Development and Validation of Phenotype Classifiers across the OHDSI Network

MEHR KASHYAP, MARTIN SENEVIRATNE, JUAN M. BANDA, NIGAM SHAH

DEPARTMENT OF BIOMEDICAL INFORMATICS, STANFORD UNIVERSITY

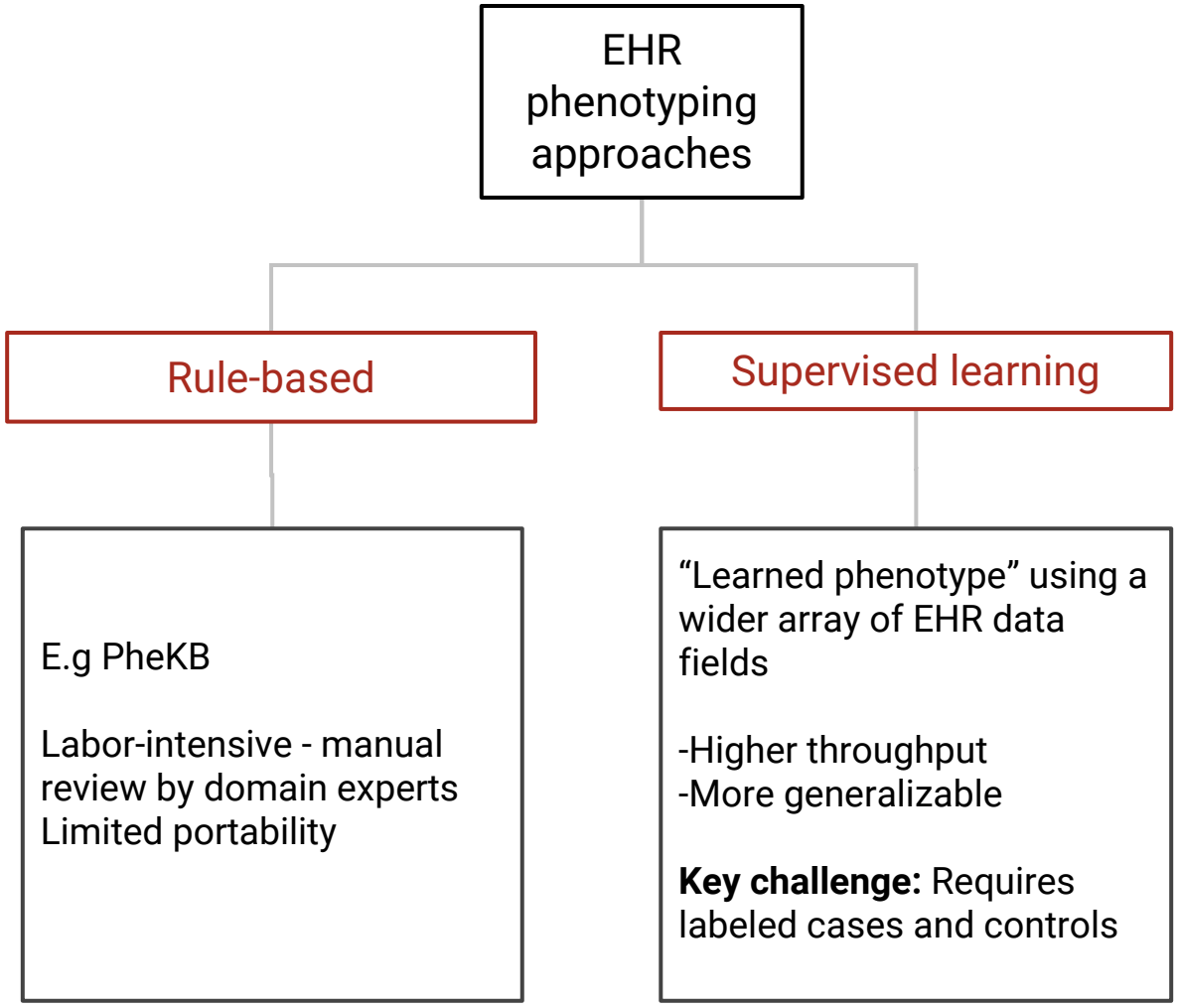
BACKGROUND: ELECTRONIC PHENOTYPING

- Electronic phenotyping – identifying patients within an electronic health record with a specific condition of interest

- Why is this important
  - Observational research
  - Pragmatic clinical trials
  - Quality improvement
  - Clinical decision support systems

