Machine learning to identify lupus phenotype in electronic health records

Yuan Luo, PhD, FAMIA

Chief AI Officer

Northwestern University Clinical and Translational Sciences Institute

Institute for Augmented Intelligence in Medicine

Associate Professor

Department of Preventive Medicine

Northwestern University

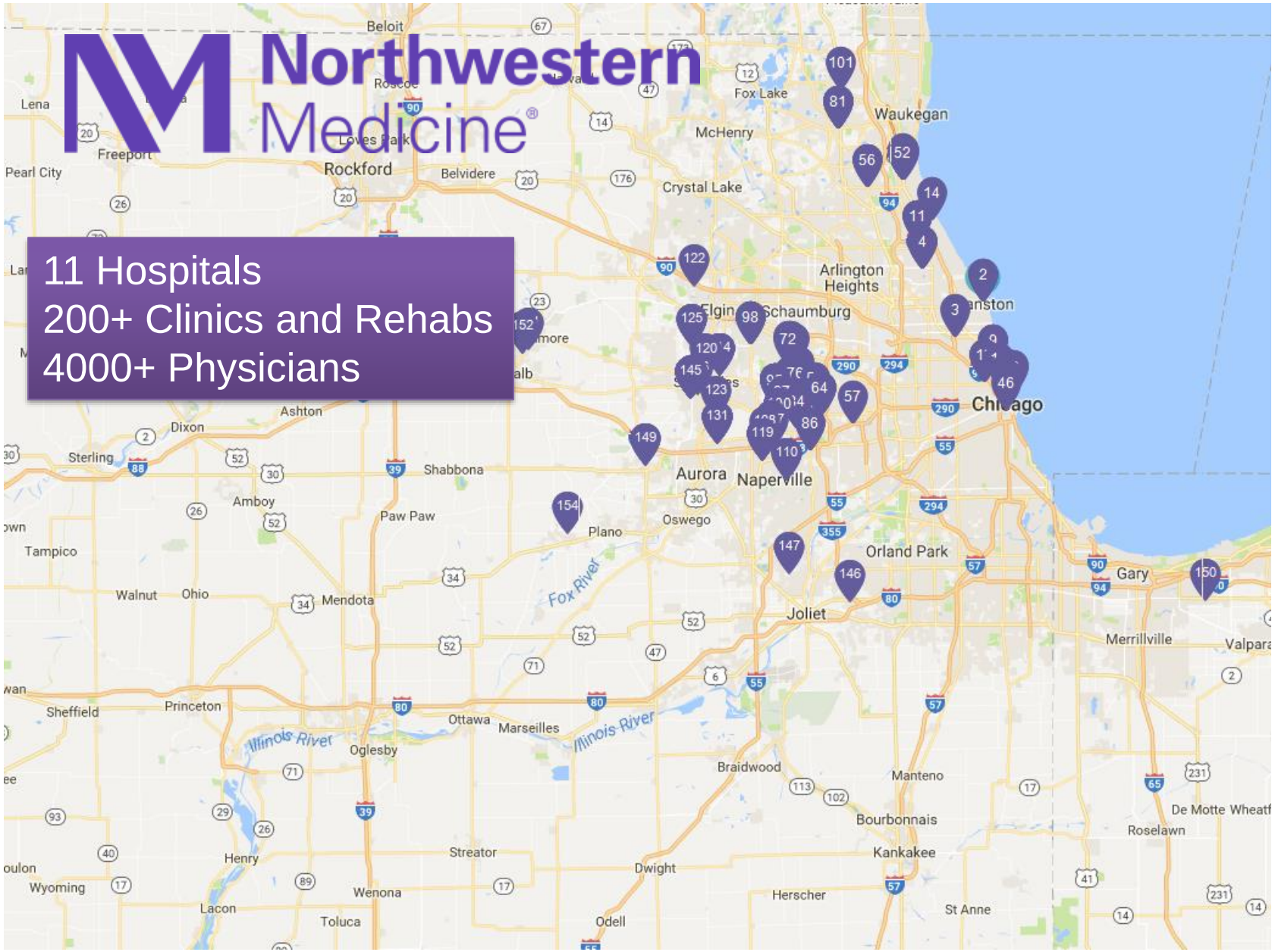
yuan.luo@northwestern.edu

@yuanhypnosluo

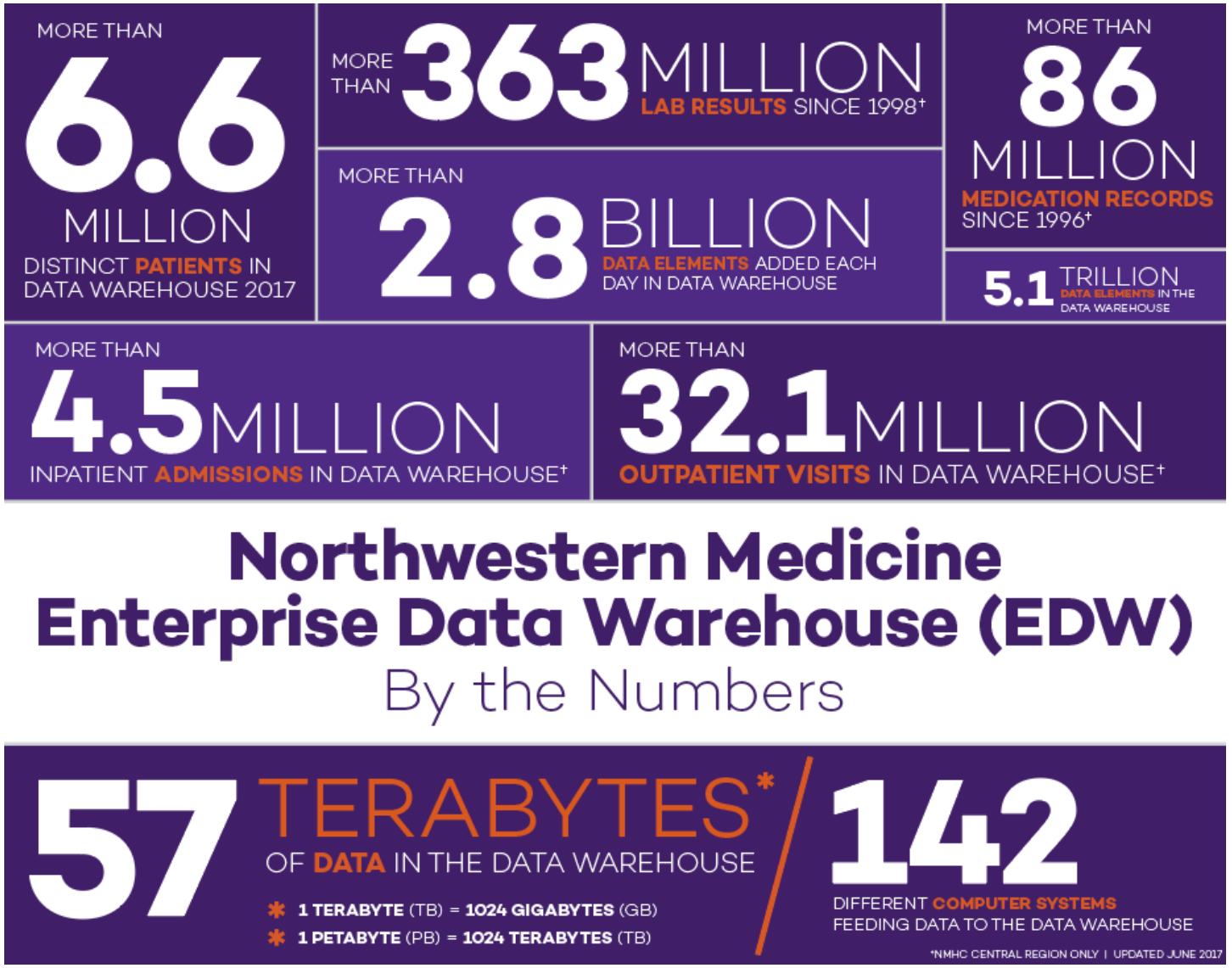
3/26/2022

NORTHWESTERN UNIVERSITY FEINBERG SCHOOL OF MEDICINE

Northwestern University


Northwestern Medicine®
11 Hospitals
200+ Clinics and Rehabs
4000+ Physicians
Beloit
Rockford
Freeport
Pearl City
Belvidere
Crystal Lake
McHenry
Fox Lake
Waukegan
Arlington Heights
Schaumburg
Elgin
Chicago
Aurora
Naperville
Oswego
Plano
Shabbona
Paw Paw
Amboy
Dixon
Sterling
Ashton
Tampico
Walnut
Ohio
Mendota
Princeton
Sheffield
Ottawa
Marseilles
Oglesby
Fox River
Illinois River
Orland Park
Joliet
Gary
Merrillville
Braidwood
Manteno
Bourbonnais
Kankakee
Dwight
Herscher
Odell
Streator
Wenona
Toluca
Lacon
Henry
Wyoming
St Anne
Roselawn

Northwestern Medicine Enterprise Data Warehouse



Northwestern Medicine Enterprise Data Warehouse

Data Engineering

Data Integration

Integrate data within the Enterprise Data Warehouse, and create custom data structures, and cubes, to support analytics for specific business functions

Data Modeling

Develop and maintain an enterprise data model that illustrates attributes and relationships for all data stored within the EDW

Data Architecture

Develop and implement data marts, tabular models, and data structures which enable self-service

