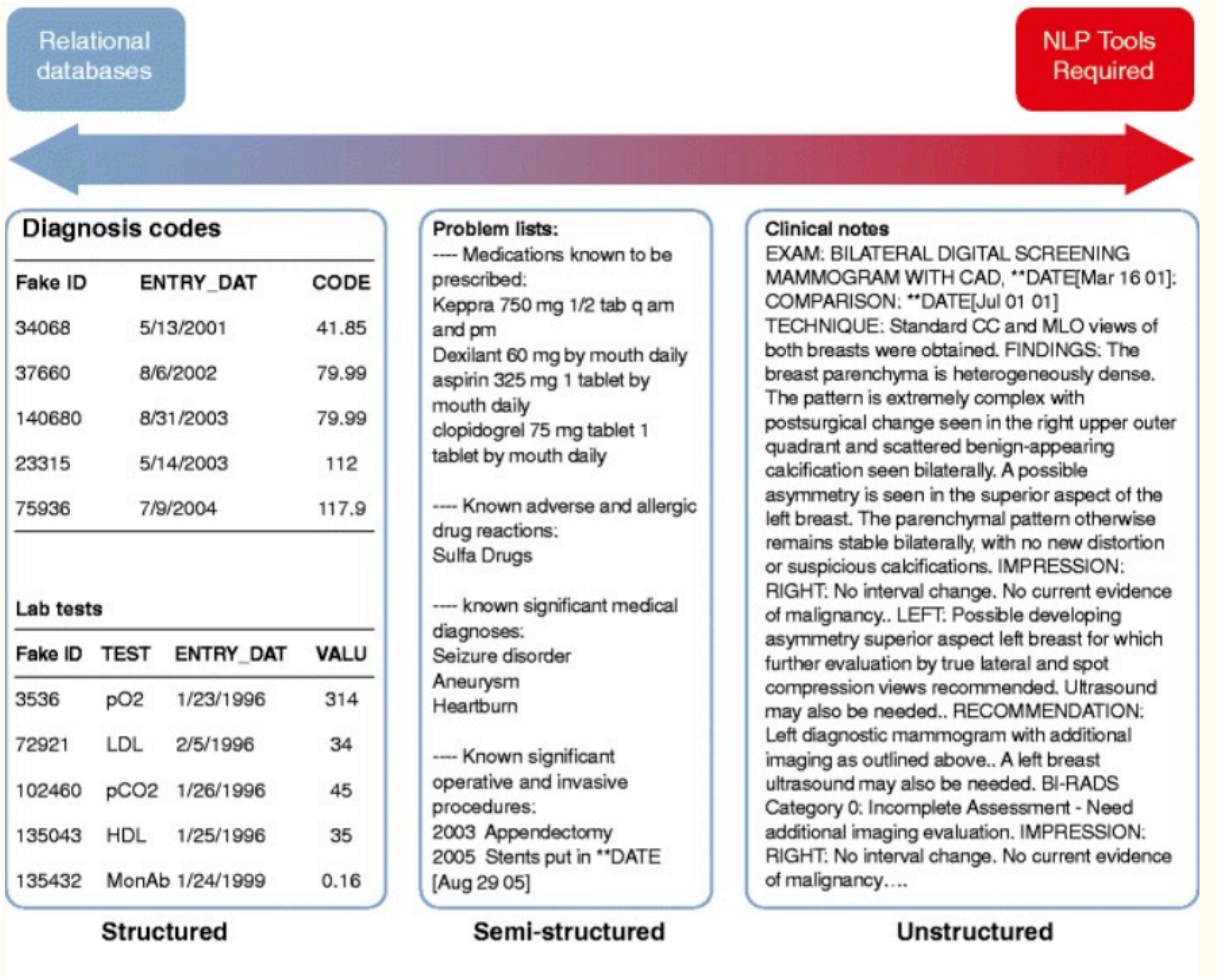
Surrogate-assisted Feature Selection for High-throughput Phenotyping

Sheng Yu, Abhishek Chakrabortty, Katherine P Liao, Tianrun Cai, Ashwin N Ananthakrishnan, Vivian S Gainer, Susanne E Churchill, Peter Szolovits, Shawn N Murphy, Issac S Kohane, and Tianxi Cai

Siyue Yang presented at EHR reading group at October 28, 2021

Electronic medical records (EMRs)

A valuable resource for research



Diagnosis codes

Fake ID	ENTRY_DAT	CODE
34068	5/13/2001	41.85
37660	8/6/2002	79.99
140680	8/31/2003	79.99
23315	5/14/2003	112
75936	7/9/2004	117.9

Lab tests

Fake ID	TEST	ENTRY_DAT	VALU
3536	pO2	1/23/1996	314
72921	LDL	2/5/1996	34
102460	pCO2	1/26/1996	45
135043	HDL	1/25/1996	35
135432	MonAb	1/24/1999	0.16

Problem lists:

---- Medications known to be prescribed:
Keppra 750 mg 1/2 tab q am and pm
Dexilant 60 mg by mouth daily
aspirin 325 mg 1 tablet by mouth daily
clopidogrel 75 mg tablet 1 tablet by mouth daily

---- Known adverse and allergic drug reactions:
Sulfa Drugs

---- known significant medical diagnoses:
Seizure disorder
Aneurysm
Heartburn

---- Known significant operative and invasive procedures:
2003 Appendectomy
2005 Stents put in **DATE [Aug 29 05]

Clinical notes

EXAM: BILATERAL DIGITAL SCREENING MAMMOGRAM WITH CAD, **DATE[Mar 16 01]: COMPARISON: **DATE[Jul 01 01] TECHNIQUE: Standard CC and MLO views of both breasts were obtained. FINDINGS: The breast parenchyma is heterogeneously dense. The pattern is extremely complex with postsurgical change seen in the right upper outer quadrant and scattered benign-appearing calcification seen bilaterally. A possible asymmetry is seen in the superior aspect of the left breast. The parenchymal pattern otherwise remains stable bilaterally, with no new distortion or suspicious calcifications. IMPRESSION: RIGHT: No interval change. No current evidence of malignancy.. LEFT: Possible developing asymmetry superior aspect left breast for which further evaluation by true lateral and spot compression views recommended. Ultrasound may also be needed.. RECOMMENDATION: Left diagnostic mammogram with additional imaging as outlined above.. A left breast ultrasound may also be needed. BI-RADS Category 0: Incomplete Assessment - Need additional imaging evaluation. IMPRESSION: RIGHT: No interval change. No current evidence of malignancy....

Source: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4416392/

Electronic medical records (EMRs)

A valuable resource for research

- Contain longitudinal patient conditions, histories, outcomes
- Widely adopted worldwide
- Faster and more inclusive to recruit patients

Opportunities for EMR-based research

- EMR data and/or biorepository
  - Genetic association studies
  - Comparative effectiveness
  - Risk stratification
  - Clinical trail recruitment
  - Patient monitoring

Opportunities for EMR-based research

- EMR data and/or biorepository
  - Genetic association studies
  - Comparative effectiveness
  - Risk stratification
  - Clinical trail recruitment
  - Patient monitoring

First, we need to get a cohort of patients …

- have the **disease**
- respond to the **treatment**
- have the relevant lifestyle **factors**

Phenotyping

- (Electronic) phenotypes
  - Patient characteristics
  - e.g. Disease status, treatment response, lifestyle factors
- Phenotyping
  - The process to extract (electronic) phenotypes from EMRs

Phenotyping

- (Electronic) phenotypes
    - Patient characteristics
    - e.g. Disease status, treatment response, lifestyle factors
- Phenotyping
    - The process to extract (electronic) phenotypes from EMRs

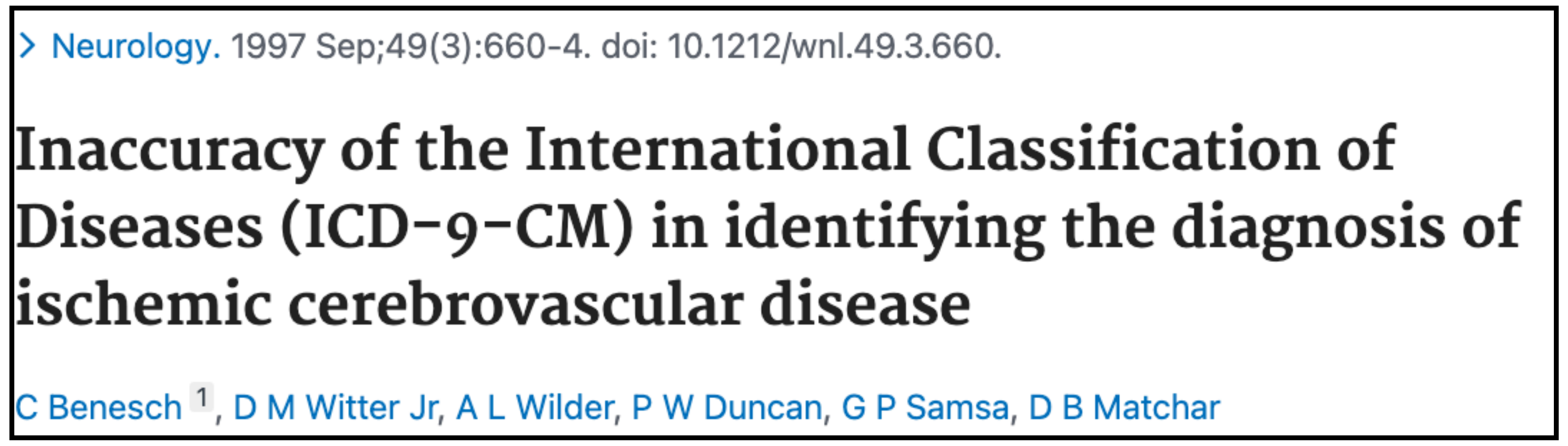
Phenotyping is fundamental in EMR-based studies!

How to extract accurate phenotypes?

- Option 1: Diagnosis codes (e.g. ICD-9)
    - Presence of related codes = having the disease

How to extract accurate phenotypes?