- Key challenges
  - EHR heterogeneity, missingness, accuracy

BACKGROUND: ELECTRONIC PHENOTYPING



BACKGROUND: NOISE-TOLERANT LEARNING

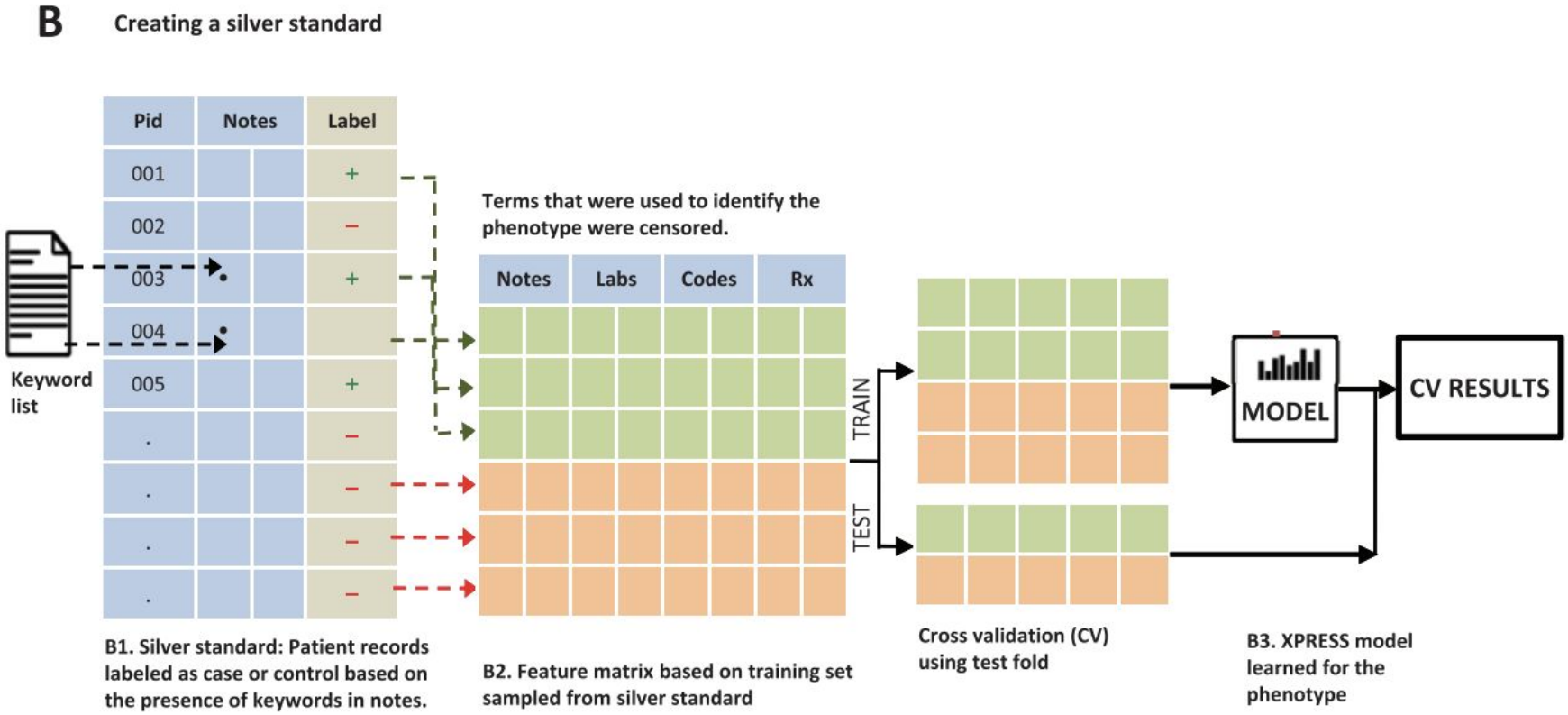
"silver-standard training sets"

a semi-supervised approach where training samples are labeled automatically (using an imperfect labeling heuristic) rather than by manual review

Intuition: noise-tolerant classifiers trained on imperfectly labeled data will abstract higher order properties of the phenotype beyond the original labeling heuristic

01	Halpern *et al.* (2016) *JAMIA*	• Anchor learning
02	Agarwal *et al.* (2016) *JAMIA*	• XPRESS
03	Banda *et al.* (2017) *AMIA Transl Summits*	• APHRODITE OHDSI package