Research Analytics

KPI Development & Reporting

Collaborate and develop reports for FSM researchers to identify patient cohorts for research studies & develop self-service reports for NUCATS and FSM Administration

Customer Support

Work with FSM researchers to create SLAs, respond to ad-hoc research requests, conduct educational sessions for power users, and administer the FSM exception policy

Advanced Analytics

Develop custom data structures in a variety of data models to enable data sharing amongst AMC peers, and support key research initatives . Develop predictive analytic applications to support outcomes research

Application Development

Research App Dev

Custom development of applications used to track research studies, patient cohorts, and research outcomes

Reporting Portal & Self-Service BI

Develop, and maintain new features on the NMEDW Reporting Portal. Administer Tableau Server, the self-service BI tool for the Health System

Customer Support & Maintenance

Service tickets related to new feature requests, or incidents and maintain an inventory of 35+ applications for NMHC & FSM.

Data Warehouse Operations

Source System Loads

Work with vendors and customers across the organizations to load new data sources into the EDW. Maintain, monitor, and troubleshoot source system loads daily

Infrastructure

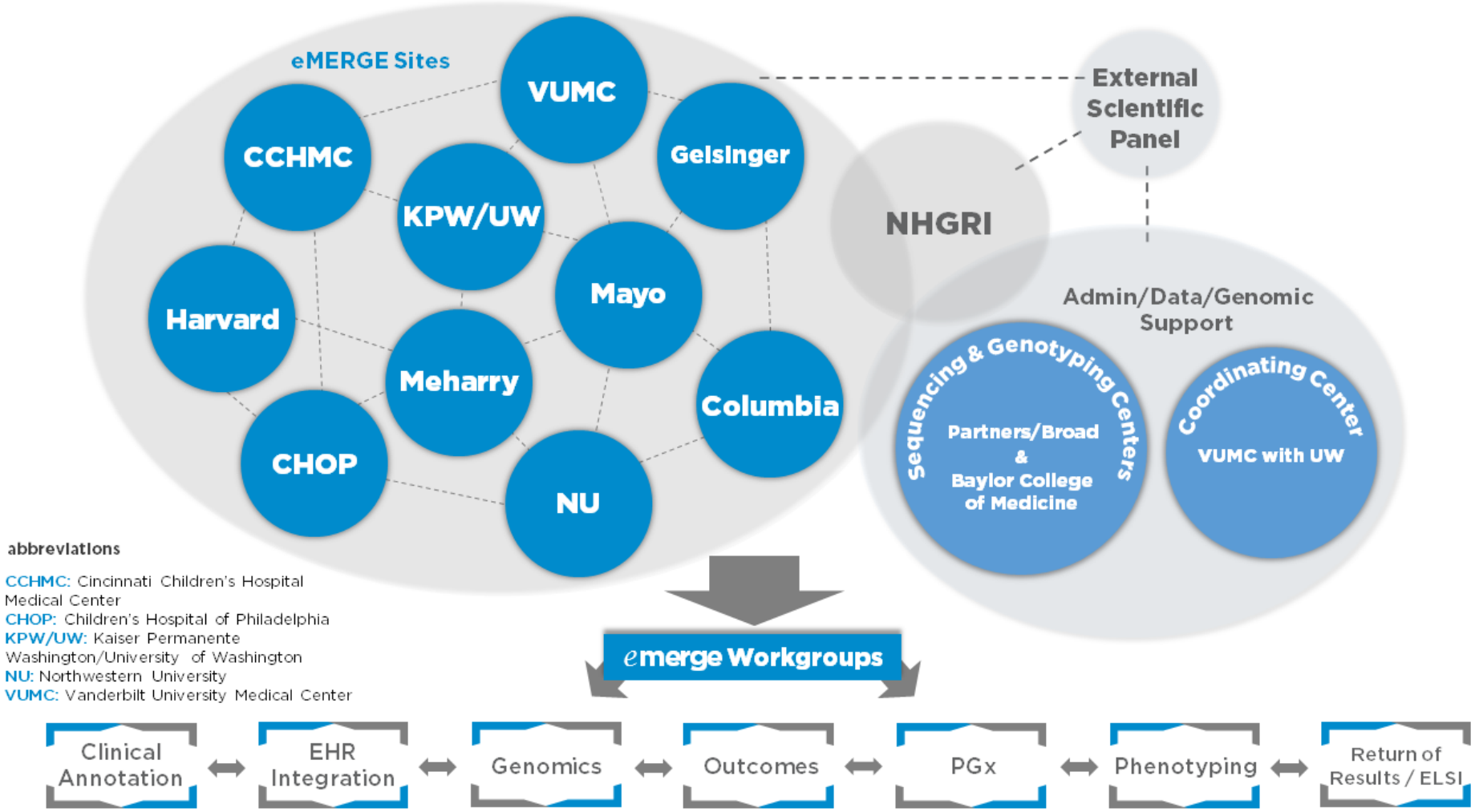
Oversee EDW infrastructure (server, storage, network, database) and collaborate with NMIS to ensure high-availability and quick query response times

