Sentinel

Welcome to the Sentinel Innovation and Methods Seminar Series

The webinar will begin momentarily

Please visit www.sentinelinitiative.org for recordings of past sessions and details on upcoming webinars.

Note: closed-captioning for today's webinar will be available on the recording posted at the link above.

Sentinel

Data-driven Automated Classification Algorithms for Acute Health Conditions: Applying PheNorm to COVID-19 Disease

Joshua C. Smith, PhD
Department of Biomedical Informatics
Vanderbilt University Medical Center

25 March 2024

Disclaimer

- This work was supported by Task Order **75F40119F19002** under Master Agreement **75F40119D10037** from the U.S. Food and Drug Administration (FDA).

- The views expressed in this presentation represent those of the presenter and do not necessarily represent the official views of the U.S. FDA.

JAMIA
A SCHOLARLY JOURNAL OF INFORMATICS IN HEALTH AND BIOMEDICINE

AMIA
INFORMATICS PROFESSIONALS. LEADING THE WAY.

JOURNAL ARTICLE FEATURED

Data-driven automated classification algorithms for acute health conditions: applying PheNorm to COVID-19 disease

Joshua C Smith, PhD ✉, Brian D Williamson, PhD, David J Cronkite, MS, Daniel Park, BS, Jill M Whitaker, MSN, Michael F McLemore, BSN, Joshua T Osmanski, MS, Robert Winter, BA, Arvind Ramaprasan, MS, Ann Kelley, MHA, Mary Shea, MA, Saranrat Wittayanukorn, PhD, Danijela Stojanovic, PharmD, PhD, Yueqin Zhao, PhD, Sengwee Toh, ScD, Kevin B Johnson, MD, MS, David M Aronoff, MD, David S Carrell, PhD

Journal of the American Medical Informatics Association, Volume 31, Issue 3, March 2024, Pages 574–582, https://doi.org/10.1093/jamia/ocad241

Published: 18 December 2023 **Article history** ▾

Introduction

- Sentinel is the U.S. FDA's active medical product safety surveillance system utilizing electronic healthcare records (EHRs) and claims data.
- One of the goals of the **Sentinel Innovation Center** is to develop, implement, and evaluate methods that incorporate unstructured EHR data to improve the performance of computable phenotype algorithms used to capture health outcomes relevant to medical product safety surveillance.

Introduction

- Sentinel Innovation Center (IC) Demonstration Project to integrate **unstructured EHR data** into Sentinel
- "Advancing scalable natural language processing approaches for unstructured electronic health record data"
- In this study, we evaluated an **automated phenotyping** method (PheNorm) applied to an acute condition, COVID-19 disease, to investigate its feasibility for rapid phenotyping and use in post-market safety studies.

Background

- **Phenotyping algorithms** are used in healthcare, epidemiological studies, and public health surveillance to distinguish between cases and non-cases.
- Methods range from the use of International Classification of Disease (ICD 9/10 codes) to the presence of multiple codes, medications, or laboratory results.
- These algorithms have typically been developed and validated using time-intensive expert curation and manually annotated gold-standard training sets, which result in **high costs**, **long development timelines**, and **limited scalability**.

Manual Feature Engineering

= Clinicians = Informaticists

Identify

Propose targets

Review knowledge

PubMed

Propose codes

ICD-10 www.cms.gov/ICD10

LOINC

cpt Current Procedural Terminology

RxNorm

Epic

Cerner

Propose terms

EHR

MAYO CLINIC

UMLS

Define

Review code lists

Validate code usage

Assemble corpus

EHR

Validate NLP

NIH MetaMap

NLP

Specify logic

Implement

Write code

sas

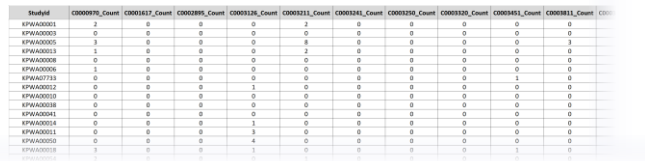
Create NLP

NIH MetaMap

NLP

Perform QC

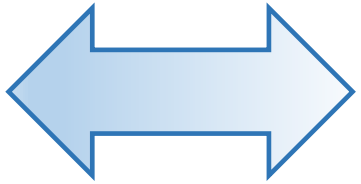
Assemble datasets



Automated Modeling: Motivation

Manual development

- *Expert*-driven
- *Manual* engineering
- Heavy reliance on *gold standard labels*
- Substantial operator dependence
- Slow



Automated development

- Data-driven
- Automated engineering
- Heavy reliance on *silver standard labels*
- Reduced operator dependence
- Fast

Automated Modeling: Approach

Principles:

1. **Clinical text** is the primary data source
2. **Published knowledge** provides expertise
3. **Data-driven** feature engineering & modeling

AFEP

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

RECEIVED 24 October 2014
REVISED 25 February 2015
ACCEPTED 24 March 2015
PUBLISHED ONLINE FIRST 30 April 2015

Sheng Yu[1,2,3,*], Katherine P Liao[2,3], Stanley Y Shaw[4], Vivian S Gainer[5], Susanne E Churchill[5], Peter Szolovits[6], Shawn N Murphy[4,5], Isaac S Kohane[3,7], Tianxi Cai[8]

ABSTRACT

Objective Analysis of narrative (text) data from electronic health records (EHRs) can improve population-scale phenot research. Currently, selection of text features for phenotyping algorithms is slow and laborious, requiring extensive an main experts. This paper introduces a method to develop phenotyping algorithms in an unbiased manner by automati informative features, which can be comparable to expert-curated ones in classification accuracy.

Materials and methods Comprehensive medical concepts were collected from publicly available knowledge source fashion. Natural language processing (NLP) revealed the occurrence patterns of these concepts in EHR narrative note informative features for phenotype classification. When combined with additional codified features, a penalized lo trained to classify the target phenotype.

Results The authors applied our method to develop algorithms to identify patients with rheumatoid arthritis and c among those with rheumatoid arthritis from a large multi-institutional EHR. The area under the receiver operating cl classifying RA and CAD using models trained with automated features were 0.951 and 0.929, respectively, compar 0.929 by models trained with expert-curated features.

Discussion Models trained with NLP text features selected through an unbiased, automated procedure achieved com curacy than those trained with expert-curated features. The majority of the selected model features were interpretable

Conclusion The proposed automated feature extraction method, generating highly accurate phenotyping algorithms significant step toward high-throughput phenotyping.