BACKGROUND: XPRESS/APHRODITE



Source: Agarwal et al, 2016, JAMIA

BACKGROUND: APHRODITE VALIDATION TO DATE

- Previously locally validated for two phenotypes
  - Type 2 diabetes mellitus
  - Myocardial infarction

	Cases	Cont.	Acc.	Recall	PPV	Cases	Cont.	Acc.	Recall	PPV
Source	**Myocardial Infarction (MI)**					**Type 2 Diabetes Mellitus (T2DM)**				
OMOP/PheKB definition [2]	94	94	**0.87**	*0.91*	**0.84**	152	152	**0.92**	*0.88*	**0.96**
XPRESS 2	94	94	0.89	0.93	0.86	152	152	0.89	0.88	0.9
APHRODITE (750)	94	94	0.91	0.93	0.90	152	152	0.91	0.95	0.88
APHRODITE (1,500)	94	94	0.92	0.93	0.91	152	152	0.92	0.95	0.89
APHRODITE (10,000)	94	94	0.92	**0.94**	0.91	152	152	0.93	**0.96**	0.89

Source: Banda et al, 2017, AMIA Jt Summits Transl Sci Proc.

OPEN QUESTIONS FOR APHRODITE

Question 1

For situations in which precision is important, can we build high-precision classifiers using a precise labeling function, without relying on text data?

Question 2

Can we use APHRODITE classifiers at other OHSDI sites? How generalizable is the pipeline or the classifiers it creates?

Question 3

Can we assess classifier performance by comparing demographics of patients identified as cases across multiple sites?