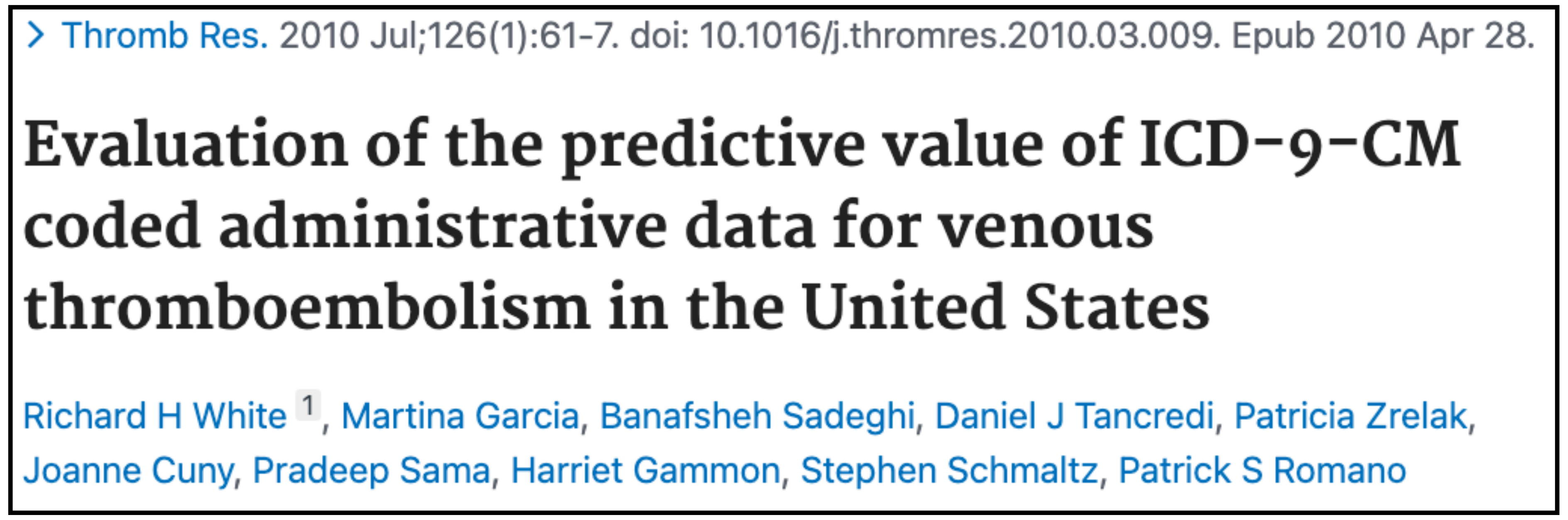
- Option 1: Diagnosis codes (e.g. ICD-9)



> Neurology. 1997 Sep;49(3):660-4. doi: 10.1212/wnl.49.3.660.

Inaccuracy of the International Classification of Diseases (ICD-9-CM) in identifying the diagnosis of ischemic cerebrovascular disease

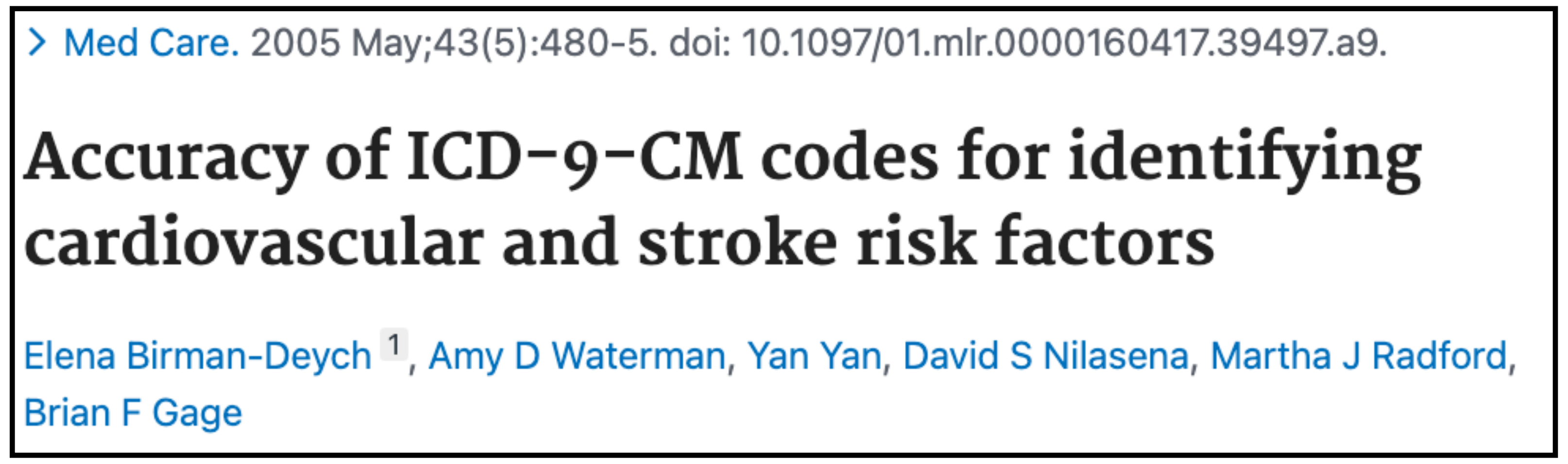
C Benesch [1], D M Witter Jr, A L Wilder, P W Duncan, G P Samsa, D B Matchar



> Thromb Res. 2010 Jul;126(1):61-7. doi: 10.1016/j.thromres.2010.03.009. Epub 2010 Apr 28.

Evaluation of the predictive value of ICD-9-CM coded administrative data for venous thromboembolism in the United States

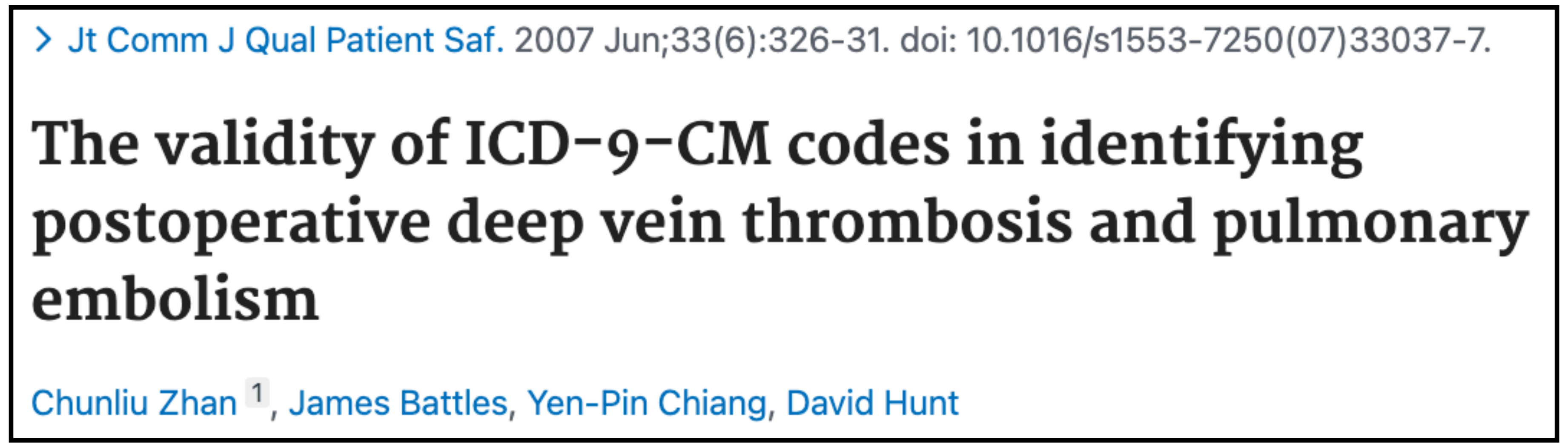
Richard H White [1], Martina Garcia, Banafsheh Sadeghi, Daniel J Tancredi, Patricia Zrelak, Joanne Cuny, Pradeep Sama, Harriet Gammon, Stephen Schmaltz, Patrick S Romano



> Med Care. 2005 May;43(5):480-5. doi: 10.1097/01.mlr.0000160417.39497.a9.

Accuracy of ICD-9-CM codes for identifying cardiovascular and stroke risk factors

Elena Birman-Deych [1], Amy D Waterman, Yan Yan, David S Nilasena, Martha J Radford, Brian F Gage



> Jt Comm J Qual Patient Saf. 2007 Jun;33(6):326-31. doi: 10.1016/s1553-7250(07)33037-7.

The validity of ICD-9-CM codes in identifying postoperative deep vein thrombosis and pulmonary embolism

Chunliu Zhan [1], James Battles, Yen-Pin Chiang, David Hunt

How to extract accurate phenotypes?

- Option 1: Diagnosis codes (e.g. ICD-9)
    - Imperfect phenotypes in subsequent genomic studies
    - Power loss + bias

How to extract accurate phenotypes?

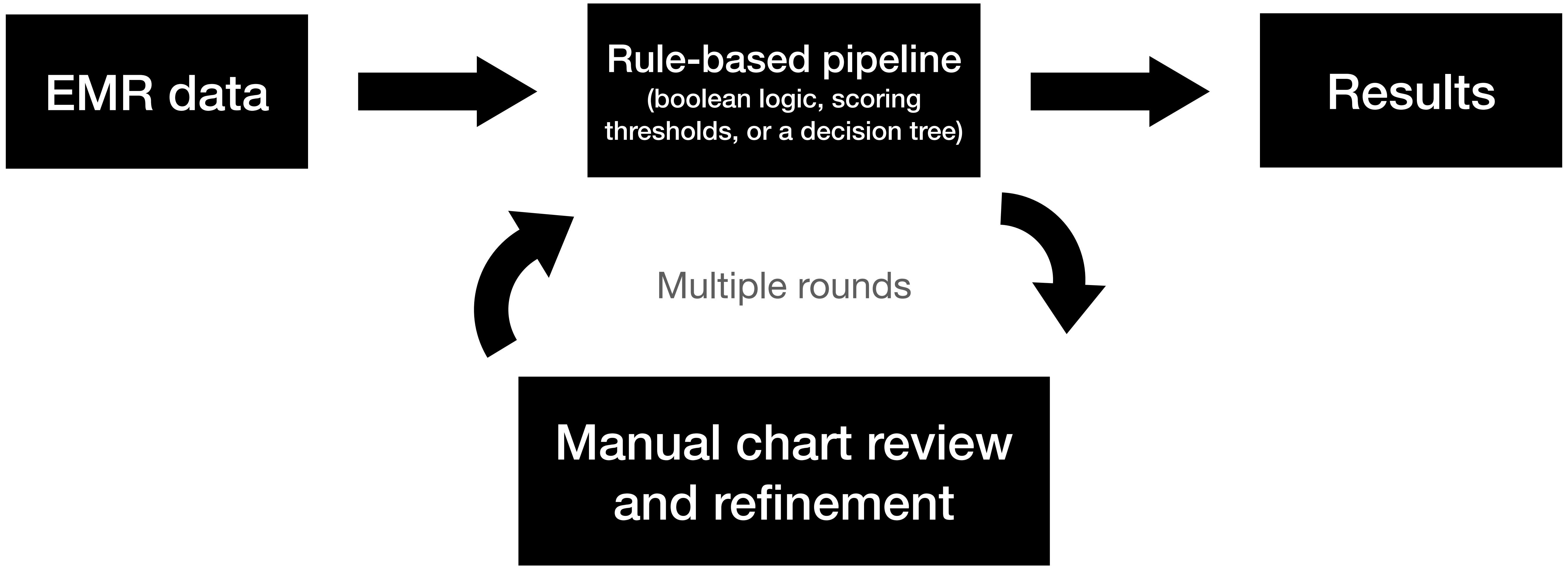
- Option 1: Diagnosis codes (e.g. ICD-9)
- Option 2: Rule-based algorithms
    - Combine ICD-9 codes and other structured data
    - Inclusion and exclusion criteria

How to extract accurate phenotypes?

- Option 1: Diagnosis codes (e.g. ICD-9)
- Option 2: Rule-based algorithms
    - Example: Type 2 diabetes
        - Presence of the diagnosis codes
        - At least one hypoglycaemic medication
        - HbA1c above certain threshold
    - More stringent, accuracy improved

How to extract accurate phenotypes?

- Option 1: Diagnosis codes (e.g. ICD-9)
- Option 2: Rule-based algorithms


EMR data
Rule-based pipeline
(boolean logic, scoring thresholds, or a decision tree)
Results
Multiple rounds
Manual chart review and refinement

How to extract accurate phenotypes?