INTRODUCTION

Electronic health record (EHR) adoption has increased dramatically in recent years. By 2013, 59% of private acute care hospitals in the United States had adopted an EHR system, up from 9% in 2008.[1] Secondary use of EHR data has emerged as a powerful approach for a variety of biomedical research, including comparative effectiveness and stratifying patients for risk of comorbidities or adverse outcomes.[2–10] More recently, the linking of genotype and biomarker data to EHR data has facilitated translational studies, such as genetic association studies.[11–17] Compared to conventionally assembled epidemiologic and genomic cohorts that require individual patient recruitment, EHR-based studies can provide large sample sizes at a lower cost and

narrative notes such as physician notes, t or pathologic studies, or hospital discha provide a rich source of complementary i processing (NLP) can efficiently extract c Occurrences of terms of clinical concept and also used as features for algorithm ing algorithms that use both codified and accuracy relative to algorithms using codi 9 billing codes).[19–22]

Today, algorithms that identify a des structed in two rather different ways. The ing on human expertise to suggest a logi

SAFE

Journal of the American Medical Informatics Association, 24(e1), 2017, e143–e149
doi: 10.1093/jamia/ocw135
Advance Access Publication Date: 15 September 2016
Research and Applications

Research and Applications

Surrogate-assisted feature extraction for high-throughput phenotyping

Sheng Yu,[1,2] Abhishek Chakrabortty,[3] Katherine P Liao,[4] Tianrun Cai,[5] Ashwin N Ananthakrishnan,[6] Vivian S Gainer,[7] Susanne E Churchill,[8] Peter Szolovits,[9] Shawn N Murphy,[7,10] Isaac S Kohane,[8] and Tianxi Cai[3]

[1]Center for Statistical Science, Tsinghua University, Beijing, China, [2]Department of Industrial Engineer Beijing, China, [3]Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, Massac Rheumatology, Brigham and Women's Hospital, Boston, Massachusetts, USA, [5]Department of Radiolog Hospital, Boston, Massachusetts, USA, [6]Division of Gastroenterology, Massachusetts General Hospital, USA, [7]Research IS and Computing, Partners HealthCare, Charlestown, Massachusetts, USA, [8]Departm matics, Harvard Medical School, Boston, Massachusetts, USA, [9]Computer Science and Artificial Intellig chusetts Institute of Technology, Cambridge, Massachusetts, USA, and [10]Department of Neurology, Hospital, Boston, Massachusetts, USA

Corresponding Author: Sheng Yu, Center for Statistical Science, Tsinghua University, Beijing, China edu.cn

PheNorm

Journal of the American Medical Informatics Association, 25(1), 2018, 54–60
doi: 10.1093/jamia/ocx111
Advance Access Publication Date: 3 November 2017
Research and Applications

Research and Applications

Enabling phenotypic big data with PheNorm

Sheng Yu,[1,2] Yumeng Ma,[3] Jessica Gronsbell,[4] Tianrun Cai,[5] Ashwin N Ananthakrishnan,[6] Vivian S Gainer,[7] Susanne E Churchill,[8] Peter Szolovits,[9] Shawn N Murphy,[7,10] Isaac S Kohane,[8] Katherine P Liao,[11] and Tianxi Cai[4]

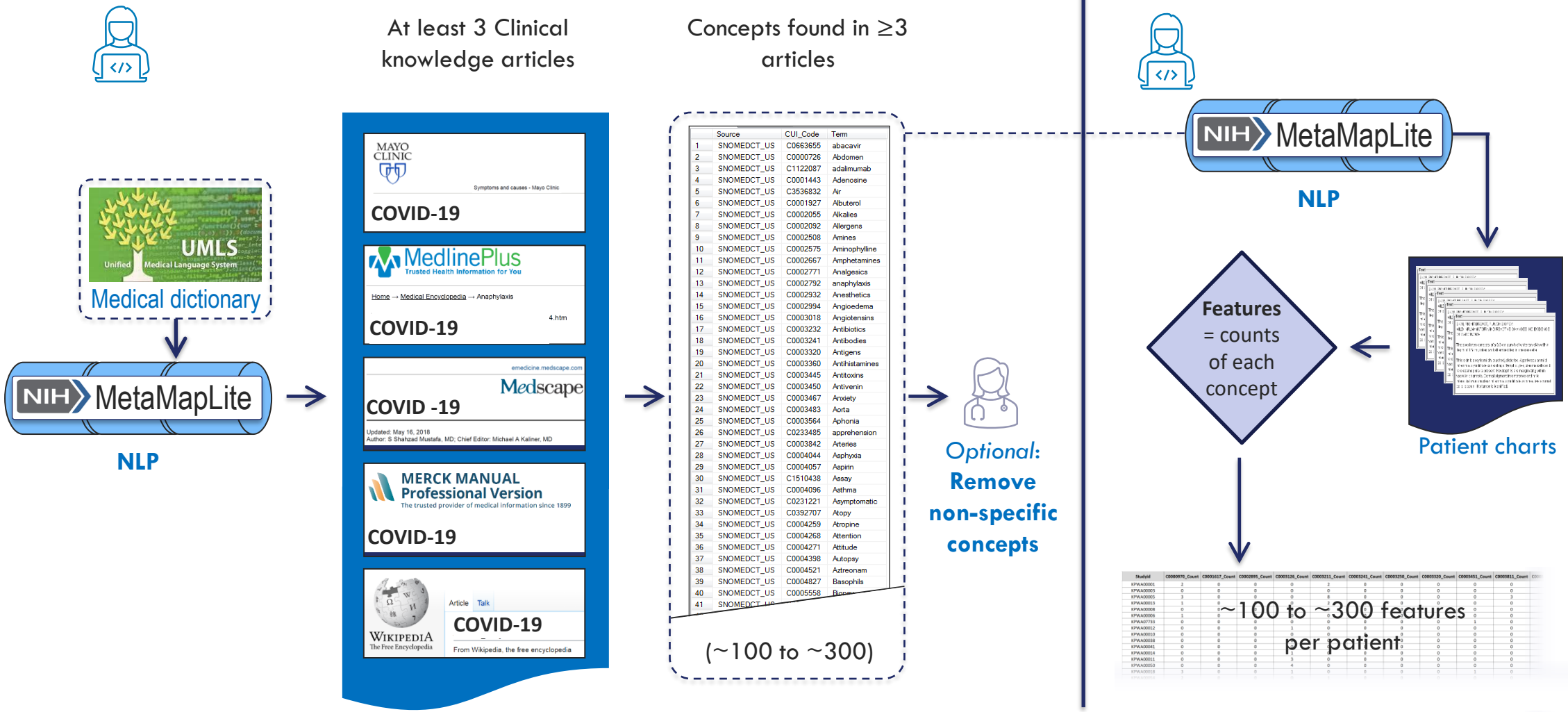
[1]Center for Statistical Science, Tsinghua University, Beijing, China, [2]Department of Industrial Engineering, Tsinghua University, Beijing, China, [3]Department of Mathematical Sciences, Tsinghua University, Beijing, China, [4]Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA, USA, [5]Department of Radiology, Brigham and Women's Hospital, Boston, MA, USA, [6]Division of Gastroenterology, Massachusetts General Hospital, Boston, MA, USA, [7]Research Information Science and Computing, Partners HealthCare, Charlestown, MA, USA, [8]Department of Biomedical Informatics, Harvard Medical School, Bos-