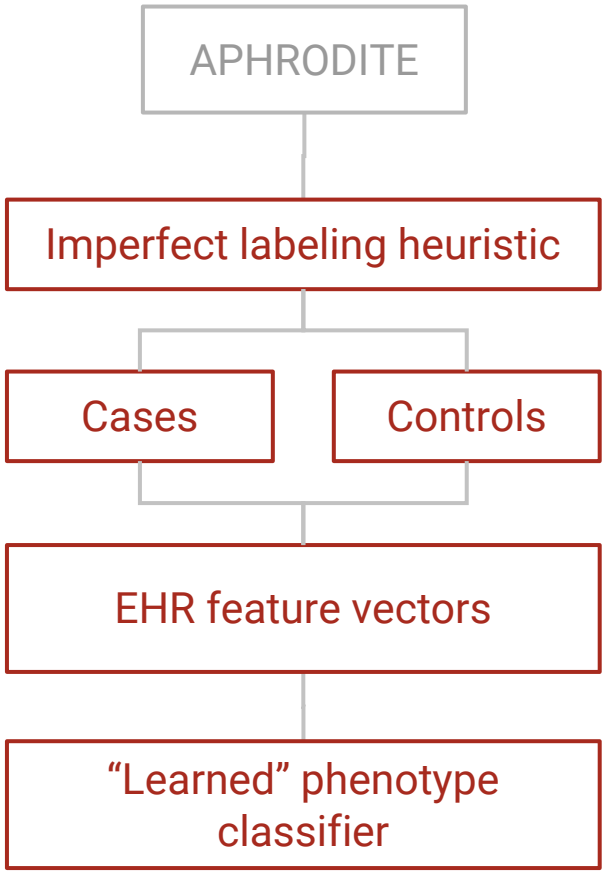
METHODS: DATA SOURCE

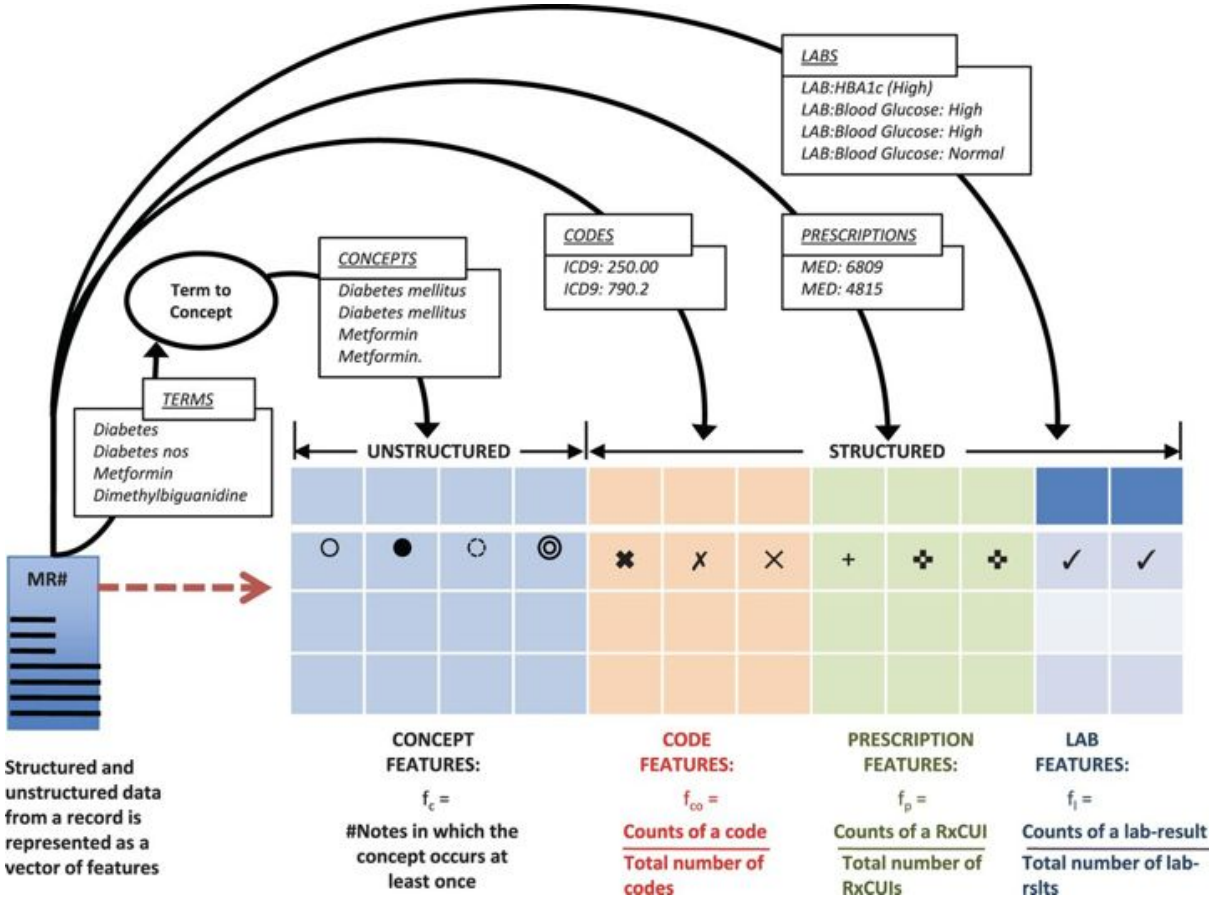
Patient extract from Stanford Medicine Research Data Repository

1.8 million patients – laboratory results, procedures, drug exposures, diagnosis codes