Manage Epic Reporting Environments

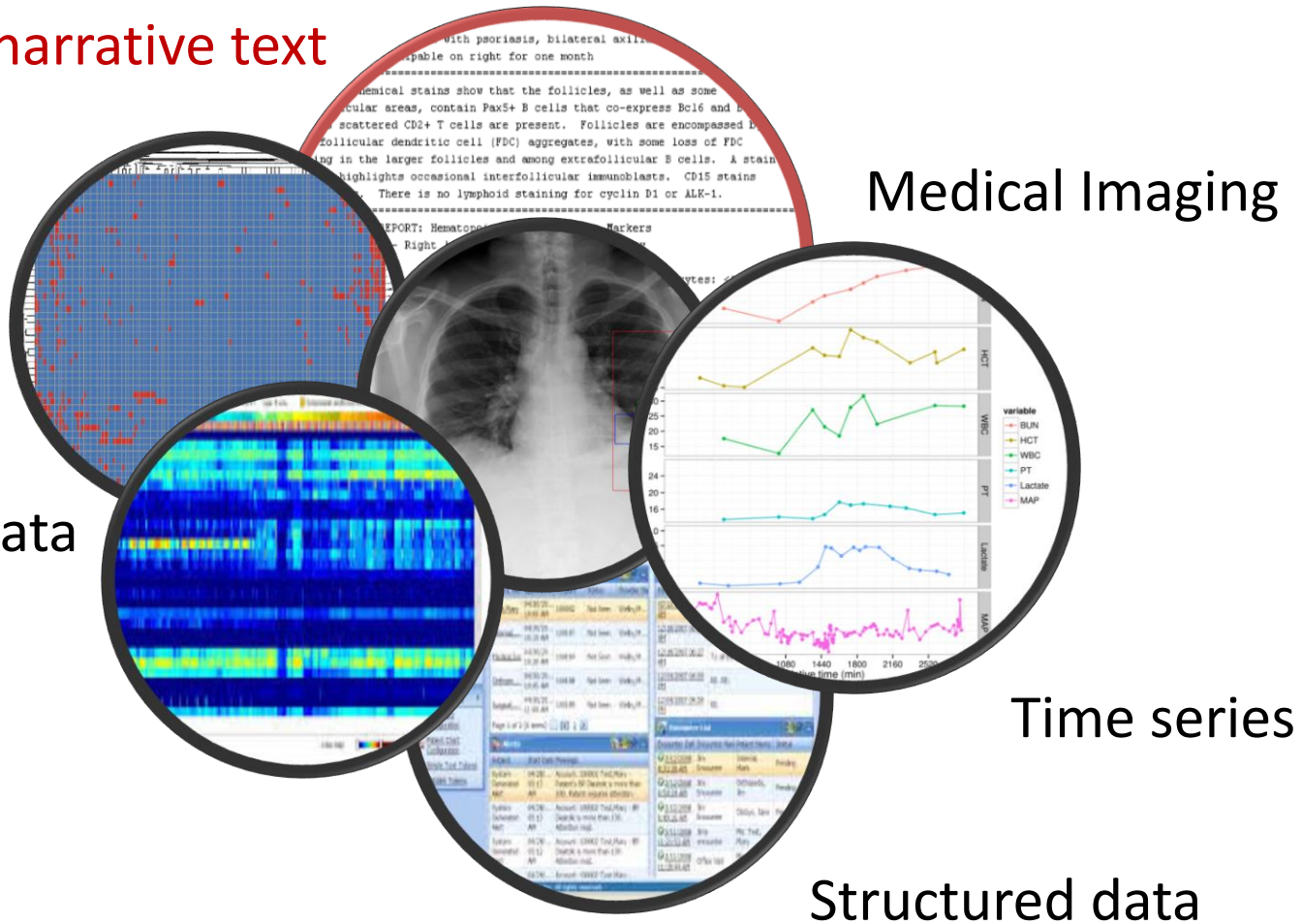
Manage, monitor, and troubleshoot all Clarity environments on campus, as well as Caboodle (Epic's DW)

Courtesy of NMEDW Director

Electronic Medical Records and Genomics (eMERGE) Network



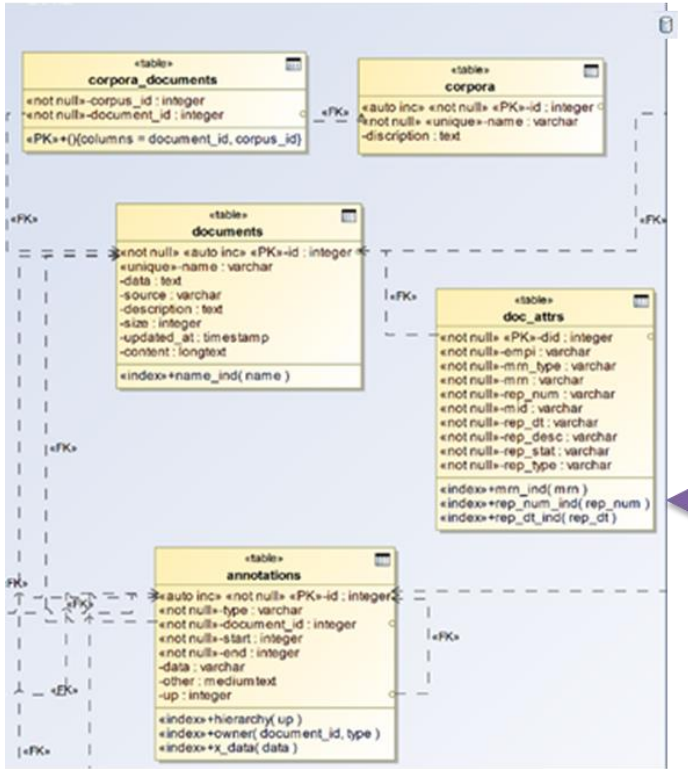
AI to Integrate Multimodal Biomedical Data