Automated Feature Engineering

= Clinicians = Informaticists

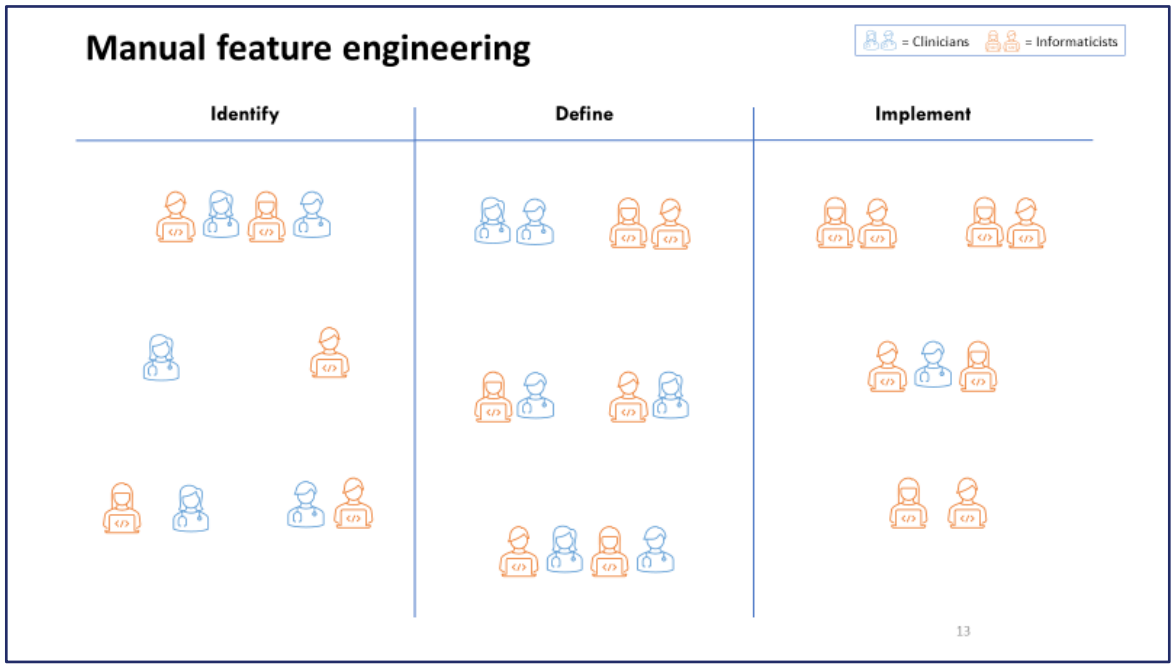
Identify & Define*

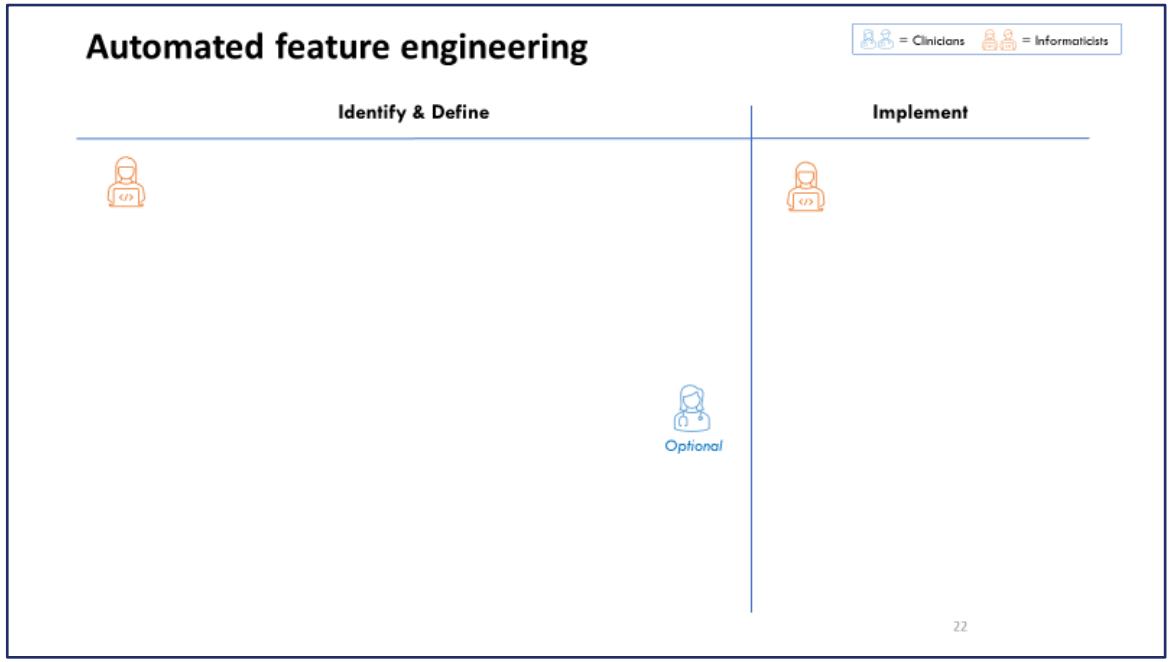
Implement



* Yu et al. Toward high-throughput phenotyping: unbiased automated feature extraction and selection from knowledge sources. JAMIA 2015

Feature Engineering: Manual vs. Automated





Advantages of automation:

- Short development time
- Low/no expenditure for domain expertise
- Reduced operator dependence
- Replicable

