($S = 0$) from the extremes for each surrogate



2. Fit a sparse regression and repeat (eg. 200 times)
   - ⋆ Only informative features recieve nonzero coefficients

3. Select those features with a non-zero coefficient $\geq 50\%$ of the time

Why SAFE works

ORIGINAL ARTICLE

WILEY Biometrics

Automated feature selection of predictors in electronic medical records data

Jessica Gronsbell[1] | Jessica Minnier[2,*] | Sheng Yu[3] | Katherine Liao[4] | Tianxi Cai[5]

SURROGATE AIDED UNSUPERVISED RECOVERY OF SPARSE SIGNALS IN SINGLE INDEX MODELS FOR BINARY OUTCOMES

By Abhishek Chakrabortty[*,†], Matey Neykov, Raymond Carroll and Tianxi Cai[†]

Details of PheCAP

1. Create the list of features
   - ⋆ S: Surrogate
   - ⋆ $\boldsymbol{X}$: Features selected from SAFE
   - ⋆ H: Healthcare utilization

2. Regress each feature against H and S and obtain the residuals as new features, $\boldsymbol{X}^*$
   - ⋆ $\boldsymbol{X}^*$ only provide information about the phenotype above and beyond S and H

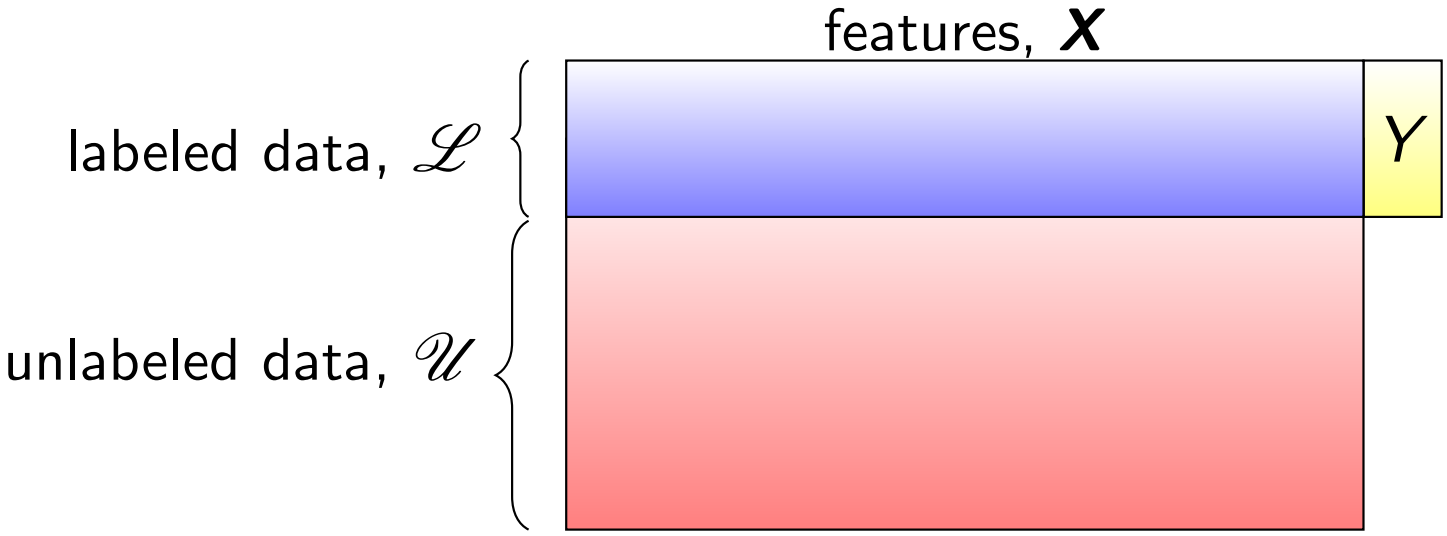
3. Fit a supervised ML model with the training set
   - ⋆ Predict the gold standard labels (Y) using the features ($S, H, \boldsymbol{X}^*$)

Module 3: PheCAP process

Going one step further...