Bulk NLP to power R&D and BI

Common Data Model Tables

- Persistent storage
- Easy to query
- Fast Execution

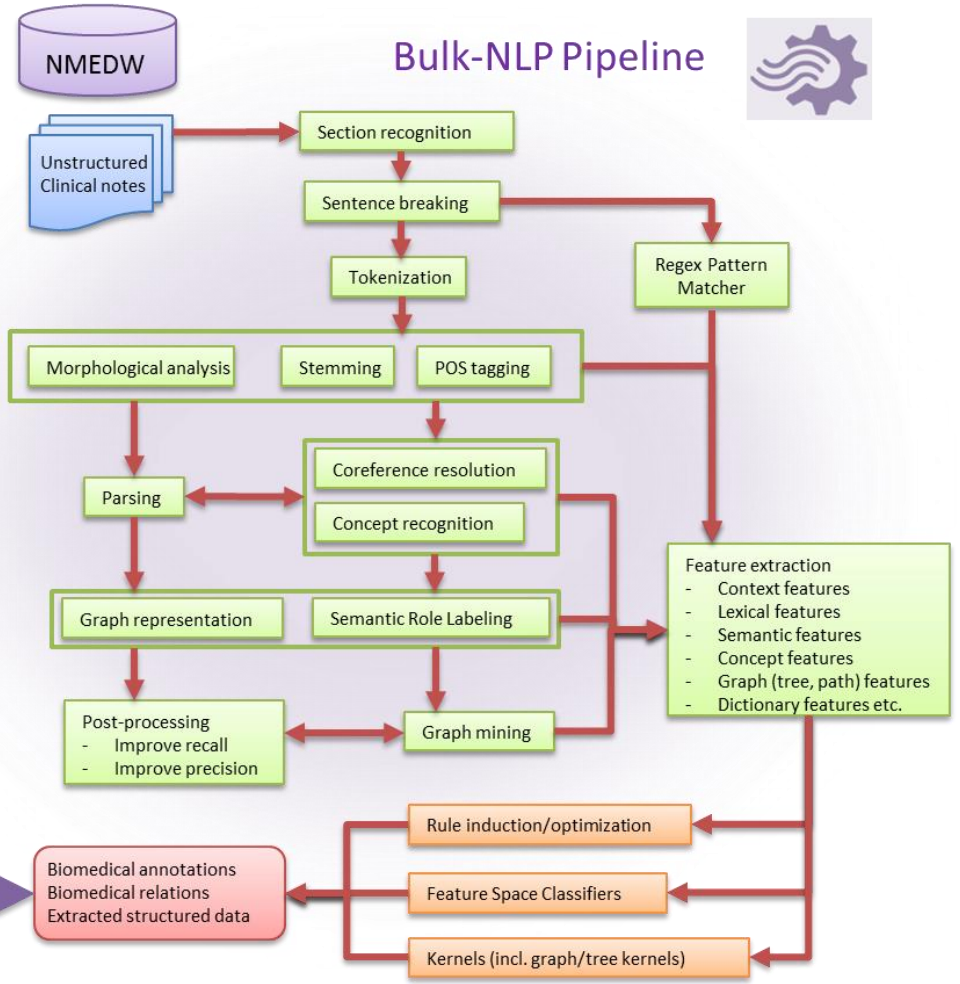


Powering R&D and BI

- Breast Cancer Phenotyping
- CVD Data Mart
- Informed Staffing

Cloud processing
NLP as a service
Reducing technical barrier
Enabling in-depth information access

Bulk-NLP Pipeline



NUCATS Clinical and Translational Sciences Institute · Northwestern · Northwestern Medicine · Ann & Robert H. Lurie Children's Hospital of Chicago · Northwestern Medicine Feinberg School of Medicine · Shirley Ryan Abilitylab

Lupus nephritis computational phenotyping

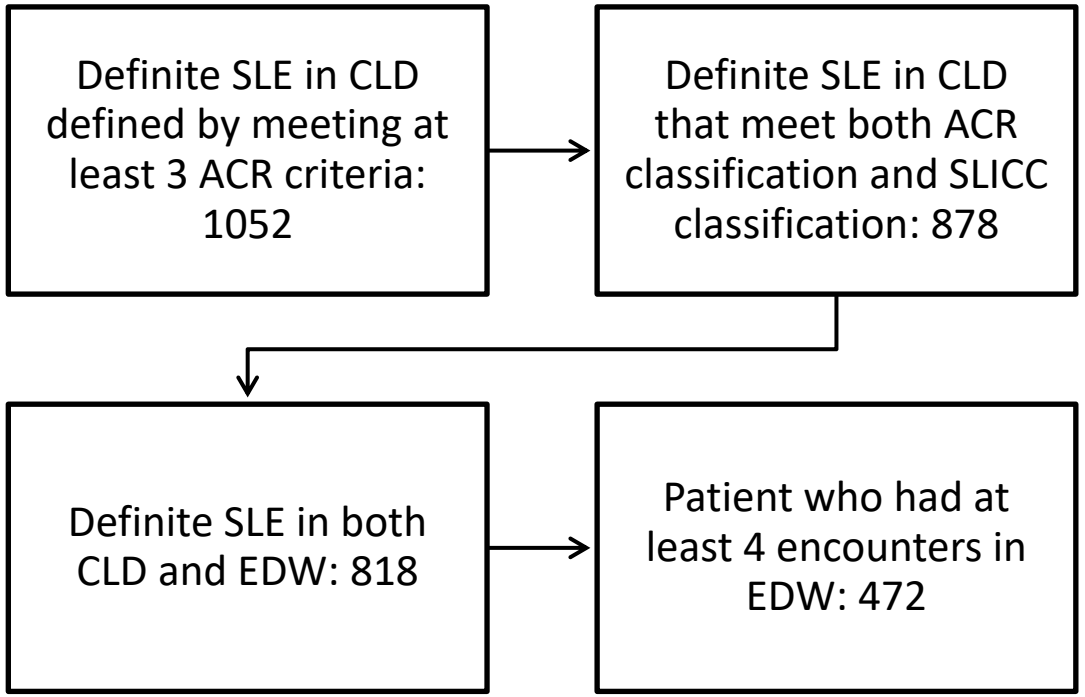
- Lupus nephritis is one of the major risk factors for systemic lupus erythematosus (SLE) mortality
- Chibnik et al. ICD codes; claims data; positive predictive value (PPV) 88%; sensitivity and specificity not mentioned; not externally validated
- Li et al. used ICD codes; good sensitivity and specificity; low PPV, 63.4%
- A need to use natural language processing (NLP) method to improve performance
- A need to validate algorithm internally and externally