- Option 1: Diagnosis codes (e.g. ICD-9)
- Option 2: Rule-based algorithms
    - Advantage: human-interpretable algorithms
    - Disadvantages
        - Significant effort, time, expertise knowledge
        - Infeasible for phenotypes not first envisioned by clinicians

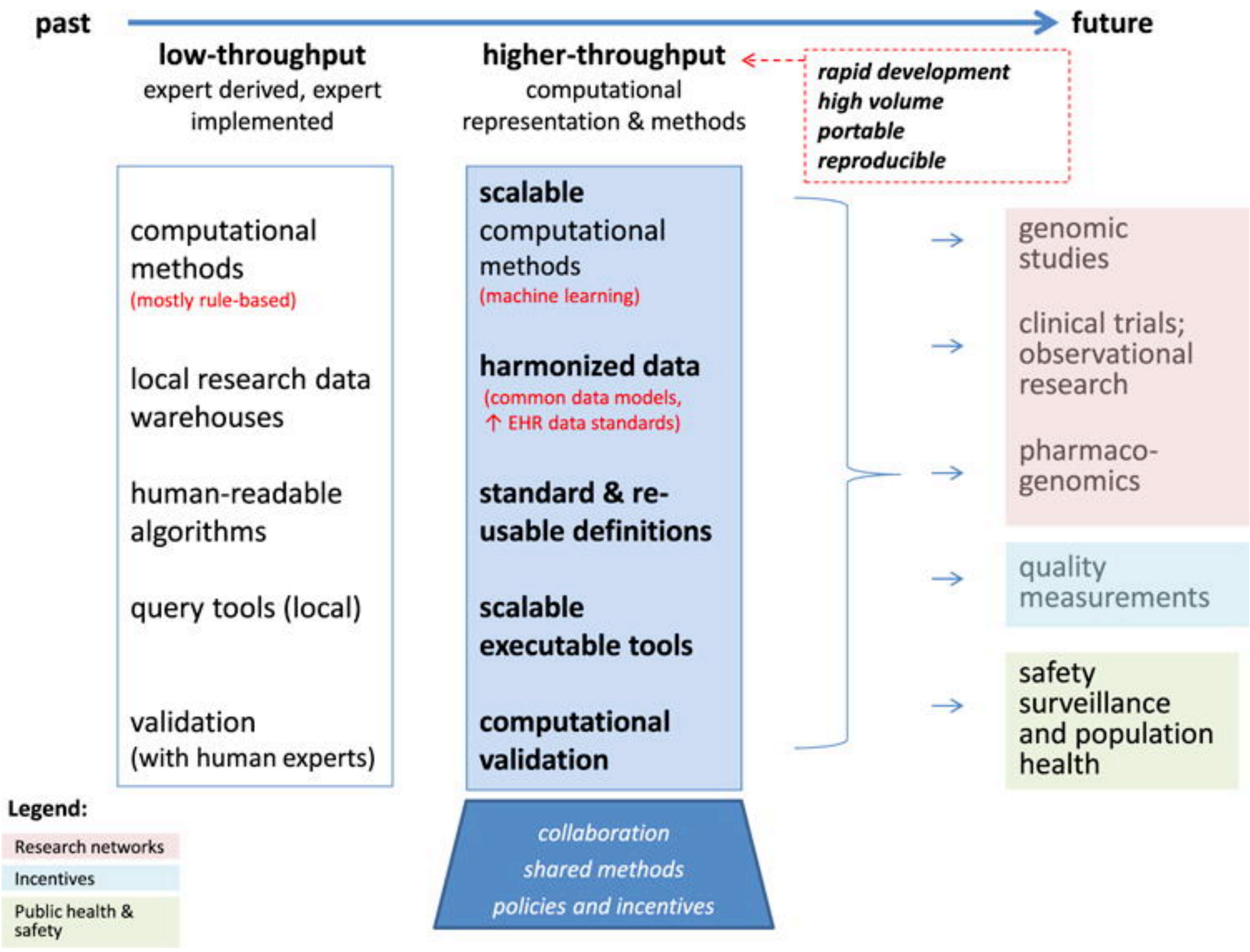
How to extract accurate phenotypes?

- Option 1: Diagnosis codes (e.g. ICD-9)
- Option 2: Rule-based algorithms
- Option 3: Machine learning algorithms
    - Data-driven

How to extract accurate phenotypes?

- Option 1: Diagnosis codes (e.g. ICD-9)
- Option 2: Rule-based algorithms
- Option 3: Machine learning algorithms
    - Data-driven
    - Reduce efforts required from domain experts
    - Towards “high-throughput phenotyping”

High-throughput phenotyping



past → future

low-throughput
expert derived, expert implemented

computational methods (mostly rule-based)

local research data warehouses

human-readable algorithms

query tools (local)

validation (with human experts)

higher-throughput
computational representation & methods

rapid development
high volume
portable
reproducible

scalable computational methods (machine learning)

harmonized data (common data models, ↑ EHR data standards)

standard & re-usable definitions

scalable executable tools

computational validation

collaboration
shared methods
policies and incentives

→ genomic studies

→ clinical trials; observational research

→ pharmaco-genomics

→ quality measurements

→ safety surveillance and population health

Legend:

- 🟥 Research networks
- 🟦 Incentives
- 🟩 Public health & safety

Source: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5480212/

Bottlenecks in high-throughput phenotyping

Feature curation and labeling

Key steps in phenotyping	Details	Rate-limiting part
Collecting informative features	Structured features: database queries Unstructured features: Natural Language Processing (NLP)	Feature curation
Developing classification algorithms with features and a gold-standard training set	Expert randomly select a subset of patients to do chart reviews	Labeling

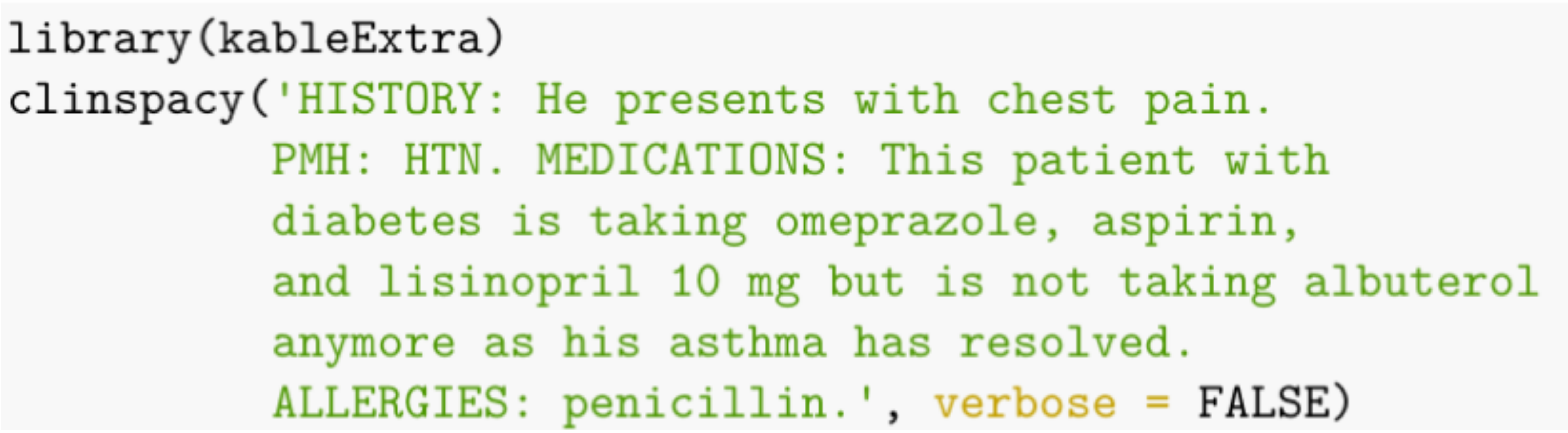
Collecting informative features

- Structured features
    - Counts of a patient’s ICD-9 codes, codes of diagnostic and therapeutic procedures, medication prescriptions, and lab codes/values
- Unstructured (NLP) features
    - Frequency of various medical concepts mentioned in patient’s notes

NLP features can be tens of thousands

For example, let’s use NLP to process three sentences

- NLP “Clinispacy” R package
- Unified Medical Language System (UMLS) concept mapping
- Negation detection



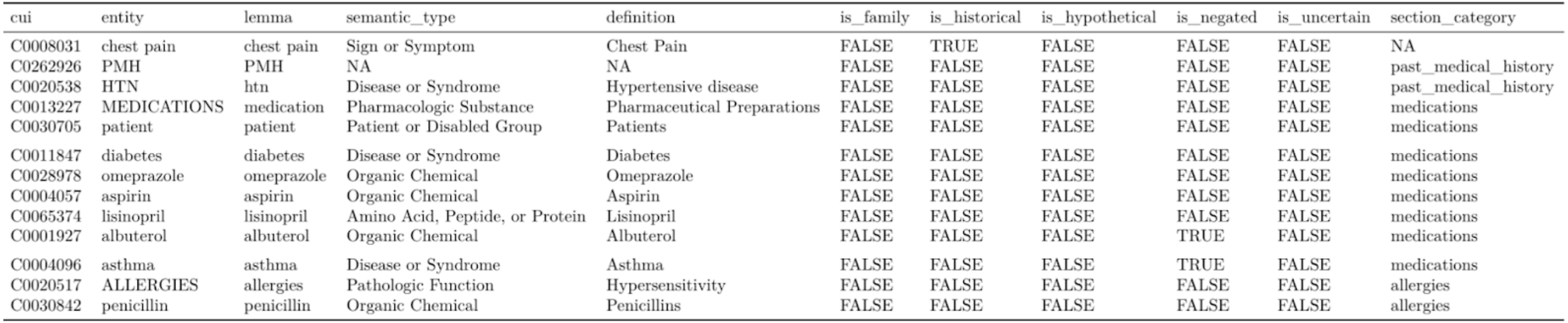
```
library(kableExtra)
clinspacy('HISTORY: He presents with chest pain.
          PMH: HTN. MEDICATIONS: This patient with
          diabetes is taking omeprazole, aspirin,
          and lisinopril 10 mg but is not taking albuterol
          anymore as his asthma has resolved.
          ALLERGIES: penicillin.', verbose = FALSE)
```

Source: https://github.com/ML4LHS/clinspacy

NLP features can be tens of thousands

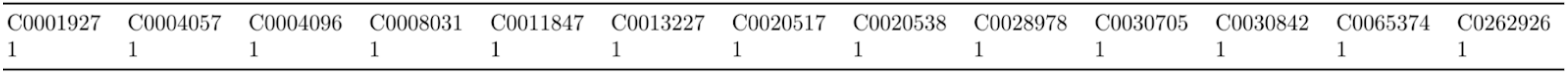
For example, let’s use NLP to process three sentences

E.g. “HISTORY: He presents with chest pain. PMH: HTN. MEDICATIONS: This patient with diabetes is taking omeprazole, aspirin, and lisinopril 10 mg but is not taking albuterol anymore as his asthma has resolved. ALLERGIES: penicillin.”