Can we think about this problem through
a slightly different lens?

Problem setting



- Observable Data
  - ⋆ Labeled: $\mathscr{L} = \{(y_i, \boldsymbol{z}_i^\mathsf{T})^\mathsf{T} = (y_i, s_i, \boldsymbol{x}_i^\mathsf{T})^\mathsf{T} \mid i = 1, \ldots, n\}$
  - ⋆ Unlabeled: $\mathscr{U} = \{\boldsymbol{z}_i \mid i = n+1, \ldots, N+n\}$
- Model

$$P(Y = 1 \mid \boldsymbol{X}, S) = g(\alpha_0 + \boldsymbol{\beta}_0^\mathsf{T}\boldsymbol{X} + \gamma_0 S) \text{ where } g(t) = \frac{1}{1+e^{-t}}$$

A neat result

Theorem

If $P(Y = 1 \mid \boldsymbol{X}, S) = g(\alpha_0 + \boldsymbol{\beta}_0^\mathsf{T}\boldsymbol{X} + \gamma_0 S)$ and

1. $S \perp \boldsymbol{X} \mid Y$
2. $\text{cov}(Y, S) > 0$

then $S|\boldsymbol{X}$ is a single index model in $\boldsymbol{\beta}_0^\mathsf{T}\boldsymbol{X}$, $S|\boldsymbol{X} \sim f(\boldsymbol{\beta}_0^\mathsf{T}\boldsymbol{X}; \epsilon)$

There exists a $\rho \neq 0$ such that $\boldsymbol{\beta}_0 = \rho\boldsymbol{\psi}_0$ where

$$(\tau_0, \boldsymbol{\psi_0}) = \underset{\tau,\boldsymbol{\psi}}{\arg\min}\, \mathsf{E}(S - \tau - \boldsymbol{\psi}^\mathsf{T}\boldsymbol{X})^2$$

provided $E(b^\mathsf{T}\boldsymbol{X}|\boldsymbol{\beta}_0^\mathsf{T})$ is linear in $\boldsymbol{\beta}_0^\mathsf{T}\boldsymbol{X}$ for all $b^\mathsf{T}\boldsymbol{X}$.

Neykov et al 2016, Gronsbell et al 2019

Implications of this result

- Simply regressing S on $\boldsymbol{X}$ can provide us with a "good" score, $\hat{\boldsymbol{\psi}}^{\mathsf{T}}\boldsymbol{X}$, for predicting Y
- Since $\boldsymbol{\beta}_0 = \rho\boldsymbol{\psi}_0$, we can rewrite our model as

$$P(Y = 1 \mid \boldsymbol{X}, S) = g(\alpha_0 + \rho\boldsymbol{\psi}_0^{\mathsf{T}}\boldsymbol{X} + \gamma_0 S)$$

Implications of this result

- Simply regressing S on $\boldsymbol{X}$ can provide us with a "good" score, $\hat{\boldsymbol{\psi}}^{\mathsf{T}}\boldsymbol{X}$, for predicting Y
- Since $\boldsymbol{\beta}_0 = \rho\boldsymbol{\psi}_0$, we can rewrite our model as

$$P(Y = 1 \mid \boldsymbol{X}, S) = g(\alpha_0 + \rho\boldsymbol{\psi}_0^{\mathsf{T}}\boldsymbol{X} + \gamma_0 S)$$

$\rightarrow$ We can estimate $\boldsymbol{\psi}_0$ with $\mathscr{U}$ and $(\alpha_0, \gamma_0, \rho)$ with $\mathscr{L}$

Two-step semi-supervised learning

1. Fit a penalized regression of S on $\boldsymbol{X}$ with $\mathscr{U}$ via

$$(\hat{\tau}, \hat{\boldsymbol{\psi}}) = \arg\min_{\tau, \boldsymbol{\psi}} \sum_{i=n+1}^{n+N} (s_i - \tau - \boldsymbol{\psi}^{\mathsf{T}} \boldsymbol{x}_i)^2 + \lambda_N \sum_{j=1}^{p} w_j |\psi_j|$$

   where $w_j = 1/|\widetilde{\psi}_j|$.