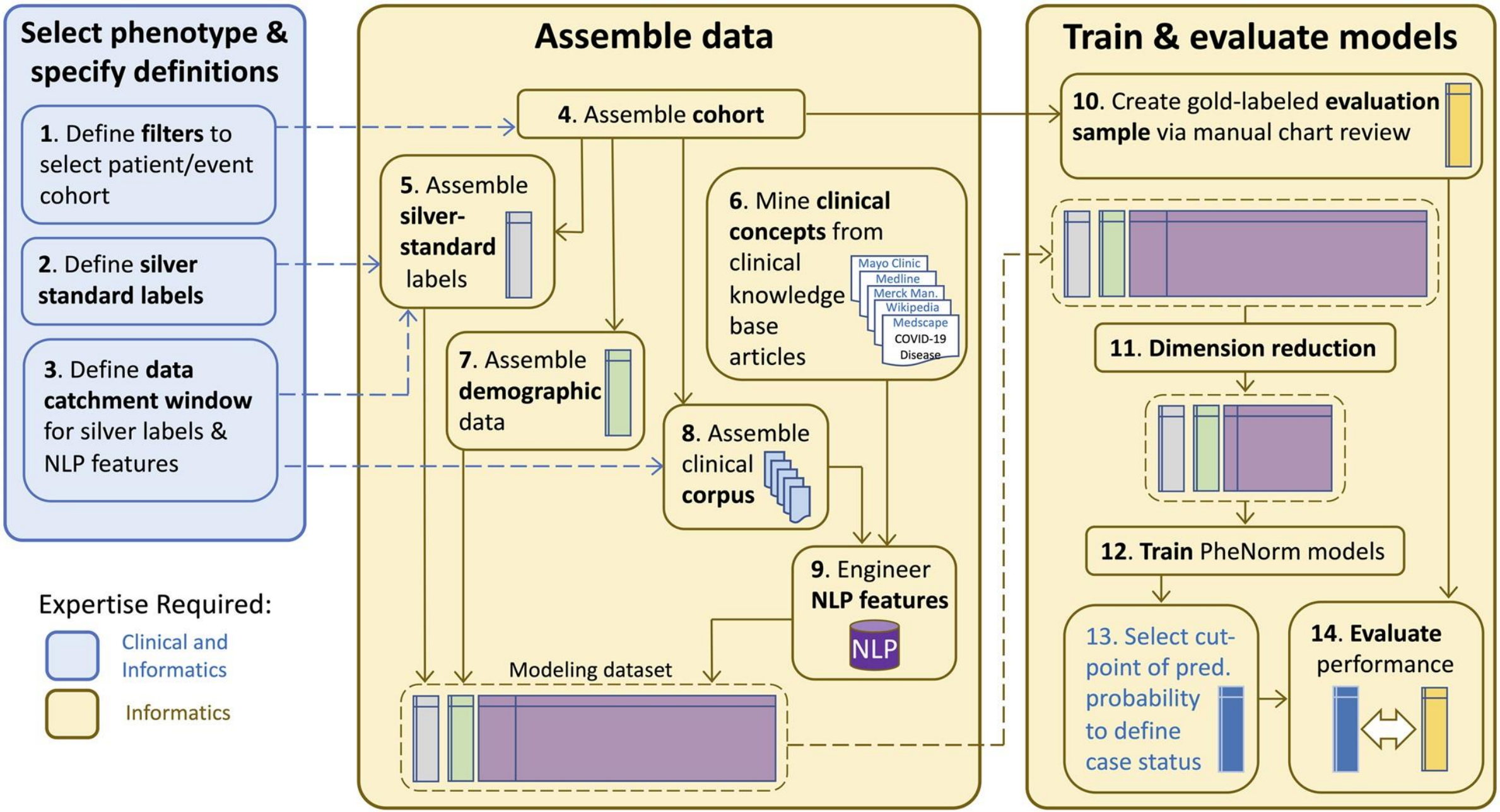
Will it work? As a starting point? As an overall solution?

Methods

Automated Modeling: Approach

- Developed by Yu, et al., **PheNorm** has been demonstrated to perform well outside Sentinel for chronic health conditions, but little was known about its performance in acute conditions.
- PheNorm is a general-purpose **automated** approach to creating computable phenotype algorithms based on natural language processing (NLP) and machine learning.
- PheNorm estimates each patient's probability of being a true case using **silver-standard labels** (readily available approximations for true case status) and NLP-derived features extracted from clinical notes.

PheNorm


Select phenotype & specify definitions
1. Define filters to select patient/event cohort
2. Define silver standard labels
3. Define data catchment window for silver labels & NLP features
Assemble data
4. Assemble cohort
5. Assemble silver-standard labels
6. Mine clinical concepts from clinical knowledge base articles
Mayo Clinic
Medline
Merck Man.
Wikipedia
Medscape
COVID-19
Disease
7. Assemble demographic data
8. Assemble clinical corpus
9. Engineer NLP features
NLP
Modeling dataset
Train & evaluate models
10. Create gold-labeled evaluation sample via manual chart review
11. Dimension reduction
12. Train PheNorm models
13. Select cut-point of pred. probability to define case status
14. Evaluate performance
Expertise Required:
Clinical and Informatics
Informatics

PheNorm Applied to COVID-19 Disease

- **Coronavirus disease 2019** (COVID-19) was first identified in December 2019. During the pandemic, diagnostic guidelines, laboratory testing, coding practices, and treatment options changed rapidly.
- We developed a phenotyping algorithm for **symptomatic COVID-19**
  - Diagnostic codes for COVID-19 have been shown* to have low accuracy, which may be due to both over-coding and under-coding.
  - Since we were interested in symptomatic disease, evidence of infection alone was insufficient since many patients who tested positive were asymptomatic.

*Lynch, et al. Positive predictive value of COVID-19 ICD-10 diagnosis codes across calendar time and clinical setting. Clin Epidemiol. 2021.

Study Cohort

- This study was performed at **Vanderbilt University Medical Center** (VUMC) and **Kaiser Permanente Washington** (KPWA).
- We identified cohorts of potential COVID-19 patients from April 2020 through March 2021 at each site.
- Cohorts included all patients with encounters accompanied by structured EHR features found to be strongly associated with COVID:
  - Six ICD-10-CM diagnosis codes for COVID-19 and related complications
  - 43 other codes (diagnoses, problems, procedures, medications, labs)
- The VUMC cohort included both inpatient and outpatient encounters; the KPWA cohort included outpatient only.