cui	entity	lemma	semantic_type	definition	is_family	is_historical	is_hypothetical	is_negated	is_uncertain	section_category
C0008031	chest pain	chest pain	Sign or Symptom	Chest Pain	FALSE	TRUE	FALSE	FALSE	FALSE	NA
C0262926	PMH	PMH	NA	NA	FALSE	FALSE	FALSE	FALSE	FALSE	past_medical_history
C0020538	HTN	htn	Disease or Syndrome	Hypertensive disease	FALSE	FALSE	FALSE	FALSE	FALSE	past_medical_history
C0013227	MEDICATIONS	medication	Pharmacologic Substance	Pharmaceutical Preparations	FALSE	FALSE	FALSE	FALSE	FALSE	medications
C0030705	patient	patient	Patient or Disabled Group	Patients	FALSE	FALSE	FALSE	FALSE	FALSE	medications
C0011847	diabetes	diabetes	Disease or Syndrome	Diabetes	FALSE	FALSE	FALSE	FALSE	FALSE	medications
C0028978	omeprazole	omeprazole	Organic Chemical	Omeprazole	FALSE	FALSE	FALSE	FALSE	FALSE	medications
C0004057	aspirin	aspirin	Organic Chemical	Aspirin	FALSE	FALSE	FALSE	FALSE	FALSE	medications
C0065374	lisinopril	lisinopril	Amino Acid, Peptide, or Protein	Lisinopril	FALSE	FALSE	FALSE	FALSE	FALSE	medications
C0001927	albuterol	albuterol	Organic Chemical	Albuterol	FALSE	FALSE	FALSE	TRUE	FALSE	medications
C0004096	asthma	asthma	Disease or Syndrome	Asthma	FALSE	FALSE	FALSE	TRUE	FALSE	medications
C0020517	ALLERGIES	allergies	Pathologic Function	Hypersensitivity	FALSE	FALSE	FALSE	FALSE	FALSE	allergies
C0030842	penicillin	penicillin	Organic Chemical	Penicillins	FALSE	FALSE	FALSE	FALSE	FALSE	allergies

- NLP features



C0001927	C0004057	C0004096	C0008031	C0011847	C0013227	C0020517	C0020538	C0028978	C0030705	C0030842	C0065374	C0262926
1	1	1	1	1	1	1	1	1	1	1	1	1

Using all possible features

A nightmare

- Avoid the need for selecting features
- Huge number of irrelevant features
- Overfitting
- Poor out-of-sample classification accuracy

How can we deal with large amount of features?

- Manual feature selection
    - Time-consuming
    - Not ideal for studies involving many phenotypes

How can we deal with large amount of features?

- Manual feature selection
  - Time-consuming
  - Not ideal for studies involving many phenotypes
- Machine learning feature selection
  - Need to create gold-standard labels
  - Time-consuming

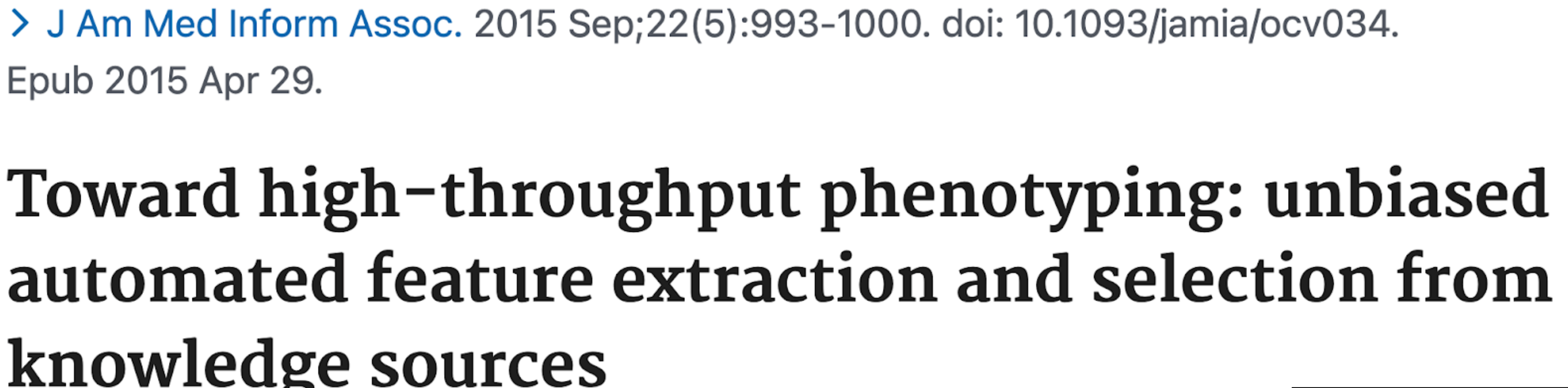
We need automated feature selection methods!

Automated feature selection

- Choose a small set of informative features
- Ideally,
    - Feature selection without using any gold-standard labels
    - Classification algorithm with
        - Selected features
        - 100-200 gold-standard labels

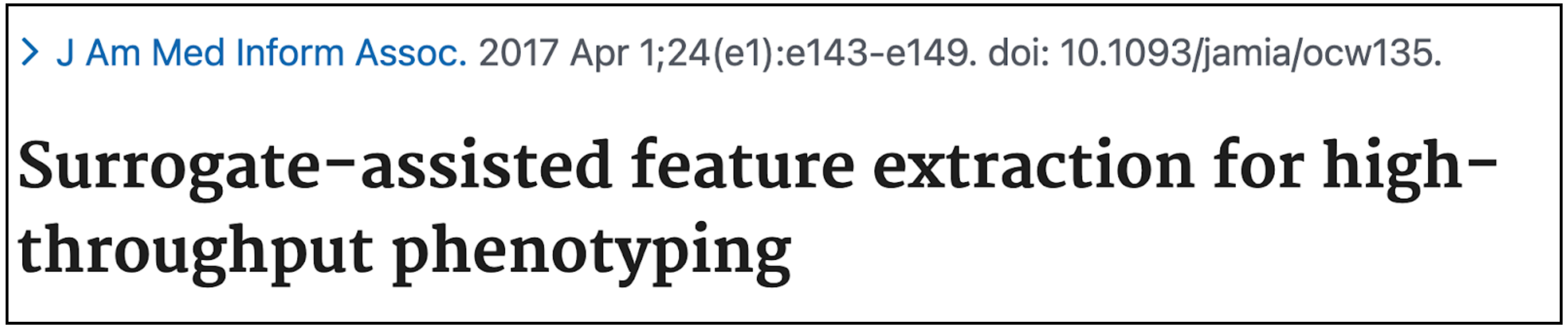
Automated feature selection

Related publications



> J Am Med Inform Assoc. 2015 Sep;22(5):993-1000. doi: 10.1093/jamia/ocv034. Epub 2015 Apr 29.

Toward high-throughput phenotyping: unbiased automated feature extraction and selection from knowledge sources



> J Am Med Inform Assoc. 2017 Apr 1;24(e1):e143-e149. doi: 10.1093/jamia/ocw135.

Surrogate-assisted feature extraction for high-throughput phenotyping

> J Biomed Inform. 2019 Mar;91:103122. doi: 10.1016/j.jbi.2019.103122. Epub 2019 Feb 7.

Feature extraction for phenotyping from semantic and knowledge resources

Comparison of the three

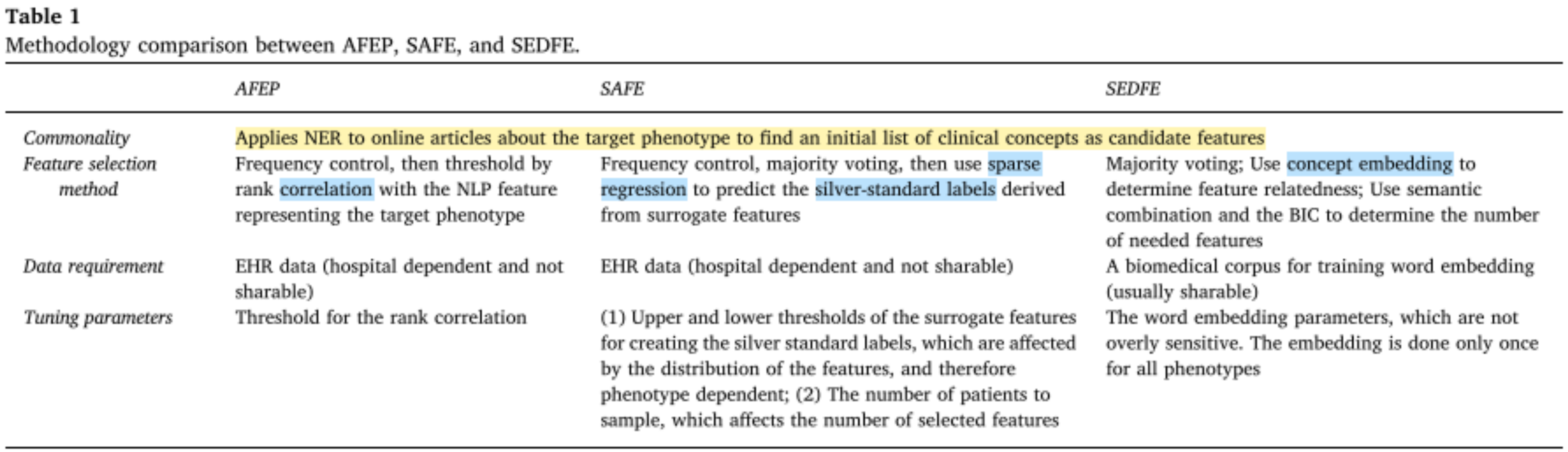


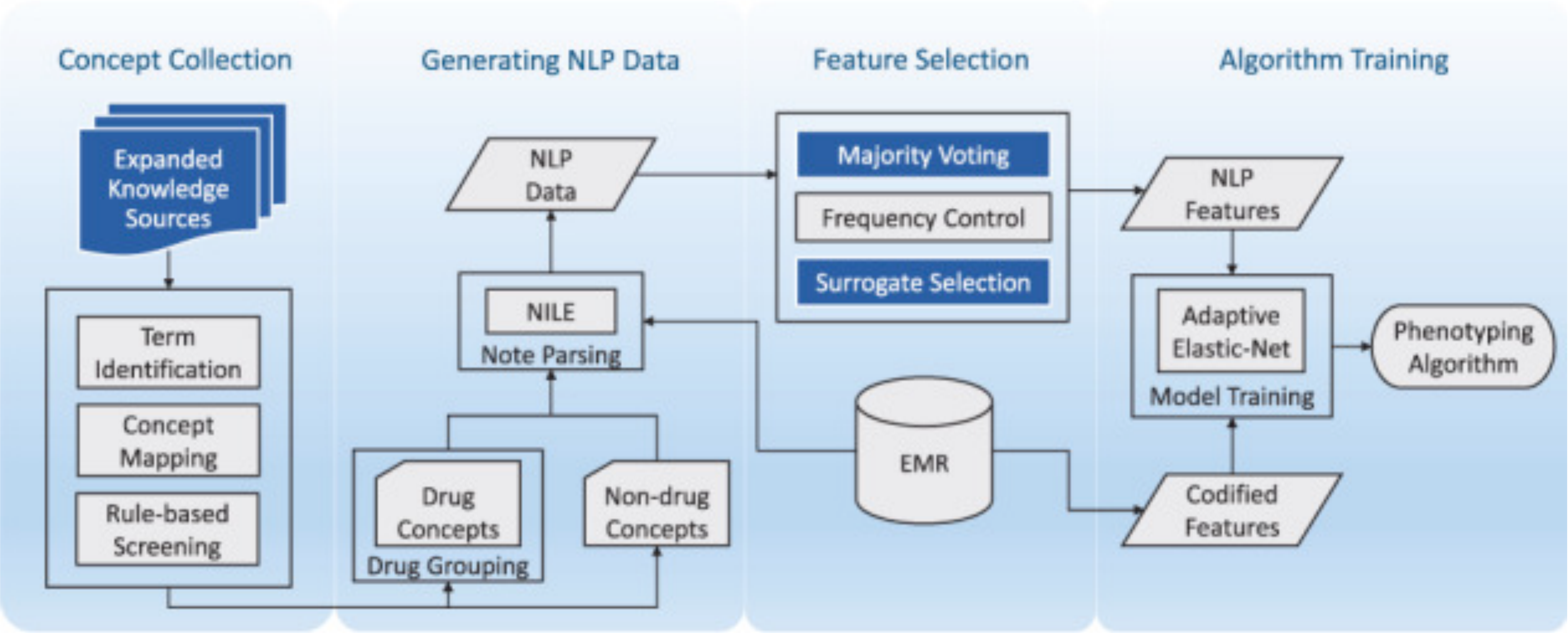
Table 1
Methodology comparison between AFEP, SAFE, and SEDFE.