Chibnik et al 2010;
Li et al 2020

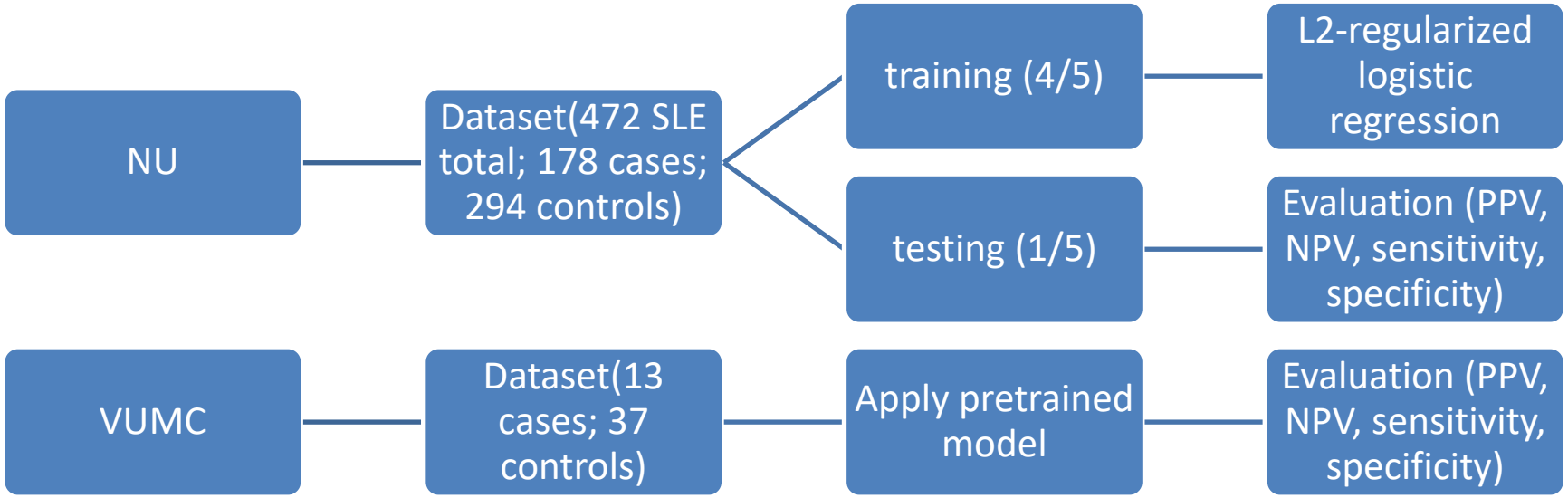
Objectives

- Develop NLP based algorithms to identify lupus nephritis phenotype among SLE patients
- Validate the algorithms using internal dataset and external dataset

SLE patient cohorts



SLE patient cohorts



Lupus nephritis phenotype: baseline model

Type	Terminology		Code	Note
Diagnosis	ICD-9		593.6	
Diagnosis	ICD-9		593.81	
Diagnosis	ICD-9		791.7	
Diagnosis	ICD-10		N28.0	
Diagnosis	ICD-10		N28.9	
Diagnosis	ICD-10		R80	This is a top-level code that contains a tree of other codes. R80 is not used for diagnosis
Diagnosis	ICD-10		R80.9	
Diagnosis	ICD-10		R82.99	This is a top-level code that contains a tree of other codes. R82.99 is not used for diagnosis
Laboratory	LOINC		2889-4	(value >500mg (/24H), >0.5g (/24H)
Laboratory	LOINC		21482-5	(value >500mg (/24H), >0.5g (/24H)
Laboratory	LOINC		51790-4	value >0/hpf or /lpf
Laboratory	LOINC		33804-6	value >0/hpf or /lpf
Laboratory	LOINC		5807-3	value >0/hpf or /lpf

Lupus nephritis phenotype: NLP for feature extraction

Lupus nephritis	Baseline model	Full metamap model (binary)	Full metamap model (counts)	Mixed filtered metamap/regex/ICD model
Features	Any positive mention of lupus nephritis in ICD9/10 or laboratory tests	- All MetaMap concepts (exclude negated concept) - C10000	- All MetaMap concepts (exclude negated concept)	- Structured data feature (1 variable): positive mention in ICD or laboratory test - NLP features - MetaMap (7 variables): C0024143, C0268757, C0268758, C4053955, C4053958, C4053959, C4054543 - Regular expression (5 variables) - Mention of any keywords related to nephritis class II - Mention of any keywords related to nephritis class III - Mention of any keywords related to nephritis class IV - Mention of any keywords related to nephritis class V - Mention of any keywords related to and proteinuria
Classification model	Rule-based	LASSO	LASSO	LASSO
Parameter tuning	NA	C(0.0001-10000); solver (newton-cg, lbfgs, sag and saga)	C(0.0001-10000); solver (newton-cg, lbfgs, sag and saga)	C(0.0001-10000); solver (newton-cg, lbfgs, sag and saga)

CUIs and their definition in mixed model

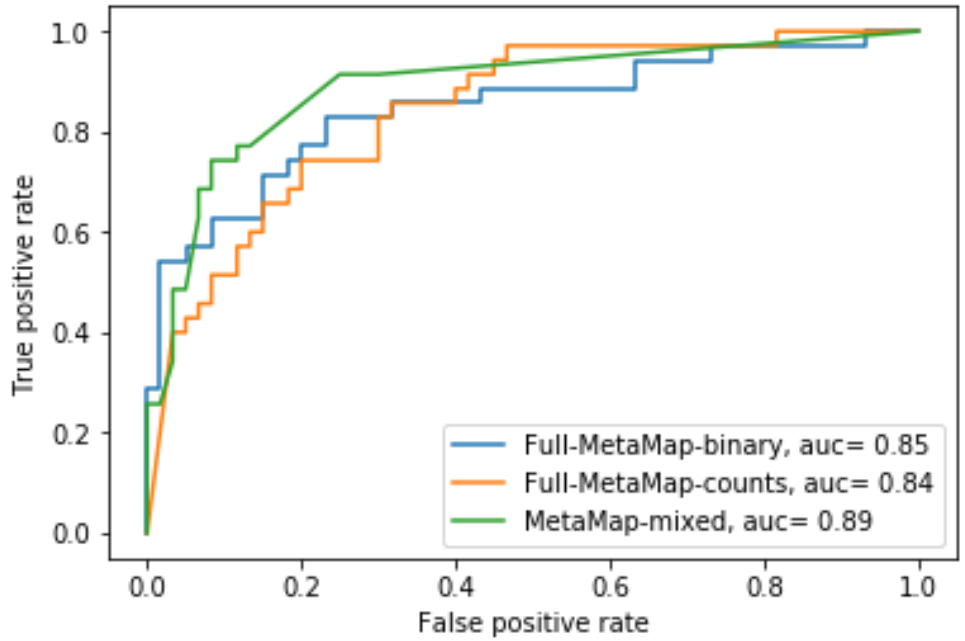
CUIs	Definition
C0024143	Glomerulonephritis in the context of systemic lupus erythematosus.
C0268757	Lupus nephritis - WHO Class IV
C0268758	Lupus nephritis - WHO Class V
C4053955	Systemic lupus erythematosus nephritis, with active or inactive diffuse, segmental or global endo- or extracapillary glomerulonephritis involving greater than or equal to 50% of all glomeruli, typically with diffuse subendothelial immune deposits, with or without mesangial alterations.
C4053958	Systemic lupus erythematosus nephritis exhibiting mesangial hypercellularity or mesangial expansion by light microscopy, with mesangial immune deposits. Isolated subepithelial or subendothelial deposits may be visible by immunofluorescence or electron microscopy, but not by light microscopy
C4053959	Systemic lupus erythematosus nephritis with active of inactive focal, segmental or global endo- or extracapillary glomerulonephritis involving less than 50% of all glomeruli, typically with focal subendothelial immune deposits with or without mesangial alterations.
C4054543	Membranous nephritis associated with systemic lupus erythematosus.