Index Date and Exclusion Criteria

- In PheNorm, a fixed data catchment period anchored to a patient-specific index date identifies data used to operationalize silver labels and features.
- The earliest encounter for each patient with any structured evidence of COVID-19 disease was used as **index date**.
- Our catchment period was **index date ±30 days**, which we consider likely to include relevant and exclude unrelated information.
- Eligible patients included adults (age 18+ years) with at least one encounter and ≥1000 characters of clinical text.

Silver Labels

- PheNorm replaces scarce, costly gold-standard data with silver-standard data during model training.
- As silver labels are imperfect representations of true-case status, the PheNorm methods suggests considering multiple alternative versions of silver labels.
- We therefore used information from each patient's data catchment period to operationalize **4 silver-standard labels** that used either structured data or NLP-derived data.

Silver Labels

1. **Structured Label 1:** Count of calendar days with a COVID-19 diagnosis code (U07.1), including both outpatient visits and inpatient days
2. **Structured Label 2:** Count of calendar days with any of 6 COVID-19-related diagnosis codes: U07.1, J12.81, J12.82, B34.2, B97.21, B97.29
3. **NLP Label 1:** Count of the number of mentions of the term “COVID-19” in chart notes
4. **NLP Label 2:** Count of chart notes with an NLP-identified UMLS concept for COVID-19 disease (C5203670)

Features

- Machine learning models use features (variable, covariates) as **input** to produce an **output** based upon training data.
- Input **features** are usually based on structured data, such as diagnosis codes or laboratory values
- PheNorm's primary features are **NLP-extracted "clinical concepts"** mentioned in the unstructured text of clinical notes.
- We processed all clinical notes within a patient's catchment window using the **MetaMap Lite NLP tool** to identify clinical concepts mentioned in the text, represented using **UMLS Concepts.**

Structured Data

PheNorm uses primarily **NLP-extracted features** as the input to the predictive models, however, useful structured data can also be included as model features.

In this study, we operationalized **two** structured data features:

- patient sex (as captured in the EHR)
- patient age (in years)

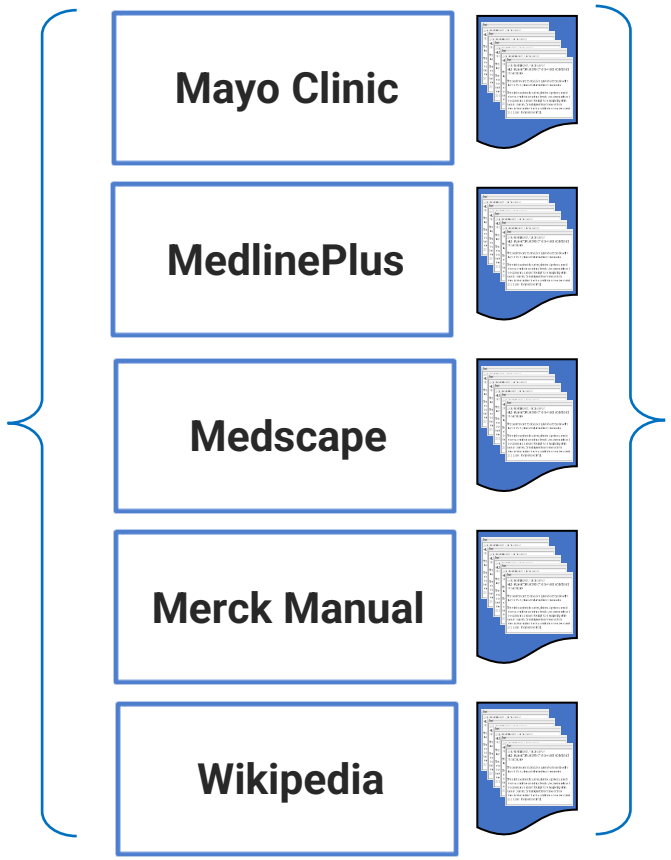
Feature Engineering

- However, “all clinical concepts” mentioned in patients’ notes is a very large features space.
- **Most of these concepts are likely uninformative.**
- Like most phenotyping algorithms, PheNorm limits the input features to those that are relevant to the Health Outcome of Interest.
- As described earlier, we utilized the AFEP approach to automated feature extraction to define a “dictionary” of relevant concepts.

Feature Engineering: NLP Dictionary Creation

Automating Feature Engineering (AFEP)

5 clinical knowledge base articles on a topic (*COVID-19*)



Corpus

MetaMap Lite NLP

	Source	CUI_Code	Term
1	SNOMEDCT_US	C0663655	abacavir
2	SNOMEDCT_US	C0000726	Abdomen
3	SNOMEDCT_US	C1122087	adalimumab
4	SNOMEDCT_US	C0001443	Adenosine
5	SNOMEDCT_US	C3536832	Air
6	SNOMEDCT_US	C0001927	Albuterol
7	SNOMEDCT_US	C0002055	Alkalies
8	SNOMEDCT_US	C0002092	Allergens
9	SNOMEDCT_US	C0002508	Amines
10	SNOMEDCT_US	C0002575	Aminophylline
11	SNOMEDCT_US	C0002667	Amphetamines
12	SNOMEDCT_US	C0002771	Analgesics
13	SNOMEDCT_US	C0002792	anaphylaxis
14	SNOMEDCT_US	C0002932	Anesthetics
15	SNOMEDCT_US	C0002994	Angioedema
16	SNOMEDCT_US	C0003018	Angiotensins
17	SNOMEDCT_US	C0003232	Antibiotics
18	SNOMEDCT_US	C0003241	Antibodies
19	SNOMEDCT_US	C0003320	Antigens
20	SNOMEDCT_US	C0003360	Antihistamines
21	SNOMEDCT_US	C0003445	Antitoxins
22	SNOMEDCT_US	C0003450	Antivenin
23	SNOMEDCT_US	C0003467	Anxiety
24	SNOMEDCT_US	C0003483	Aorta
25	SNOMEDCT_US	C0003564	Aphonia
26	SNOMEDCT_US	C0233485	apprehension
27	SNOMEDCT_US	C0003842	Arteries
28	SNOMEDCT_US	C0004044	Asphyxia
29	SNOMEDCT_US	C0004057	Aspirin
30	SNOMEDCT_US	C1510438	Assay
31	SNOMEDCT_US	C0004096	Asthma
32	SNOMEDCT_US	C0231221	Asymptomatic
33	SNOMEDCT_US	C0392707	Atopy
34	SNOMEDCT_US	C0004259	Atropine
35	SNOMEDCT_US	C0004268	Attention
36	SNOMEDCT_US	C0004271	Attitude
37	SNOMEDCT_US	C0004398	Autopsy
38	SNOMEDCT_US	C0004521	Aztreonam
39	SNOMEDCT_US	C0004827	Basophils
40	SNOMEDCT_US	C0005558	Biopsy
41	SNOMEDCT_US		
4			