2. Fit a logistic regression of Y on $\hat{\boldsymbol{\psi}}^{\mathsf{T}} \boldsymbol{X}$ and S with $\mathscr{L}$ via

$$(\hat{\alpha}, \hat{\gamma}, \hat{\rho}) = \arg\min_{\alpha, \rho, \gamma} \sum_{i=1}^{n} \ell(y_i, \alpha + \rho \hat{\boldsymbol{\psi}}^{\mathsf{T}} \boldsymbol{x}_i + \gamma s_i)$$

   and take $g(\hat{\alpha} + \hat{\rho} \hat{\boldsymbol{\psi}}^{\mathsf{T}} \boldsymbol{x} + \hat{\gamma} s)$ as the prediction for y.

Comparison to PheCAP

Advantages

- Aims to unify feature selection & estimation
- Doesn't rely on an extreme assumption

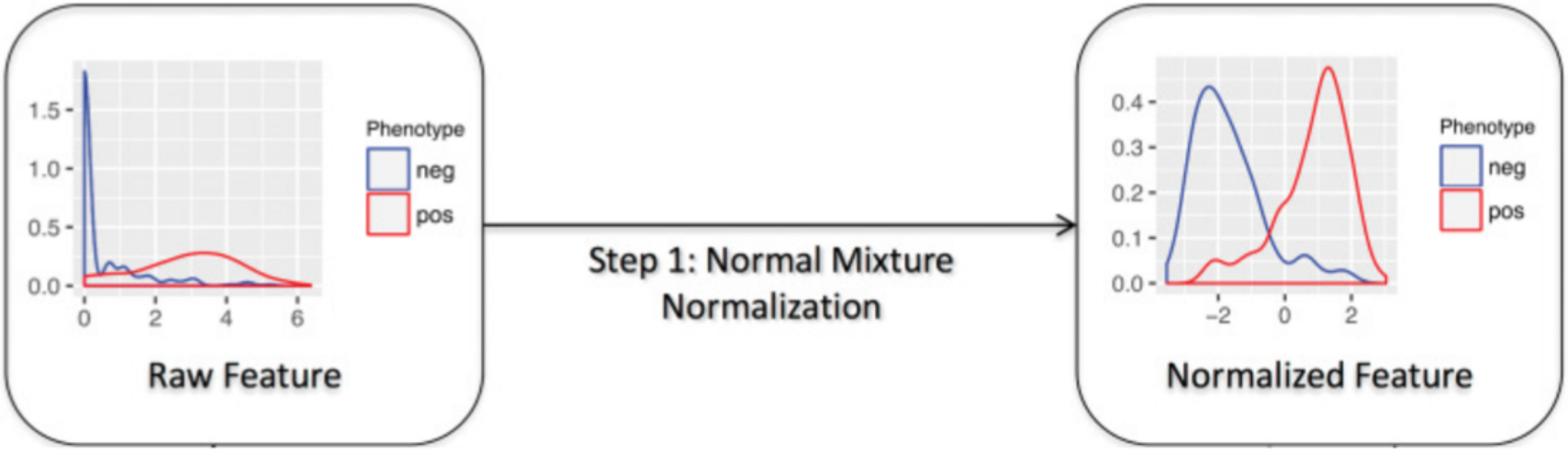
Disadvantages

- Restricted to the logistic regression model
- $S \perp \boldsymbol{X} | Y$ may not hold in practice

Going another step further...

Can we build a phenotyping model without any gold standard labeled data?