Results

Phenotype	Dataset	Algorithm	Sensitivity	Specificity	PPV	NPV	F Measure
Lupus nephritis	NU (testing set)	Baseline	0.43	0.6	0.39	0.64	0.41
Lupus nephritis	NU (testing set)	Full MetaMap (binary)	0.63	0.92	0.82	0.81	0.71
Lupus nephritis	NU (testing set)	Full MetaMap (counts)	0.6	**0.95**	**0.88**	0.8	0.71
Lupus nephritis	NU (testing set)	MetaMap mixed	**0.74**	0.92	0.84	**0.86**	**0.79**
Lupus nephritis	VUMC	Baseline	0.92	0.61	0.46	0.96	0.62
Lupus nephritis	VUMC	MetaMap mixed	**1**	**0.97**	**0.93**	**1**	0.96

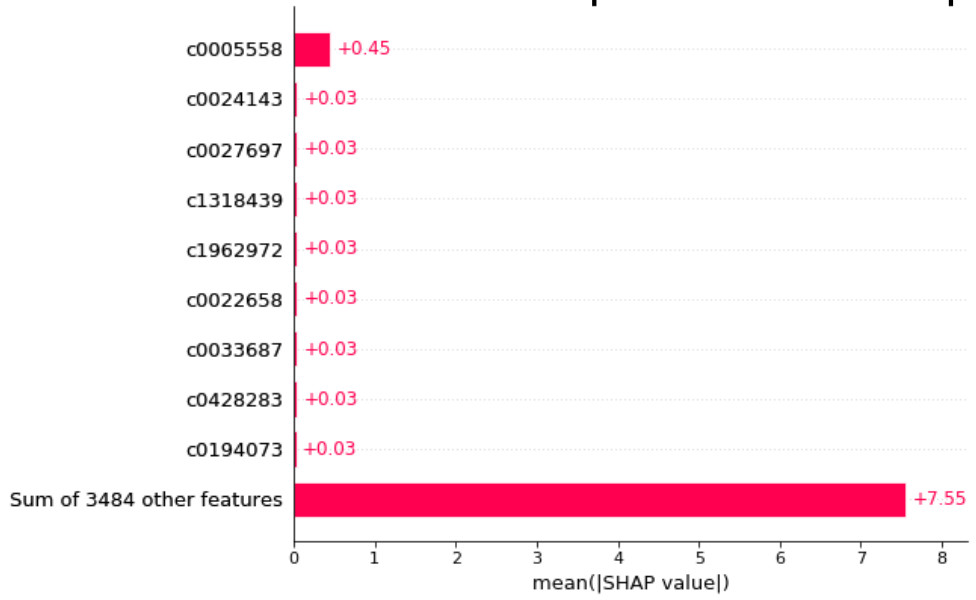
Results

Area under curve (AUC) for Full MetaMap (binary), Full MetaMap (counts), and MetaMap mixed model in NU testing set



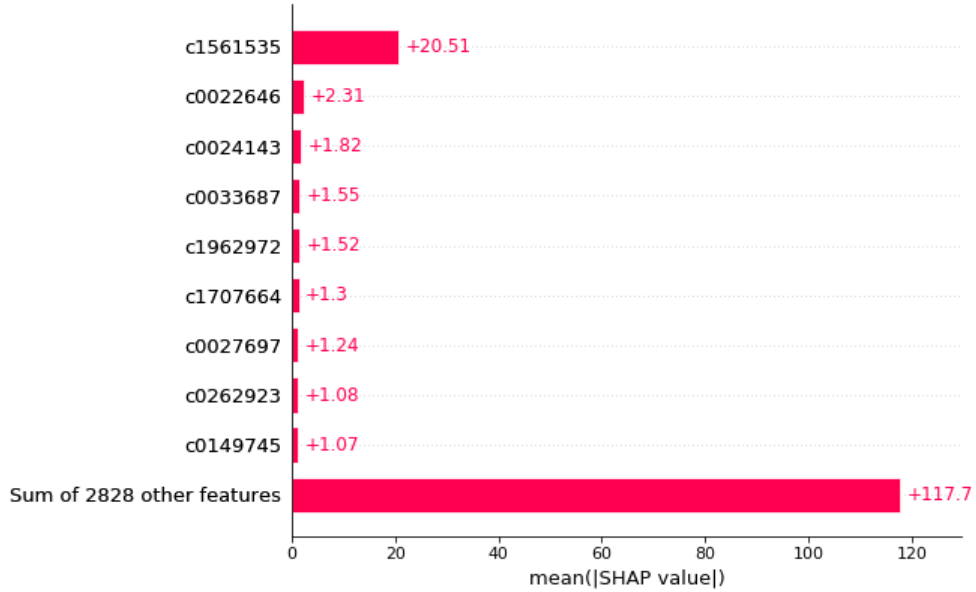
SHAP plot for Full MetaMap (binary) model

- SHAP feature importance measured as the mean absolute Shapley values
  - C0005558 was the most important feature, changing the predicted absolute lupus nephritis probability on average by 0.45
  - C0005558: Patient required removal of tissue or fluid specimen to establish a diagnosis
  - C0024143: Glomerulonephritis in the context of systemic lupus erythematosus
  - C0027697: Inflammation of renal tissue
  - C1318439: A quantitative measurement of the total amount of creatinine present in a sample of urine
  - C1962972: Proteinuria, CTCAE 3.0
  - C0022658: Kidney Disorder
  - C0033687: Proteinuria
  - C0428283: Urine creatinine level finding
  - C0194073: Kidney Biopsy



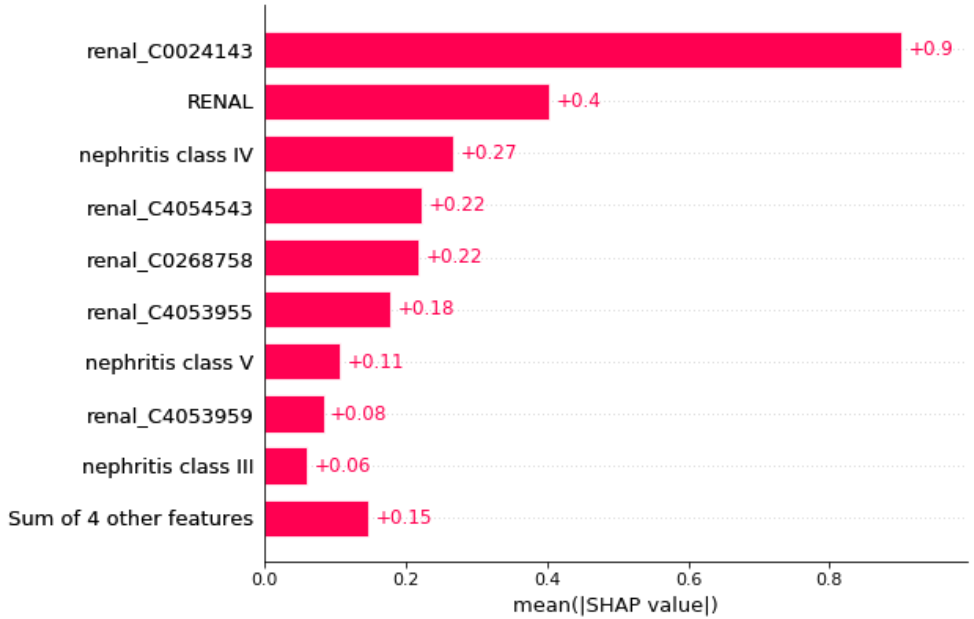
SHAP plot for MetaMap (counts) model

- SHAP feature importance measured as the mean absolute Shapley values
  - C1561535 was the most important feature, changing the predicted absolute lupus nephritis probability on average by 20.51
  - C1561535: Creatinine, CTCAE
  - C0022646: Kidney
  - C0024143: Lupus Nephritis
  - C0033687: Proteinuria
  - C1962972: Proteinuria, CTCAE 3.0
  - C1707664: Delayed Release Dosage Form
  - C0027697: Nephritis
  - C0262923: Urine protein test
  - C0149745: Oral Ulcer