295 candidate UMLS Concepts (CUIs) appeared in *≥3 articles*

***158 CUIs** retained for the dictionary after manual review*

MD Reviewer

Yu et al. Toward high-throughput phenotyping: unbiased automated feature extraction and selection from knowledge sources. JAMIA 2015

Feature Engineering: Additional Options

- NLP Features for PheNorm are basically counts of mentions of clinical concepts within patient notes (in the catchment window). However, additional options can be considered:
- **Negation**: Count concepts negated in text? (e.g., “No fever”)
- **Normalization**: Longer notes have more concepts; is that information useful, or misleading?
- **Dimension Reduction**: May yield simpler models without sacrificing performance by removing duplicative or less-informative features.

Feature Engineering: Additional Options

	Feature engineering options			
Model set	**Exclude negated mentions**	**Normalize by patient's chart length**	**Dimension reduction pre-modeling**	**Scientific question**
1	No	No	No	Does simple feature engineering yield sufficient model performance?
2	Yes	No	No	Does excluding NLP negation improve performance (vs Model set 1)?
3	No	Yes	No	Does normalizing features improve performance (vs Model set 1)?
4	No	No	Yes	Is performance preserved in models based on reduced feature sets?
5	Yes	Yes	Yes	Do all feature engineering options combined improve performance?

Modeling

- We developed models for all 8 logical combinations of those options (negation, normalization, and dimension reduction).
- Each of these **8 model sets** included **5 PheNorm models**,
  - One for each of the 4 silver labels
  - A fifth aggregate model is the average of the predicted probabilities from the 4 silver-label models
  - **40 models total**
- We trained these models using data from patients without gold-standard case labels and evaluated the models using data from a set-aside sample of patients with gold-standard labels.

Gold-Standard Sample

- PheNorm uses silver labels for training, but gold-standard data is necessary for evaluation.
- We used manual chart review to create the gold-standard data used to evaluate our PheNorm models from a stratified random sample of patients.
- Trained chart abstractors following written guidelines assigned **phenotype positive** labels to patients with evidence of at least possible SARS-COV-2 infection and at least symptomatic COVID-19 disease, and **phenotype negative** labels to all other patients.
- Inter-rater agreement assessed using two reviewers at each site.

Methods – Evaluation & Outcomes

Evidence of COVID-19 Infection

Definite or highly probable infection

- PCR-positive or explicit positive assertion

Probable or possible infection

- Symptoms are consistent with a diagnosis of COVID-19 and absence of an explicit *alternative* diagnosis

Unlikely infection

- Explicit *alternative* diagnosis or statement ruling-out COVID-19 and absence of relevant symptoms/labs

Not infected

- No indication in the EHR of infection

Insufficient Information

Severity of Illness Scale (NIH)

SEVERITY LEVEL	SIGN/SYMPTOM
Asymptomatic	No symptoms
Mild	Fever (>=100.4F)
	Cough
	Sore throat
	Malaise/fatigue
	Headache
	Muscle pain
	Nausea
	Vomiting
	Diarrhea
	Loss of sense of taste or smell
Moderate	Shortness of breath (SpO2 >=94%)
	Dyspnea (SpO2 >=94%)
	Abnormal chest imaging (SpO2 >=94%)
Severe	SpO2 <94%
	PaO2/FiO2* <300 mm Hg
	Respiratory freq >30 breaths/min
	Lung infiltrates >50%
Critical	Respiratory failure
	Septic shock
	Multiple organ dysfunction

Results

Results

Study Cohorts

- VUMC: **24,177 patients**, approximately 1.1 million notes
- KPWA: **8,329 patients**, 143,584 notes

Gold-standard evaluation sample

- VUMC: **419 patients** (Cohen's kappa 0.951)
- KPWA: **437 patients** (Cohen's kappa 0.802)

Results

Characteristics of the study cohorts at VUMC and KPWA

	VUMC		KPWA	
Characteristic	**Count**	**Percent**	**Count**	**Percent**
All patients	24 177	100	8329	100
Sex is female	14 025	58	4837	58
Age group				
18-29 years	5645	23	1104	13
30-49 years	8131	34	2503	30
50-69 years	7433	31	3126	38
70+ years	2968	12	1596	19
Race is White	16 407	68	5335	64
Ethnicity is Hispanic	1018	4	756	9

Results

- The AUCs across all PheNorm models ranged from 0.770 to 0.804 at VUMC and 0.801 to 0.853 at KPWA.
- Model PPVs (at maximum F1 score) ranged from 0.858 to 0.903 at VUMC and 0.772 to 0.876 at KPWA.
- The VUMC model with highest AUC was trained on Structured Label 2, without excluding negated mentions, feature normalization, or dimension reduction.
- The highest-AUC KPWA model was also trained on Structured Label 1, with feature normalization, without excluding negated mentions or dimension reduction.