Key observation



Normalized surrogates approximately
follow a mixture model

MAP: Multimodal Automated Phenotyping

- Input
  1. S_{ICD}: Count of the main ICD
  2. S_{NLP}: Count of the main NLP
  3. $S_{\text{ICD+NLP}}$: Count of the main ICD and NLP
  4. H: Health utilization feature

- Output
  - ⋆ Predicted probability for the phenotype

- A *weakly-supervised* learning approach
  - ⋆ We utilize surrogates, not gold standard labels for model training

Overview of MAP

1. Fit multiple mixture models adjusting for H for $S_{\text{count}} \in \{S_{\text{ICD}}, S_{\text{NLP}}, S_{\text{ICD+NLP}}\}$:

$$S_{\text{count}} \mid Y = y \sim \text{Poisson}\{\alpha \log(H+1) + \lambda_y\},$$
$$\log(S_{\text{count}}+1) \mid Y = y \sim \text{Normal}\{\alpha \log(H+1) + \mu_y, \sigma_y^2\},$$

   where

   - α adjusts for the total health utilization
   - λ_y and μ_y are the cluster-specific means
   - σ_y^2 is the cluster-specific variance

2. Ensemble the fitted models with model averaging to produce the combined prediction scores

Comparison to the semi-supervised approaches

Advantages

- No labeled data required
- Less data processing

Disadvantages

- Loses some interpretability
- Strongly depends on the strength of surrogate

Module 4: Semi- and weakly-supervised learning

Part VI
Ongoing phenotyping research

Recap: The ideal phenotyping algorithm

- **Accurate**: Can precisely identify the phenotype
- **Efficient**: Can be developed quickly
- **Portable**: Can be implemented across healthcare settings

Accurate phenotyping for more conditions

- Deep learning
  - ⋆ Leveraging temporal and narrative information
- Time-to-event modeling
  - ⋆ Semi-supervised methods for current status data (Ongoing work with Dr. Xuan Wang)
- Semi-supervised methods for model evaluation

Accurate phenotyping for more conditions

- Deep learning
  - ⋆ Leveraging temporal and narrative information
- Time-to-event modeling
  - ⋆ Semi-supervised methods for current status data (Ongoing work with Dr. Xuan Wang)
- Semi-supervised methods for model evaluation

Original Article

Semi-supervised approaches to efficient evaluation of model prediction performance

Jessica L. Gronsbell, Tianxi Cai

First published: 23 December 2017 | https://doi.org/10.1111/rssb.12264 | Citations: 7

Accurate phenotyping for more conditions

- Deep learning
  - ⋆ Leveraging temporal and narrative information
  - ⋆ Questions of interpretability, portability, etc.
- Time-to-event modeling
  - ⋆ Semi-supervised methods for current status data

  (Ongoing work with Dr. Xuan Wang)
- Semi-supervised methods for model evaluation
  - ⋆ Methods to evaluate algorithmic fairness

  (Ongoing work with Siyue Yang)
- Accounting for phenotyping errors
  - ⋆ Correction methods for GWAS

  (Ongoing work with Jianhui Gao & Dr. Zack McCaw)

Making phenotyping more efficient

- Weakly-supervised learning methods
  - ⋆ Positive unlabeled learning
  - ⋆ Joint prediction of related phenotypes (sureLDA)
- Alternative sampling designs

Making phenotyping more efficient

- Weakly-supervised learning methods
  - ⋆ Positive unlabeled learning
  - ⋆ Joint prediction of related phenotypes (sureLDA)
- Alternative sampling designs

Making phenotyping more efficient

- Weakly-supervised learning methods
  - ⋆ Positive unlabeled learning
  - ⋆ Joint prediction of related phenotypes (sureLDA)
- Alternative sampling designs