SHAP plot for MetaMap Mixed model

- SHAP feature importance measured as the mean absolute Shapley values
  - C002413 (glomerulonephritis in the context of systemic lupus erythematosus) was the most important feature, changing the predicted absolute lupus nephritis probability on average by 0.9
  - RENAL: renal indictor from structured data
  - renal_C4054543: Membranous Lupus Nephritis
  - renal_C0268758: SLE glomerulonephritis syndrome, WHO class V
  - renal_C4053955: Systemic Lupus Erythematosus Nephritis Class IV
  - renal_C4053959: Systemic Lupus Erythematosus Nephritis Class III



Discussion

- Developed and validated three NLP algorithms to identify lupus nephritis phenotype among SLE patients in EHR
- All three NLP algorithms outperformed the baseline algorithm in the internal validation dataset
- The MetaMap/regex/ICD mixed model (NLP based) outperformed baseline model in both internal and external validation sets (0.79 vs 0.41; 0.96 vs 0.62)
- Limitations: limited sample size of external validation dataset

Y Deng, J Pacheco, A Chung, C Mao, J Smith, J Zhao, WQ Wei, A Barnado, C Weng, C Liu, A Cordon, J Yu, Y Tedla, A Kho, R Ramsey-Goldman, T Walunas, Y Luo, Natural language processing to identify lupus nephritis phenotype in electronic health records, *under review*

Latent Class Analysis for SLE Sub-phenotype Identification

- SLE is a systemic auto-immune disease
- SLE affects many organs
- Heterogeneity in nature
- Clinical trial on lupus treatment has been challenging
- Stratification on SLE may potentially improve clinical trial results

ACR (American College of Rheumatology) criteria

ACR Criteria [1,2]	Explanation
Malar Rash	Flat or raised erythema, often sparing the nasolabial folds
Discoid Rash	Raised erythematous patches with keratotic scaling, follicular plugging, and atrophic scarring
Photosensitivity	By patient history or physician observation
Oral Ulcers	Oral or nasopharyngeal ulceration, usually painless
Arthritis	Involving ≥2 peripheral joints, with tenderness and swelling
Serositis	Pleuritic pain or rub or evidence of pleural effusion
Renal Disorder	Confirmed by ECG, rub or evidence of pericardial effusion, Persistent proteinuria > 0.5 g per day; OR Cellular casts (RBC, granular, mixed)
Neurologic Disorder	Seizures - in the absence of other causes or drugs; OR Psychosis - in the absence of other causes or drugs
Hematologic Disorder	Hemolytic anemia - with reticulocytosis; OR Leukopenia: < 4000/ mm; OR Lymphopenia: < 1500/ mm; OR Thrombocytopenia:
Immunological Disorder	Positive anti-DNA; OR positive anti-Sm; OR positive test for antiphospholipid antibodies
Antinuclear antibodies	By immunofluorescence or ELISA, at any point in time

Objectives

- To evaluate if latent class analysis (LCA) [3] could identify distinct SLE subtypes based on ACR criteria
- To evaluate the potential association of these identified clusters with the risk of mortality

Methods

Latent cluster analysis (LCA)

Model selection: optimal number of clusters (BIC)

Cluster1

Cluster2

… …

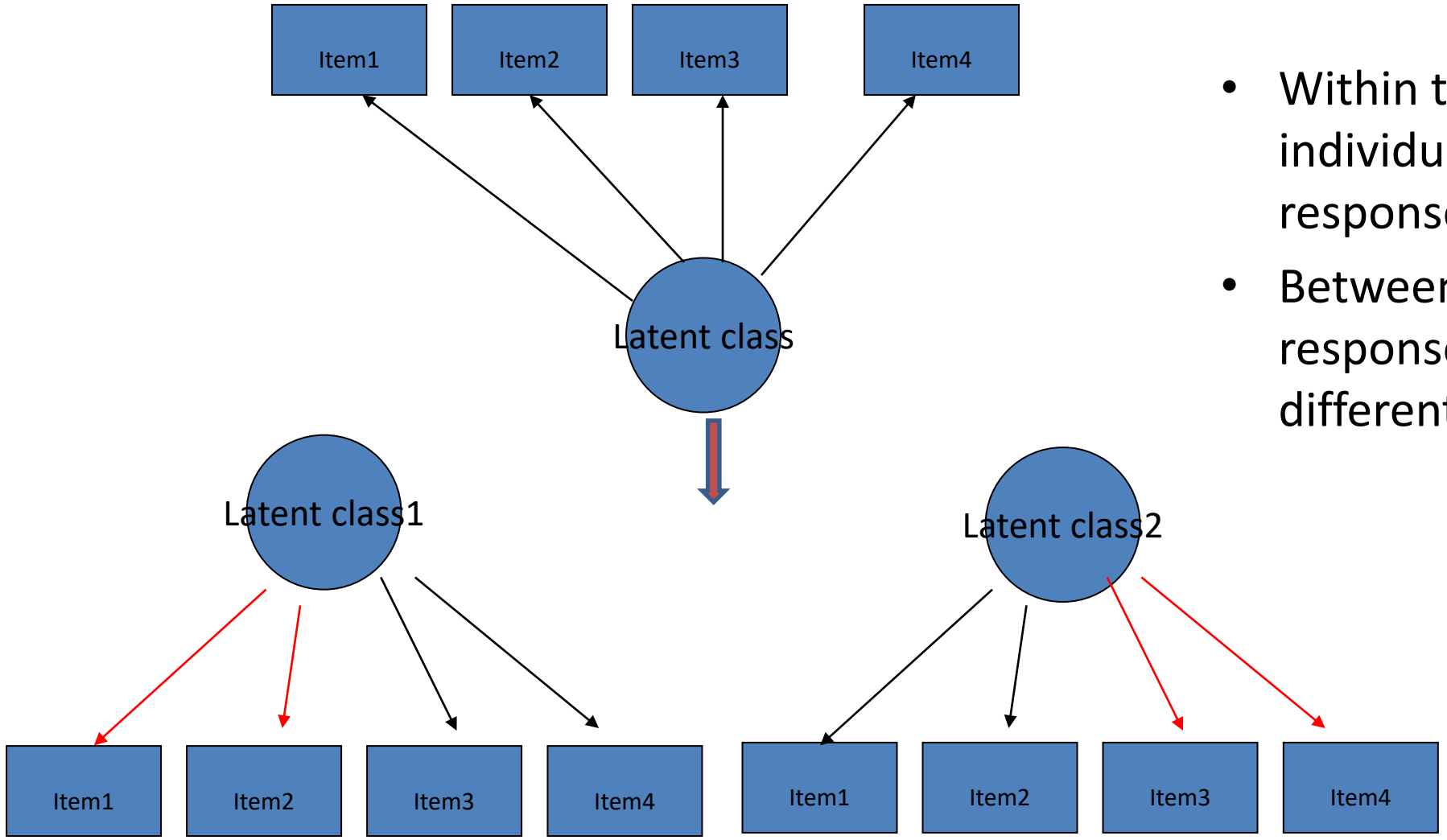
Model Evaluation: bootstrapping + Jaccard score