Results

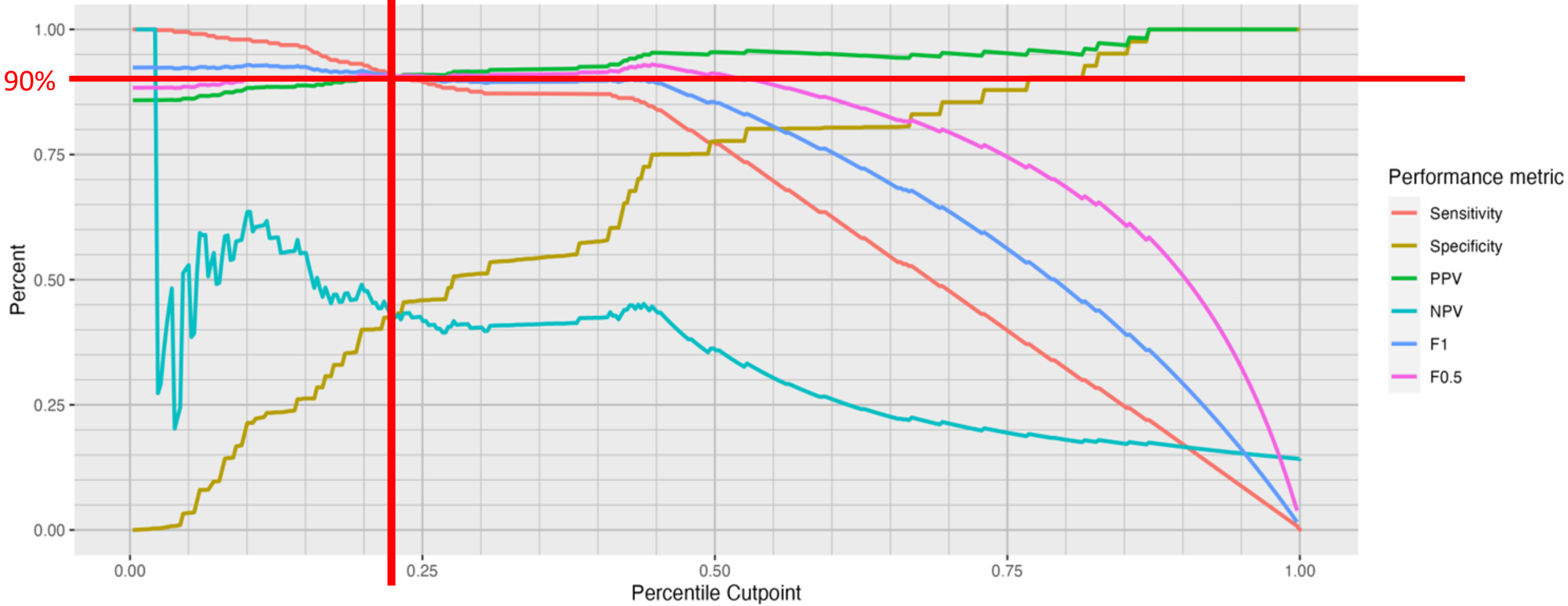
Best performing model sets at each site when maximizing F1 Score

Study site (model set)	Silver label	AUC	Max. F1	Sensitivity	Specificity	PPV	NPV
VUMC (model set 1) Negation: NO Normalize: NO Dim. Reduc: NO	Struc. 1	0.802	0.927	0.976	0.214	0.883	0.597
	Struc. 2	**0.804**	0.929	0.976	0.234	0.885	0.617
	NLP 1	0.788	**0.937**	0.982	0.309	0.896	0.743
	NLP 2	0.775	**0.937**	0.982	0.306	0.896	0.741
	Agg.	0.786	**0.937**	0.982	0.306	0.896	0.741
KPWA (model set 3) Negation: NO Normalize: YES Dim. Reduc: NO	Struc. 1	**0.853**	0.865	0.879	0.662	0.851	0.713
	Struc. 2	0.851	0.862	0.875	0.662	0.850	0.706
	NLP 1	0.819	0.861	0.945	0.451	0.791	0.789
	NLP 2	0.833	**0.869**	0.949	0.482	0.801	0.812
	Agg.	0.847	0.867	0.949	0.472	0.798	0.809