ORIGINAL ARTICLE

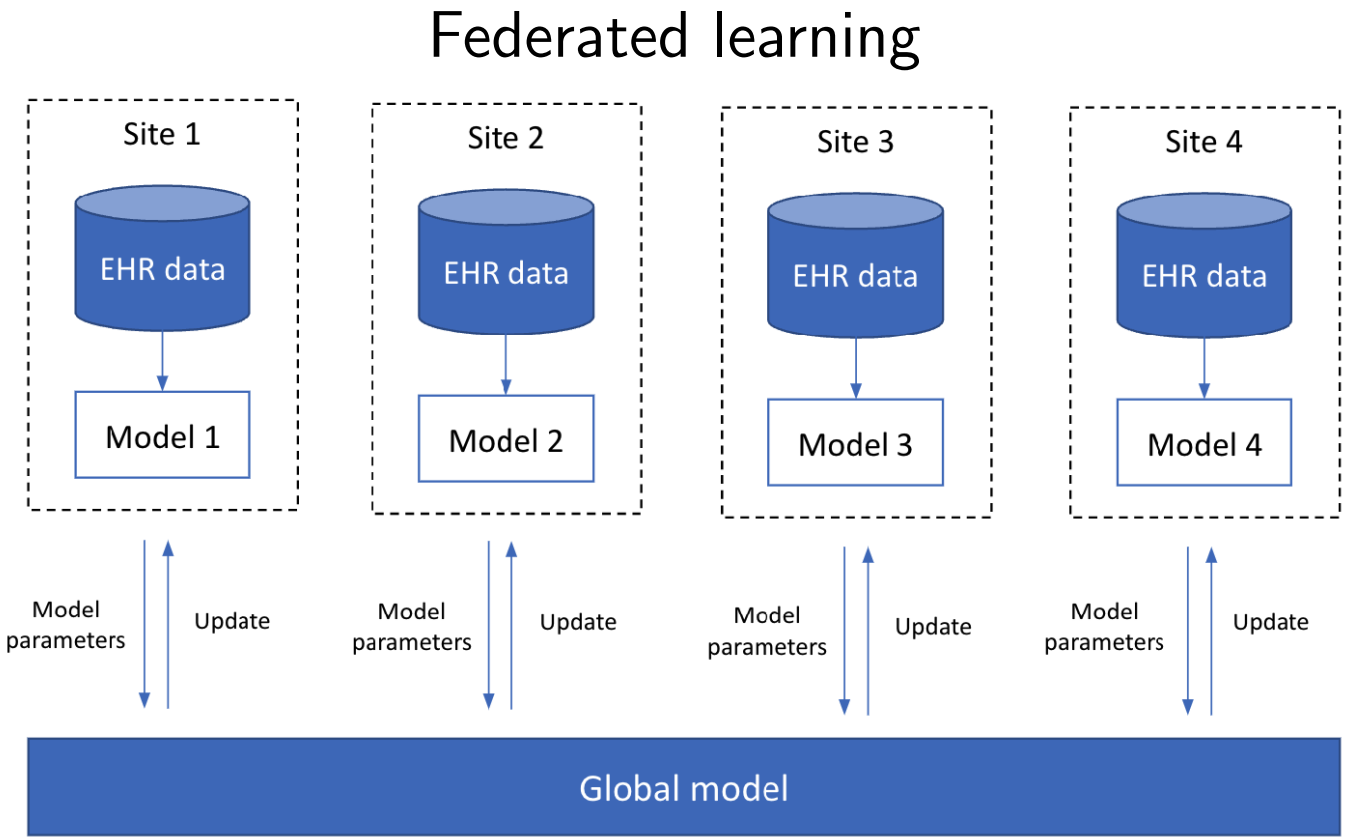
Efficient evaluation of prediction rules in semi-supervised settings under stratified sampling

Jessica Gronsbell, Molei Liu, Lu Tian, Tianxi Cai

First published: 26 April 2022 | https://doi.org/10.1111/rssb.12502

Jessica Gronsbell and Molei Liu are equal contributors to this work.

Improving phenotyping algorithm portability



Thank you!

j.gronsbell@utoronto.ca