	AFEP	*SAFE*	*SEDFE*
Commonality	Applies NER to online articles about the target phenotype to find an initial list of clinical concepts as candidate features		
Feature selection method	Frequency control, then threshold by rank correlation with the NLP feature representing the target phenotype	Frequency control, majority voting, then use sparse regression to predict the silver-standard labels derived from surrogate features	Majority voting; Use concept embedding to determine feature relatedness; Use semantic combination and the BIC to determine the number of needed features
Data requirement	EHR data (hospital dependent and not sharable)	EHR data (hospital dependent and not sharable)	A biomedical corpus for training word embedding (usually sharable)
Tuning parameters	Threshold for the rank correlation	(1) Upper and lower thresholds of the surrogate features for creating the silver standard labels, which are affected by the distribution of the features, and therefore phenotype dependent; (2) The number of patients to sample, which affects the number of selected features	The word embedding parameters, which are not overly sensitive. The embedding is done only once for all phenotypes

The goal of this paper

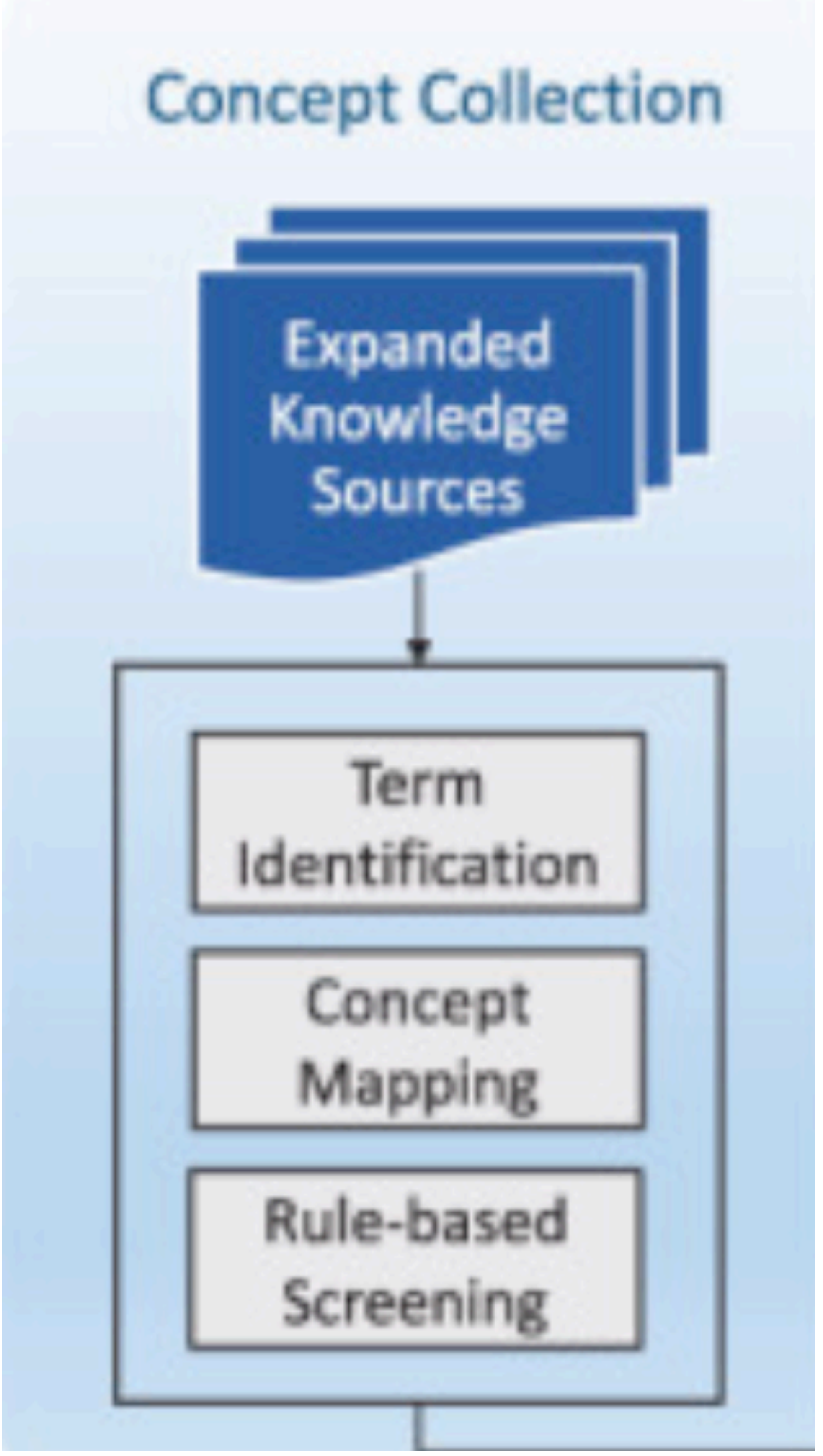
Develop automated feature selection methods for high-throughput phenotyping through the use of easily available but noisy surrogates

Methods


Concept Collection
Generating NLP Data
Feature Selection
Algorithm Training
Expanded Knowledge Sources
Term Identification
Concept Mapping
Rule-based Screening
NLP Data
NILE
Note Parsing
Drug Concepts
Drug Grouping
Non-drug Concepts
Majority Voting
Frequency Control
Surrogate Selection
EMR
NLP Features
Adaptive Elastic-Net
Model Training
Codified Features
Phenotyping Algorithm

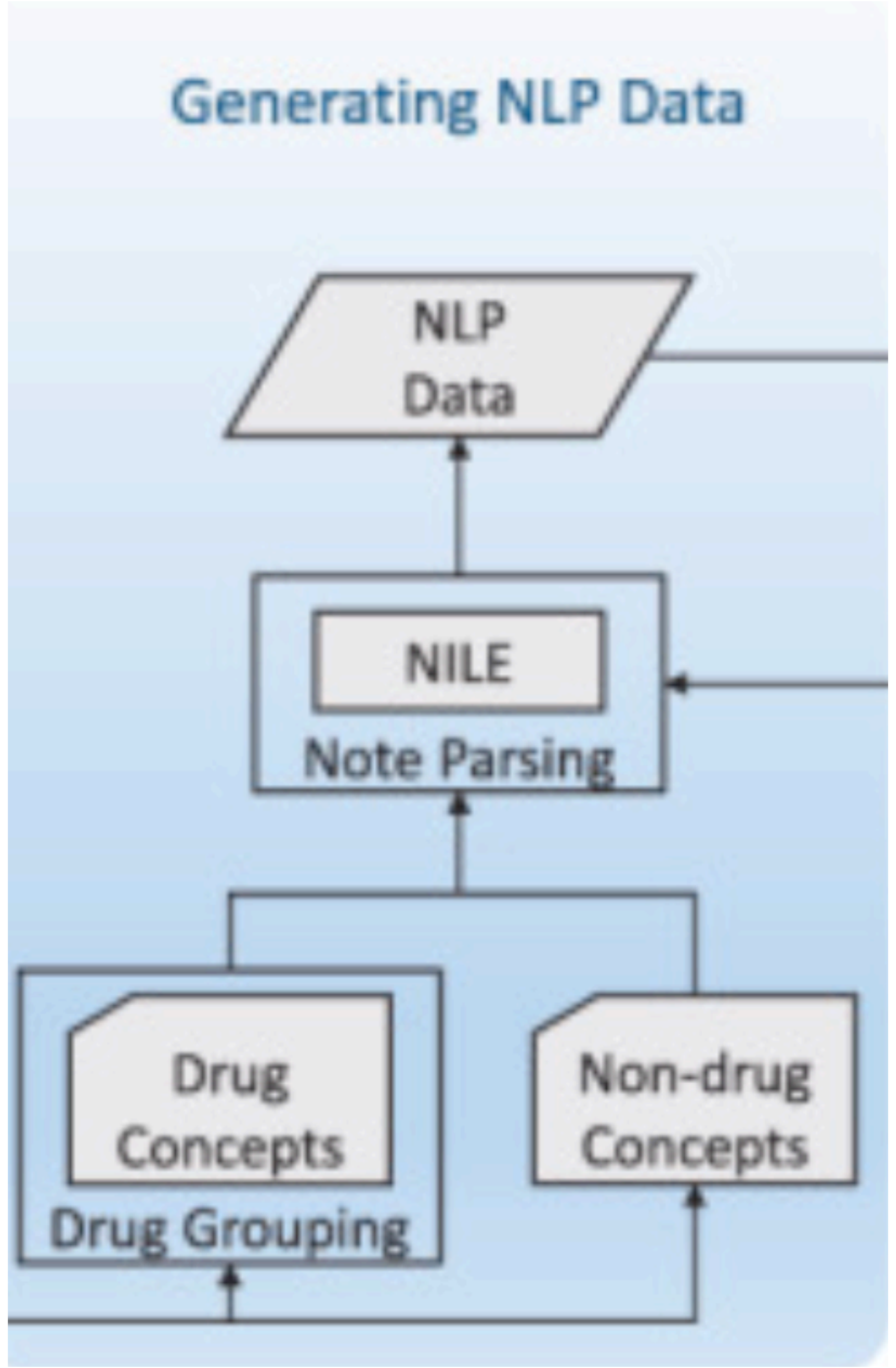
1. Concept collection

- Publicly available data source
  - Wikipedia, Medscape, Merk Manuals Professional Edition, Mayo Clinic Diseases and Conditions, and MedlinePlus Medical Encyclopedia
- Candidate features: ~1000 UMLS concepts



2. Generating NLP data

- Mentions of the candidate concepts
- Summarised in patient-level counts
- Only positive mentions
- Not include negated assertions, family histories, and conditional problems