Model Perfo

PPV = 0.90
Sens = 0.90

A) VUMC, Model Set 1, Structured 2 Silver Label

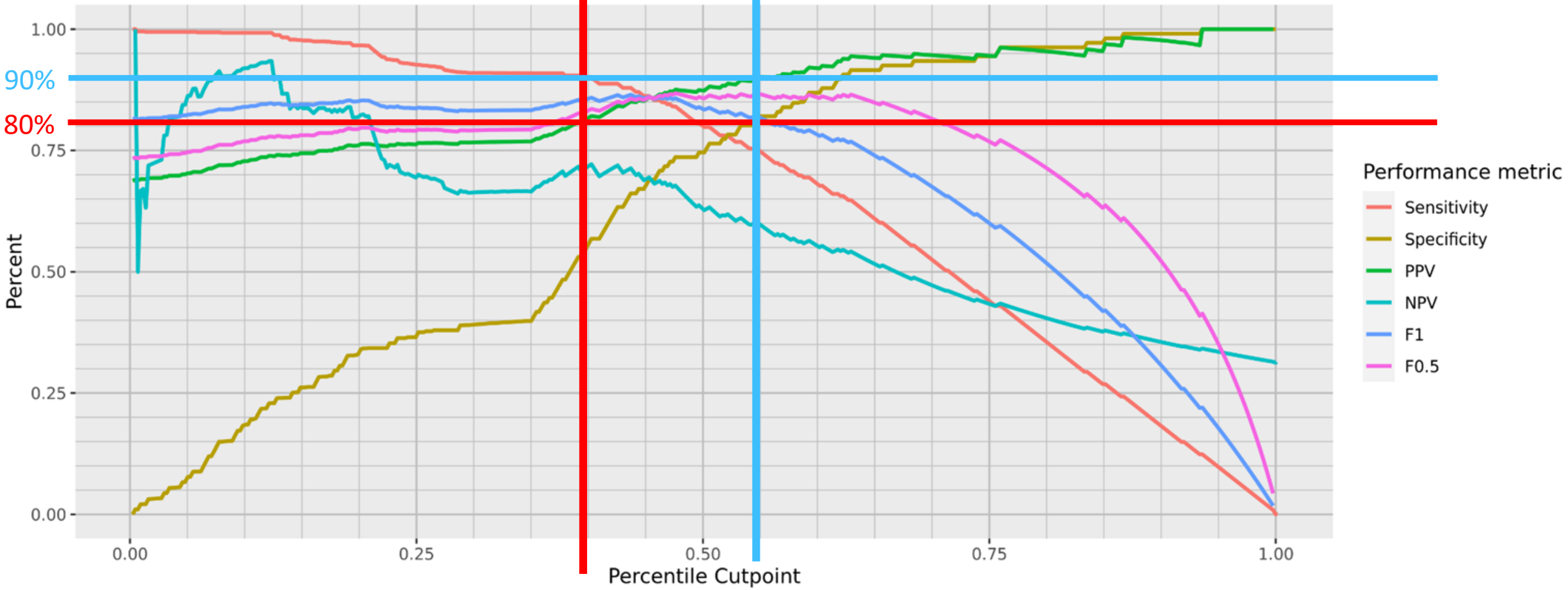


Model Performance

PPV = 0.80
Sens = 0.90

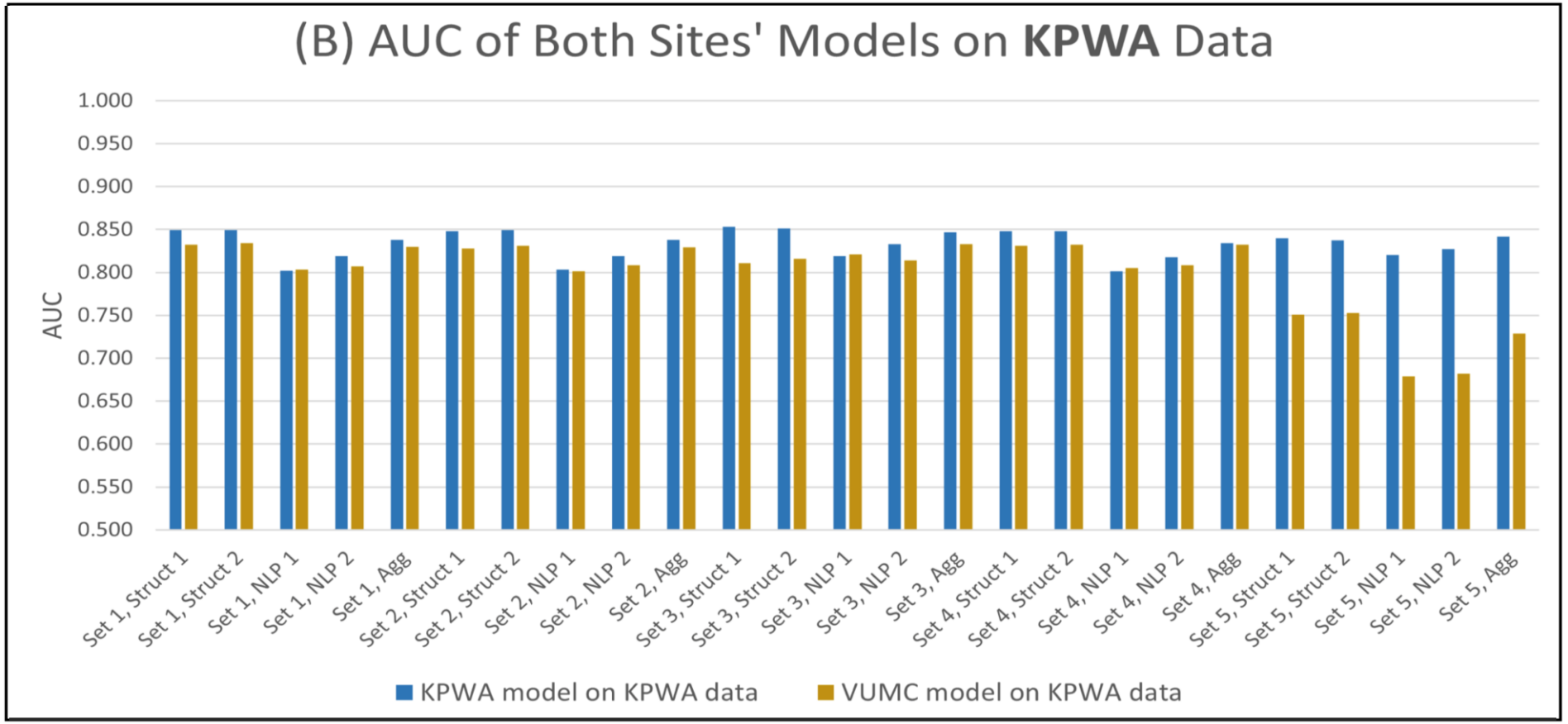
PPV = 90
Sens = 0.75

B) KPWA, Model Set 3, Structured 1 Silver Label



Model Transportability

In addition to testing local models on local data, we also tested each others' models on local data, producing surprisingly good results.



Discussion

- Model performance varied by silver label, but models trained on structured data labels generally had higher AUCs at both sites.
- Performance also varied when using alternative feature engineering options, but all yielded strong performance.
- Excluding negated mentions and normalizing feature counts had little impact on model performance; and dimension reduction produced models with strong performance based on fewer features.
- Overall, these changes/additions only minorly affected performance.

Discussion

- Using cut-points of model-predicted probability that yielded greater than or equal to 80% PPV (a commonly used “benchmark”) yields sensitivities of 0.999 in the best VUMC model and 0.905 in the best KPWA model.
- Performance metrics and levels suitable for addressing different specific scientific questions may be achieved by selecting different cut-points of predicted probability.

Discussion

- At both study sites the performance of externally trained models was generally similar to that of internally trained models.
- At VUMC, the AUC of the best externally trained model was 0.804, compared to 0.817 for the best locally trained model.
- At KPWA, the AUC of the best-performing externally trained model was 0.834, compared to 0.853 for the best locally trained model.
- At least for this phenotype, this evidence of transportability of models is promising.

Limitations

- We used data from early in the COVID-19 pandemic, which may introduce idiosyncrasies relative to other phenotypes and time periods.
- We used data from only 2 healthcare settings which, though diverse, may not be representative of other settings.
- The positive predictive value of the COVID-19 ICD code at VUMC was higher than expected (85%), but PheNorm still resulted in improved performance.
- More work should be done to assess PheNorm performance on other acute phenotypes .

Conclusions

- The PheNorm approach can successfully identify an acute health condition, COVID-19 Disease.
- Tools such as PheNorm, utilizing unstructured EHR data, can support rapid phenotyping for public health surveillance.
- Preliminary results indicate that models trained at one site may be transportable to other sites with little decrease in performance.
- The simplicity of the PheNorm approach allows it to be applied at multiple study sites with substantially reduced overhead compared to traditional phenotyping.

Acknowledgements

- **FDA Sentinel Initiative**
- **Vanderbilt University Medical Center:**
  Daniel Park, Jill M. Whitaker, Michael F. McLemore, Elizabeth E. Hanchrow, Dax Westerman, Joshua T. Osmanski, Robert Winter, David M. Aronoff
- **Kaiser Permanente Washington Research Institute:**
  David S. Carrell, Brian D Williamson, Arvind Ramaprasan, Ann Kelley, Mary Shea, David Cronkite
- **Sentinel Innovation Center Workgroup:**
  Saranrat Wittayanukorn, Danijela Stojanovic, Yueqin Zhao, Darren Toh, Kevin B. Johnson, Rishi Desai

Thank You

joshua.smith@vumc.org