Mapped to OMOP Common Data Model v5

METHODS: CLASSIFIER BUILDING PIPELINE



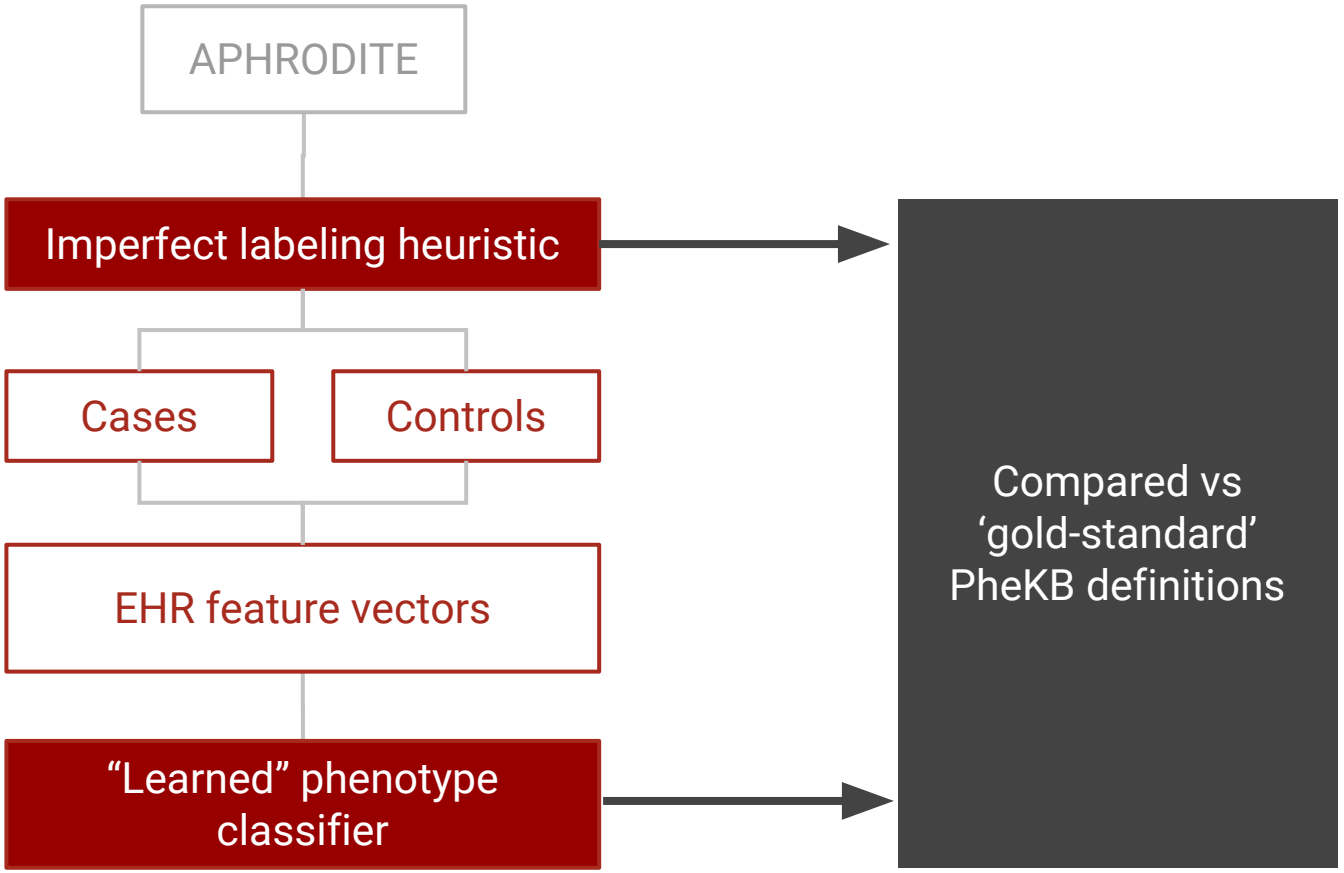


Source: Agarwal et al, 2016, JAMIA

METHODS: IMPERFECT LABELING HEURISTIC

- Labeling heuristic – **multiple mentions of disease specific code**
- Patients with **2+ mentions of relevant codes** considered cases
- **High precision, low recall**

METHODS: CLASSIFIER EVALUATION



EXPERIMENT 1: LOCAL RESULTS

Phenotype	Prevalence of cases in test set	Multiple mentions of SNOMED code			APHRODITE classifier		Recall boost using classifier	Precision loss using classifier
		No. of mentions	Recall	Precision	Recall	Precision		
Appendicitis	0.05	2	0.31	1.00	0.97	0.99	**+0.66**	**-0.01**
T2DM	0.14	4	0.24	0.99	0.60	0.91	**+0.36**	**-0.08**
Cataracts	0.17	4	0.07	0.97	0.63	0.93	**+0.56**	**-0.04**
Heart Failure	0.02	4	0.33	0.94	0.99	0.56	**+0.66**	**-0.38**
Abdominal Aortic Aneurysm	0.04	4	0.22	0.99	0.53	0.97	**+0.31**	**-0.02**
Epileptic seizure	0.02	4	0.06	1.00	0.22	0.94	**+0.17**	**-0.06**
PAD	0.05	4	0.18	0.98	0.91	0.91	**+0.72**	**-0.07**
Adult onset obesity	0.36	4	0.20	1.00	0.29	0.91	**+0.09**	**-0.09**
Glaucoma	0.01	4	0.08	1.00	0.22	0.88	**+0.14**	**-0.12**
VTE	0.01	4	0.03	1.00	0.69	0.22	**+0.66**	**-0.78**

EXPERIMENT 1: DISCUSSION

- Classifiers retain **high precision + recall boost** relative to the labeling heuristic:
  - Learning algorithm able to generalize such that final model has higher recall than original labeling function
  - Suitable for phenotyping tasks where precise cohorts required
- **Real-world prevalence used**
- **Ten phenotype classifiers developed** – construction time ~1.5 hrs/phenotype
- This method **does not rely on textual data**