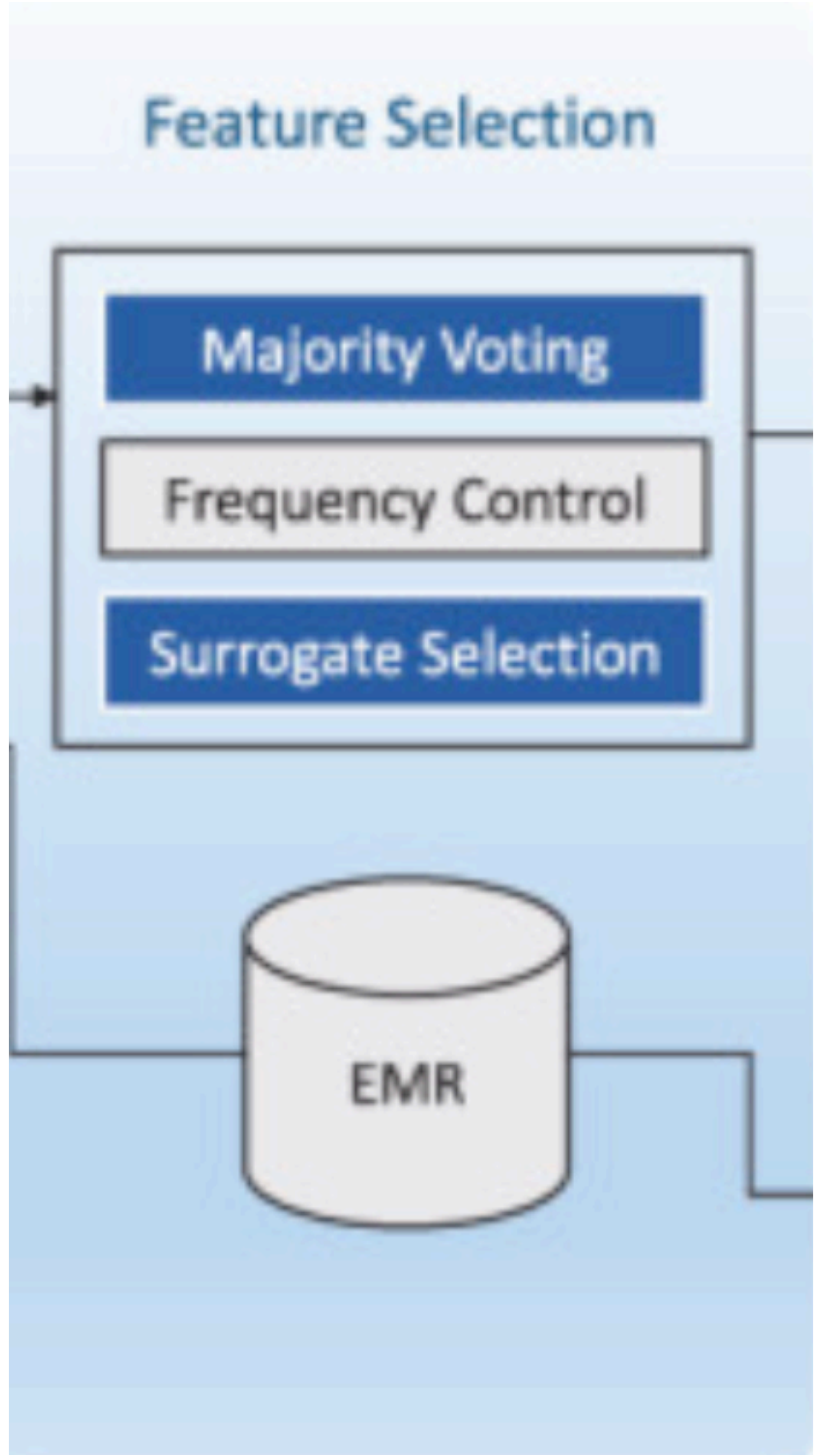
Short discussion

Why using only positive mentions?

- Double count
- Avoid too many features
- Other considerations?

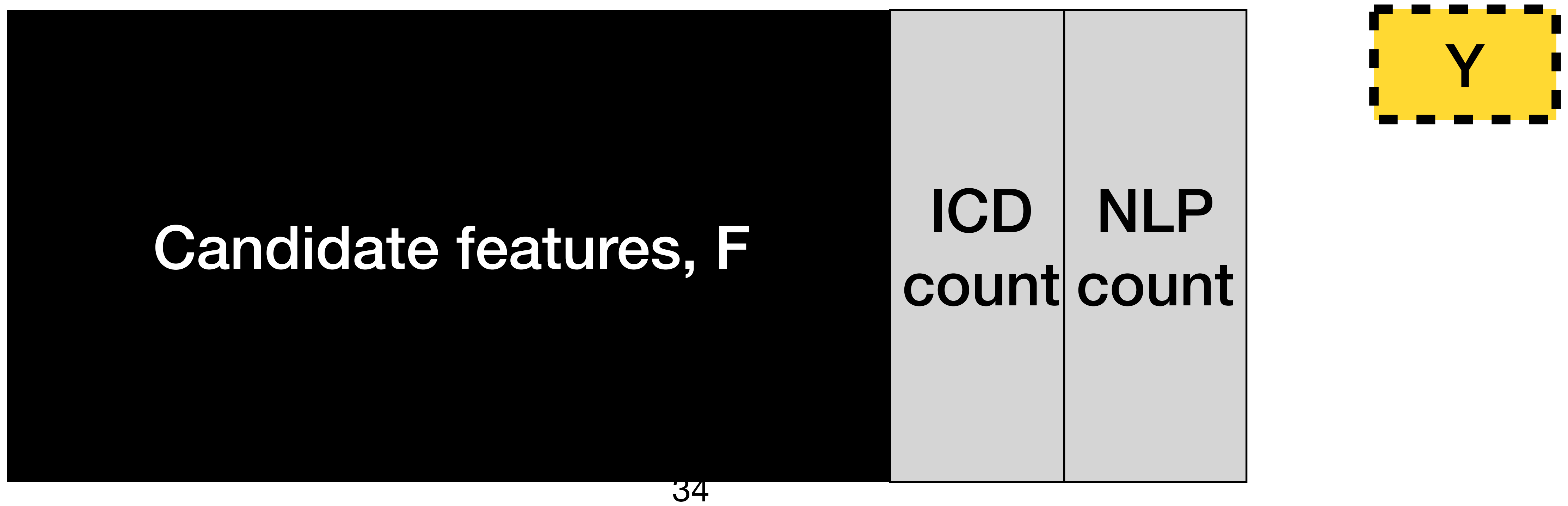
3. Feature selection

- Majority voting
- Frequency control
  - at least 5% notes
  - no more than 50% of all patients



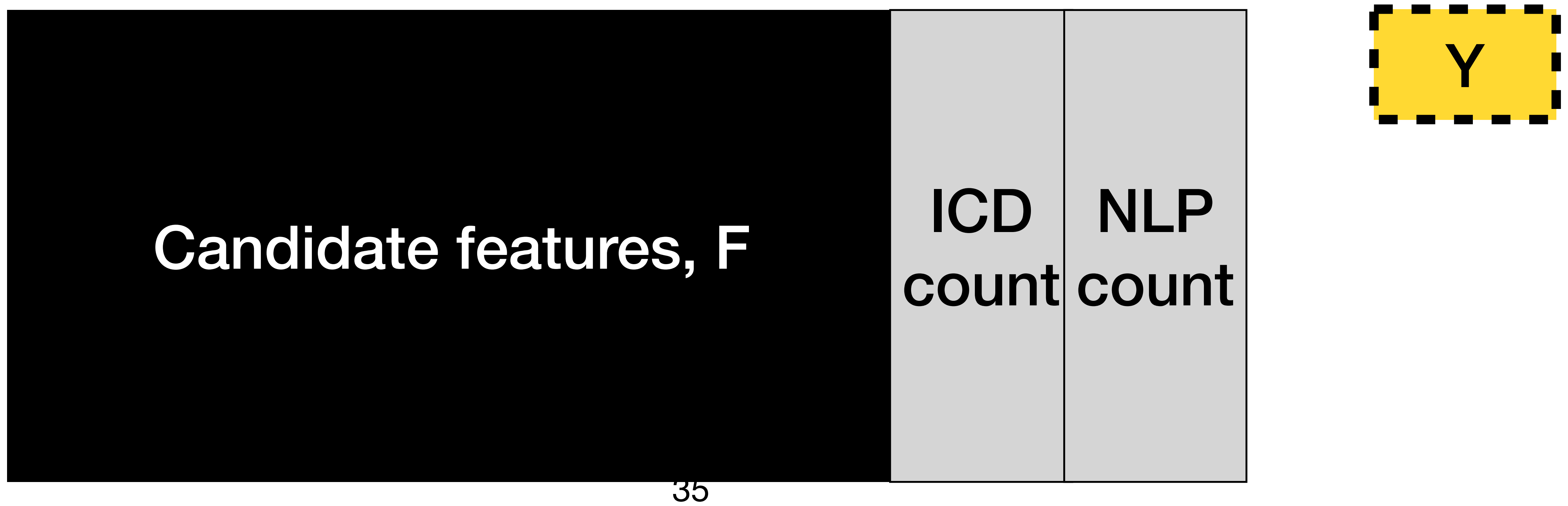
3. Feature selection - surrogate selection

- Data we have for feature selection
  - **main counts** of ICD-9 codes (codes of all subtypes)
  - **main counts** of NLP (UMLS concepts)
  - candidate features pass the 2 steps, F_{cand}
- Our goal is to find a subset of F_{cand} that is related to true disease status, Y



We have limited gold-standard labels

- If Y is available for each patient
  - Machine learning feature selection is straightforward
  - e.g. Sparse logistic regression of Y against F_{cand}
- but our goal is **not to use Y** to achieve **full automated feature selection**

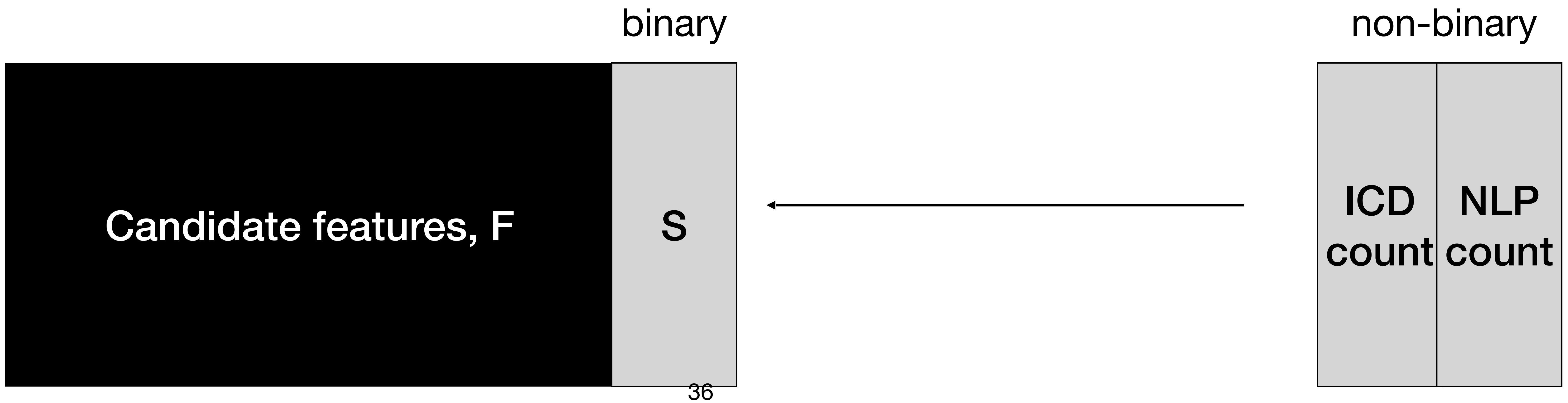


How to select features without Y?