EXPERIMENT 2: PERFORMANCE ACROSS OHDSI NETWORK

Development site	Stanford			Columbia	SNUBH	Stanford			Columbia	SNUBH
Validation site	Stanford	Columbia	SNUBH	Stanford	Stanford	Stanford	Columbia	SNUBH	Stanford	Stanford
Phenotype	Recall					Precision				
Appendicitis	0.97	0.9	0.09	0.97	0.52	0.99	0.9	0.56	0.92	0.13
T2DM	0.6	0.63	0.77	0.45	0.67	0.91	0.86	0.75	0.94	0.51
Cataracts	0.63	0.45	0.84	0.2	0.35	0.93	0.79	0.85	0.96	0.42
Heart Failure	0.99	0.97	0.8	0.01	0.71	0.56	0.67	0.75	1	0.11
Abdominal Aortic Aneurysm	0.53	0.24	0.54	0.68	0.33	0.97	0.75	0.87	0.96	0.13
Epileptic seizure	0.22	0.3	0.28	0.57	0.46	0.94	0.87	0.55	0.74	0.08
PAD	0.91	0.89	0.57	0.89	0.46	0.91	0.87	0.68	0.91	0.24
Adult onset obesity	0.29	0.33	0.07	0.04	0.39	0.91	0.93	0.73	0.9	0.68
Glaucoma	0.22	0.18	0.11	0.65	0.22	0.88	0.78	0.65	0.66	0.06
VTE	0.69	0.34	0.2	0.05	0.46	0.2	0.71	0.83	0.63	0.05

EXPERIMENT 2: DISCUSSION

- Models trained at Stanford work very well at Columbia and reasonably well at SNUBH
- Conversely, models trained at Columbia work well at Stanford
- However, models trained at SNUBH do not work well at Stanford

EXPERIMENT 3: COHORT DEMOGRAPHICS

Phenotype	Characteristic	Stanford	Columbia	SNUBH
Appendicitis	Male (%)	0.52	0.52	0.51
	Mean age	36.36	34.14	44.49
T2DM	Male (%)	0.53	0.44	0.56
	Mean age	63.65	67.77	66.47
Cataracts	Male (%)	0.49	0.37	0.43
	Mean age	67.02	75.54	68.22
Heart Failure	Male (%)	0.56	0.50	0.48
	Mean age	65.58	70.69	72.9
Abdominal Aortic Aneurysm	Male (%)	0.75	0.68	0.79
	Mean age	76.60	78.24	77.41
Epileptic seizure	Male (%)	0.43	0.47	0.53
	Mean age	43.83	45.57	32.27
PAD	Male (%)	0.54	0.48	0.69
	Mean age	70.29	73.29	66.22
Adult onset obesity	Male (%)	0.44	0.30	0.61
	Mean age	57.73	45.18	35.9
Glaucoma	Male (%)	0.50	0.40	0.58
	Mean age	67.96	80.10	62.97
VTE	Male (%)	0.53	0.39	0.51
	Mean age	38.67	65.02	68.46
All patients	**Male (%)**	**0.45**	**0.45**	**0.48**
	Mean age	**39.40**	**39.90**	**40.41**

EXPERIMENT 3: DISCUSSION

- Phenotype **models with high precision identify similar patients** across different sites
- Comparing demographics of cases identified by classifiers across different sites may serve as a **proxy for model validation**
  - This could be useful in the absence of manually labeled or rule-based evaluation sets
- Unclear when variation in demographics simply represents underlying patient differences

CONCLUSIONS & NEXT STEPS

- Classifiers built using precise training data retain high precision while improving recall relative to the labeling heuristic
  - Classifiers are able to generalize such that final model has higher recall than original labeling function
- Classifiers generally perform well across OHDSI sites though this may be limited by regional discrepancies in mapping of EHR data
  - Should we be sharing "recipes" rather than completed models?
- In the absence of manually labeled or rule-based evaluation sets, comparing demographics of cases identified by classifiers may serve as a proxy for model validation

ACKNOWLEDGEMENTS

- Martin Seneviratne
- Juan Banda
- rohit vashisht
- Ken Jung
- Steven Bagley
- Stephen Pfohl
- Vibhu agarwal
- Nigam Shah
- Shah lab members

OHDSI Collaborators

- Thomas Falconer
- Sooyoung Yoo
- Borim Ryu
- OHDSI Community

Contact: mkashyap@stanford.edu

Stanford University