- Data we have for feature selection

create “silver-standard” labels, S

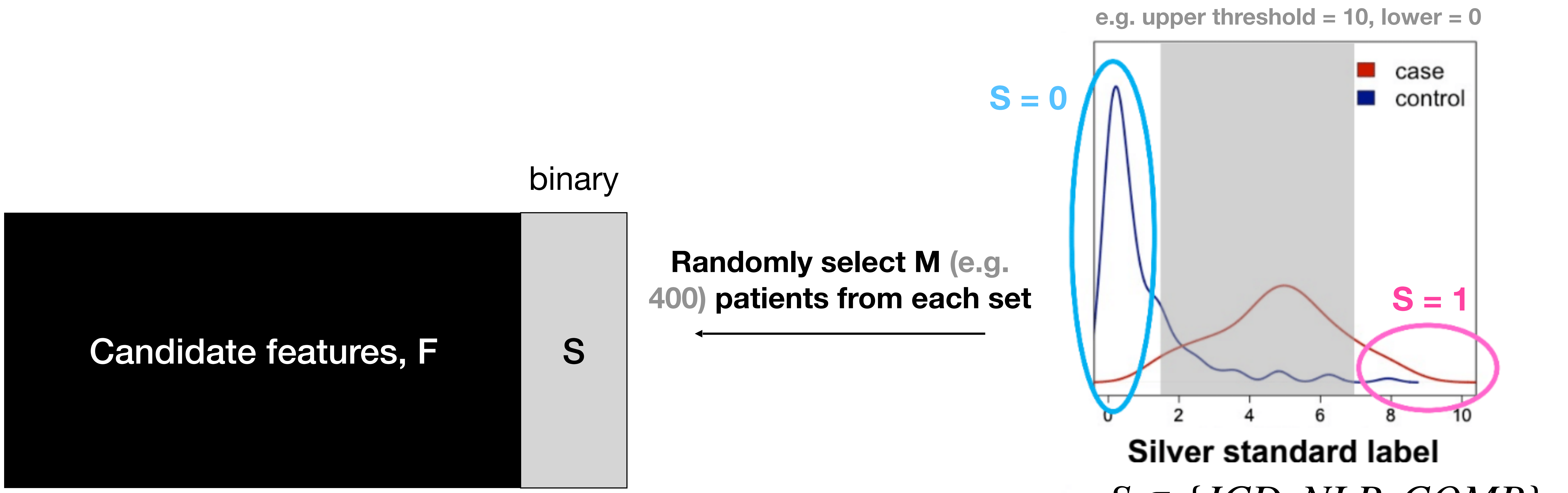
  - **main counts** of ICD-9 codes (codes of all subtypes)
  - **main counts** of NLP (UMLS concepts)
  - candidate features pass the 2 steps, F_{cand}



Intuition behind the surrogate selection

- Our goal is to identify a subset of F that is predictive of Y
- Y can be inferred from S by
    - Patients with **high** main ICD-9 or NLP counts generally have the phenotypes
    - Patients with **extremely** low counts are unlikely to have the phenotype
- Can we identify features related to Y with those related to S?

How to create binary silver-standard labels?



Randomly select M (e.g. 400) **patients from each set**

Silver standard label

$$S \in \{ICD, NLP, COMB\}$$

(See detailed choices of S from papers)

Surrogate selection

Model fitting details

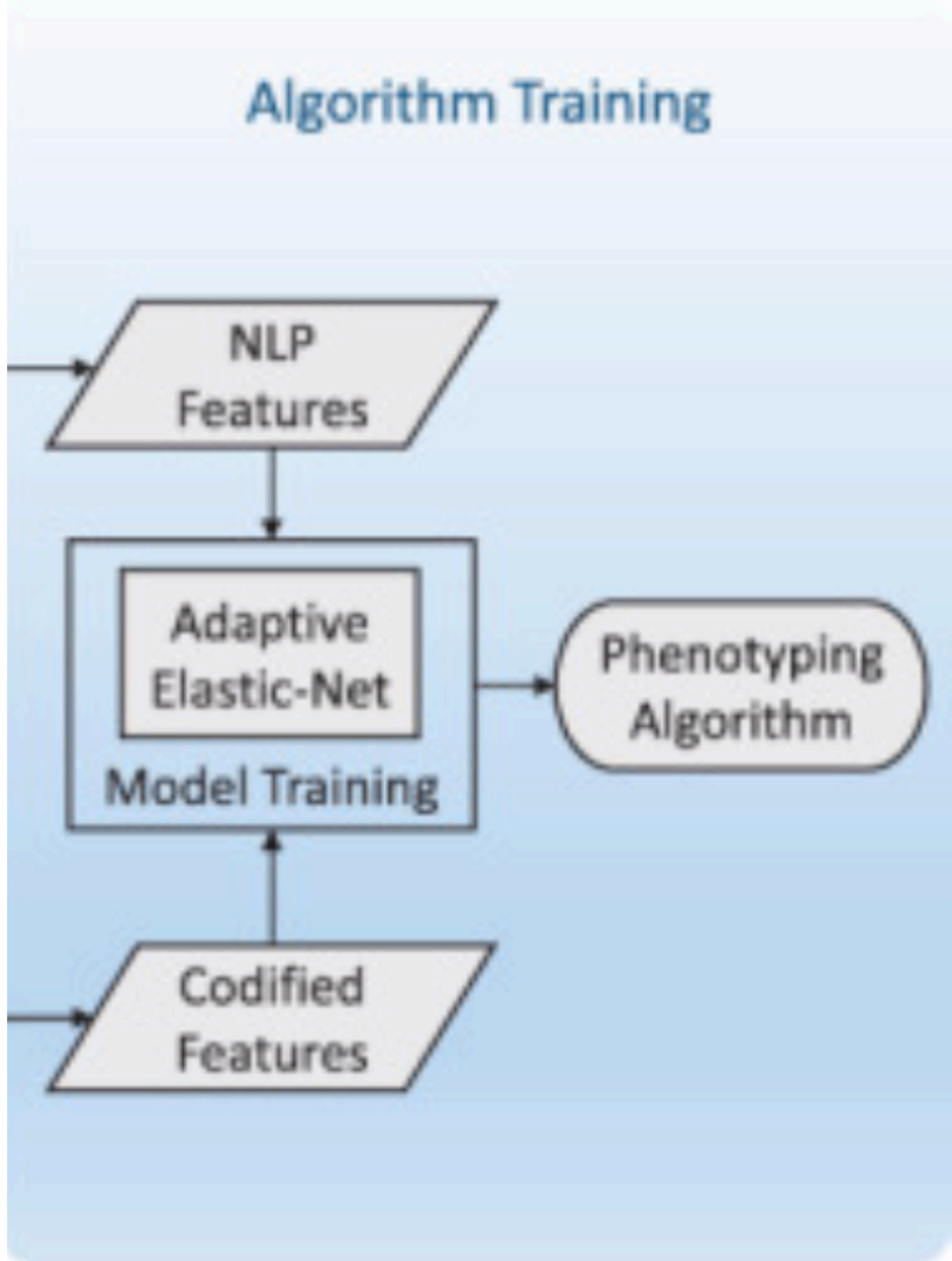
- Transform F_{cand} using $x \rightarrow log(x+1)$
- Adaptive elastic-net penalised logistic regression model S against $F_{transform}$
    - When S_{ICD} as response, exclude main ICD counts in the predictors; and so on
    - Tuning parameters choosing via BIC
- Repeat many times
- Include features $\neq 0$ at least 50% of the time

Short discussion

- Reasons for transforming the counts?
- Reasons for using adaptive elastic-net?
- Reasons for repeating?
- Reasons for using S_{COMB}?
- How to select a good surrogate, S?

4. Algorithm training

- Features
  - Selected NLP features
  - Codified features (total number of notes, age, gender, etc)
- Gold-standard labels



Algorithm evaluation

Data descriptions

- 4 phenotypes
  - Coronary artery disease (CAD), rheumatoid arthritis (RA), Crohn’s disease (CD), and ulcerative colitis (UC)
- 2 datamarts from Partners HealthCare
  - RA datamart: 46 568 patients with at least 1 ICD-9 codes of RA and other inflammatory polyarthropathies or had been tested for a diagnostic marker for RA
    - 435 gold-standard labels for RA, 758 for CAD
  - Inflammatory bowel disease (IBD) datamart: 34 033 patients with at least 1 ICD-9 codes of regional enteritis or ulcerative enterocolitis
    - 600 gold-standard labels for UC and CD, respectively

Evaluation metrics

- Out-of-sample accuracy
  - Metrics: area under the receiver operating characteristic curve (AUC) and F-score
  - At the 95% specificity level
- Size of training set: n = 100, 150, 200, 250, and 300
- Size of evaluation set: the rest of the labels
- (Stably) estimates by averaging the results randomly sampled 200 times

Results

Different combination of building blocks

- SAFE selects fewer features than AFEP and domain experts

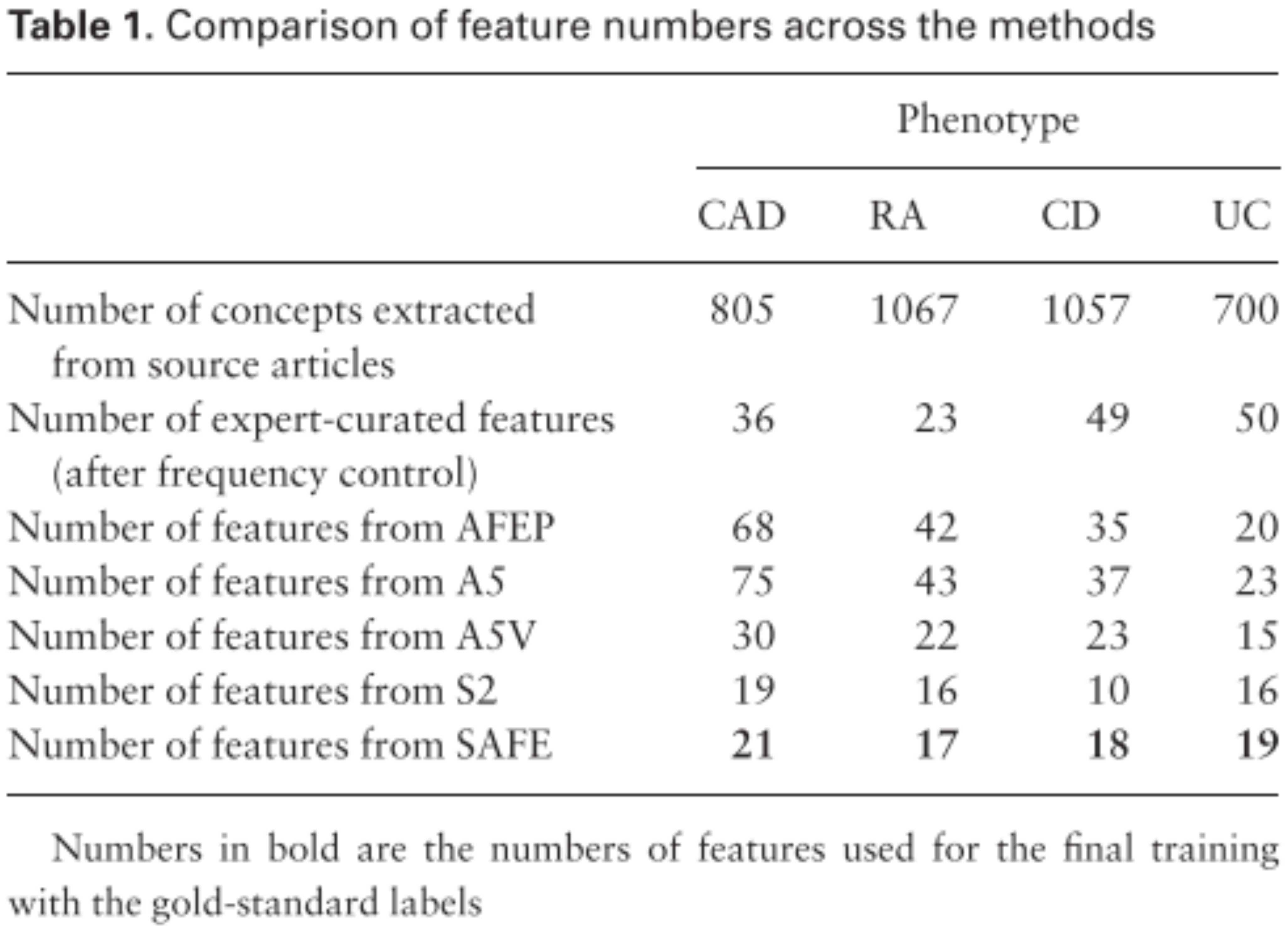


Table 1. Comparison of feature numbers across the methods

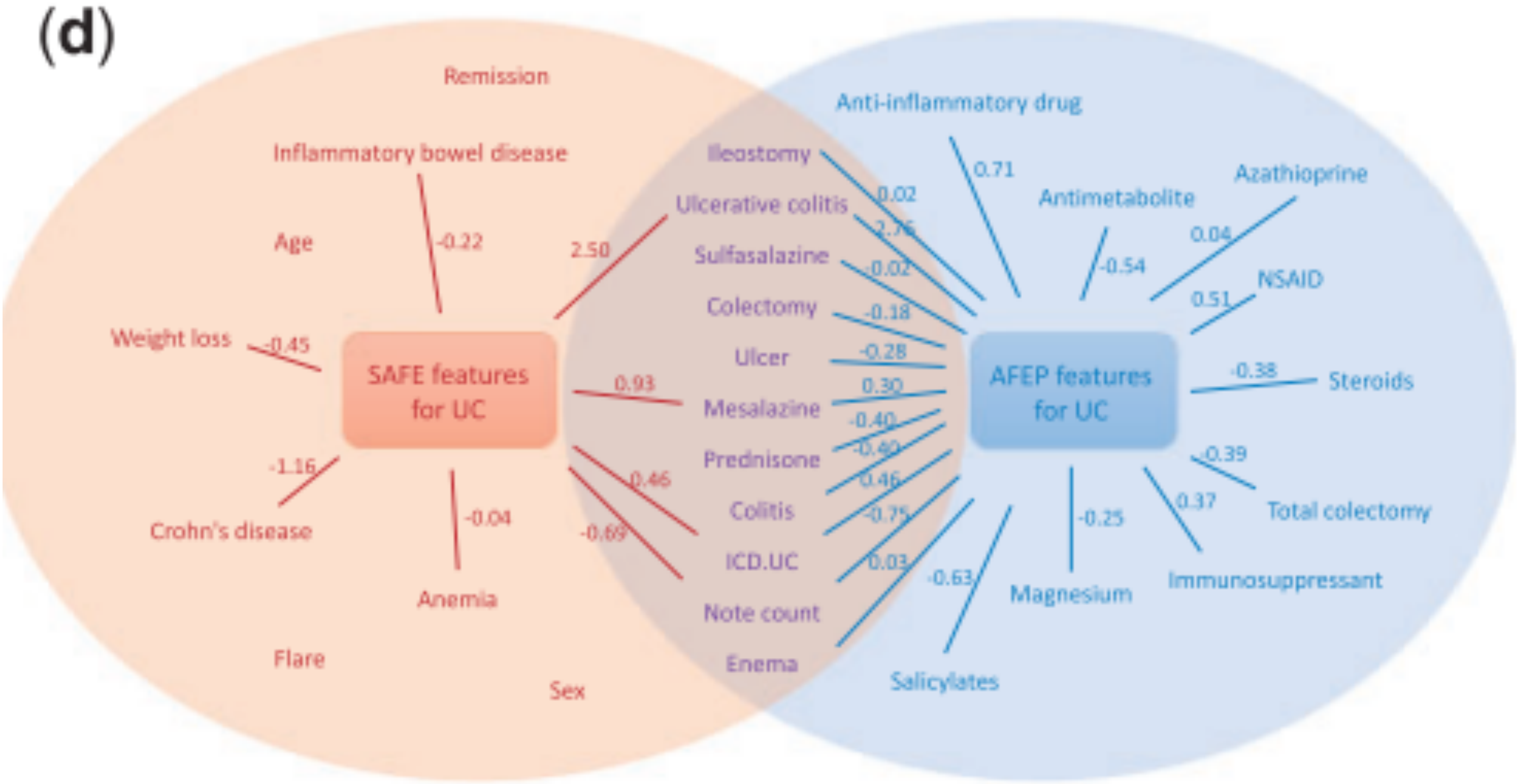
	Phenotype			
	CAD	RA	CD	UC
Number of concepts extracted from source articles	805	1067	1057	700
Number of expert-curated features (after frequency control)	36	23	49	50
Number of features from AFEP	68	42	35	20
Number of features from A5	75	43	37	23
Number of features from A5V	30	22	23	15
Number of features from S2	19	16	10	16
Number of features from SAFE	**21**	**17**	**18**	**19**

Numbers in bold are the numbers of features used for the final training with the gold-standard labels

Results

Compare SAFE and AFEP

- SAFE select more clinically meaningful features
    - AFEP missed “Crohn’s disease” and “weight loss”, expert missed “weight loss”
    - “Crohn’s disease” is a differential diagnosis of UC
    - “Weight loss” is a common symptom for CD, but not for UC

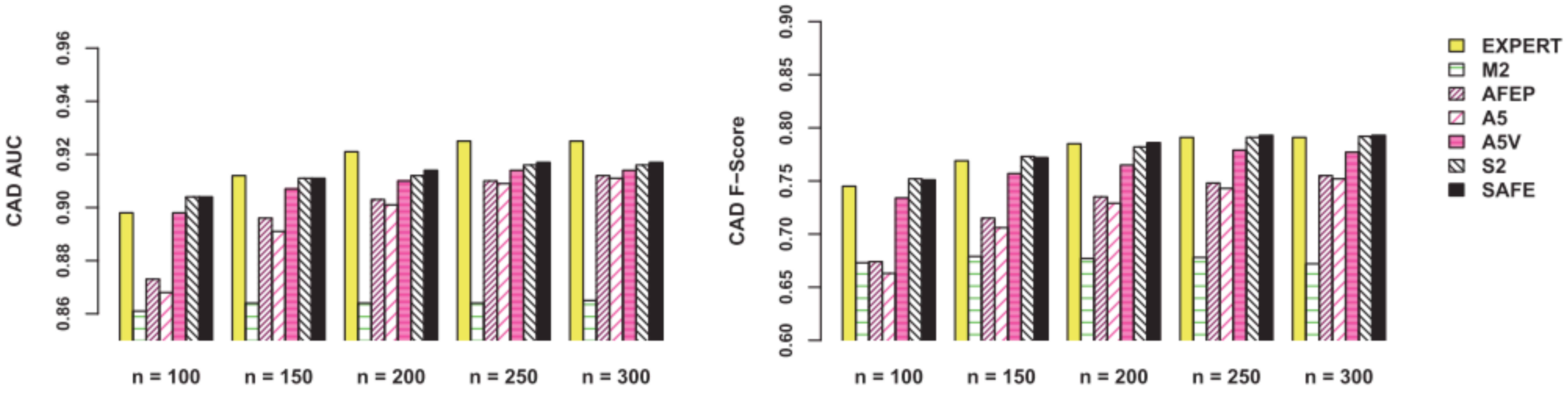


(**C**) CD, and (**D**) UC. Left and right circles include features from SAFE and AFEP

Results

Different combination of building blocks

- SAFE has higher out-of-sample AUC and F-scores than AFEP
- SAFE has comparable performance to expert curation

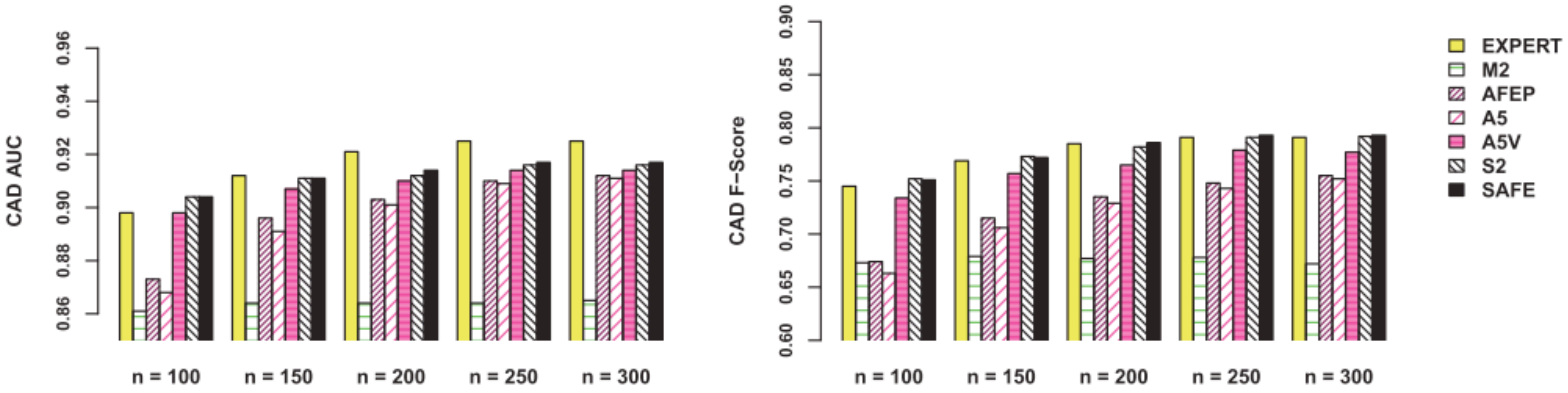


*Note: Expert curation has slightly higher AUCs when n is larger since expert created a feature covering CAD-specific procedures

Results

Different combination of building blocks

- Advantages of using SAFE more evident when n is small
    - Since overfitting less concerning for larger n



*Note: Expert curation has slightly higher AUCs when n is larger since expert created a feature covering CAD-specific procedures

Results

Different combination of building blocks

- SAFE not sensitive to the choice of upper/lower threshold in defining S

Short discussion

Motivation of the paper

- What is the problem being solved?
- Why is it important?

Short discussion

Approach of the paper

- What methods were used and why?
- What datasets were used and why?

Short discussion

Results of the paper

- How well did the approach solve the problem with simulated and/or real data?
- How did the approach compare to other solutions?
- What conclusions can be drawn?

Short discussion

Contribution of the paper

- How does this work compare to previous work?
- What makes the paper “new” or “novel”?

Short discussion

Limitation of the paper

- What might the issues be in applying the approach to another dataset or problem?
- What results are missing from the paper?
- Are the author’ conclusions well-informed?

Thank you!