Patient characteristics comparison among clusters: Chi-square test/ANOVA

Survival outcome comparison: Kaplan Meier

Cluster clinical interpretation

The LCA Model



- Within the same latent class, individual has similar response pattern
- Between latent classes, response patterns are different

The LCA Model

- Assumption:
  - Mixture of C classes [3]
    - $P(\mathbf{y}) = \sum_{x=1}^{C} P(X = x)P(y|X = x)$

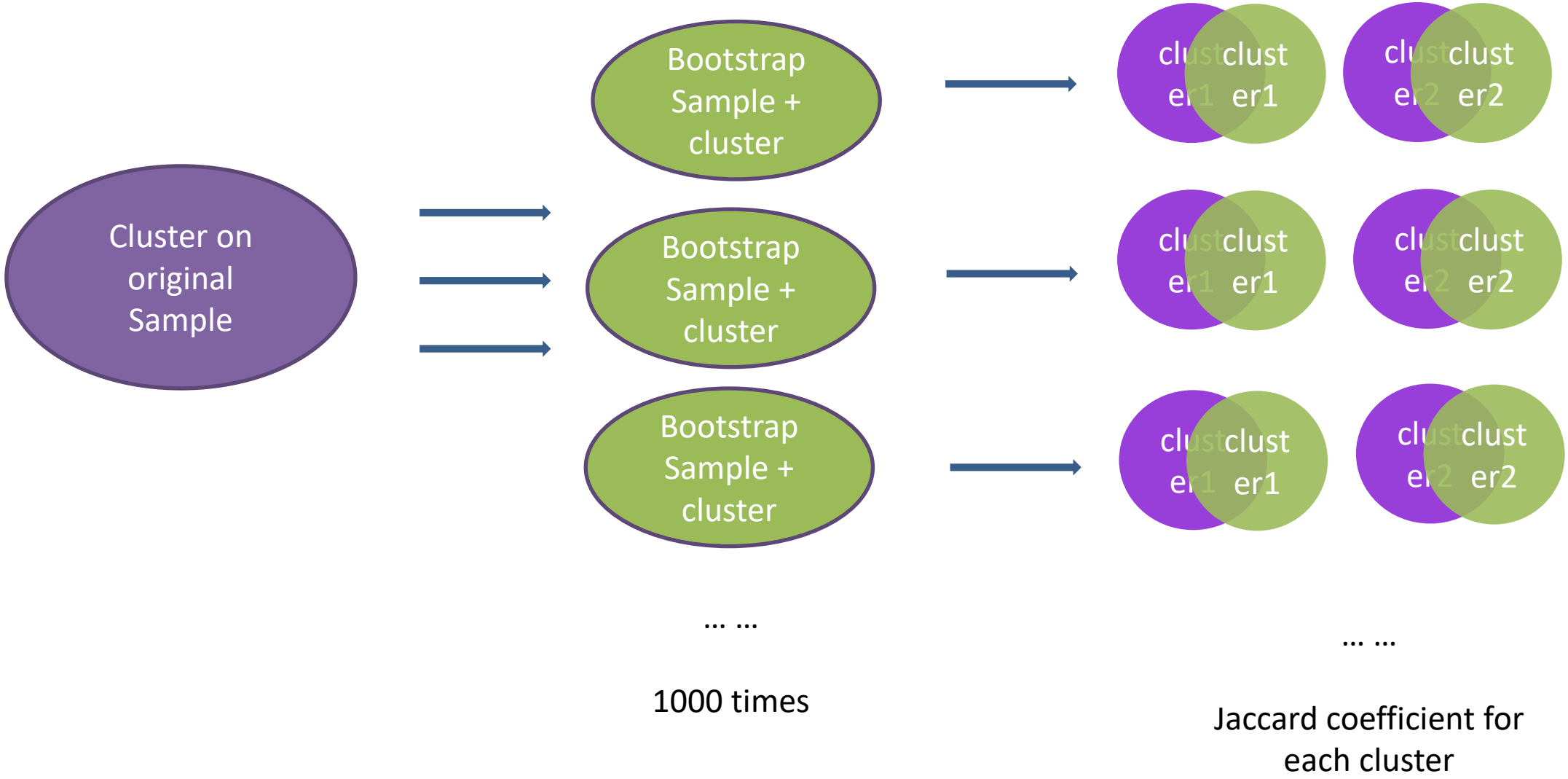
      $$P(y_1, y_2, y_3) = P(X = 1)P(y_1, y_2, y_3 \mid X = 1) + P(X = 2)P(y_1, y_2, y_3 \mid X = 2)$$
  - Independency within classes
    - $P(\mathbf{y} \mid X = x) = \prod_{k=1}^{K} P(y_k \mid X = x)$

$$P(y_1, y_2, y_3 \mid X = 1) = P(y_1 \mid X = 1)P(y_2 \mid X = 1)P(y_3 \mid X = 1)$$

$$P(y_1, y_2, y_3 \mid X = 2) = P(y_1 \mid X = 2)P(y_2 \mid X = 2)P(y_3 \mid X = 2)$$

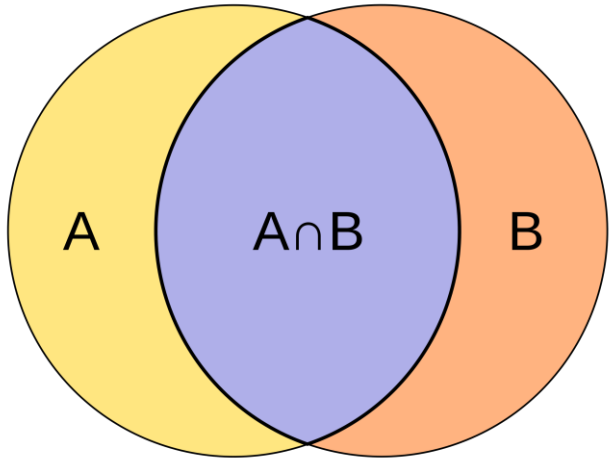
- Likelihood: $P(\mathbf{y}) = \sum_{x=1}^{C} P(X = x) \prod_{k=1}^{K} P(y_k \mid X = x)$

Model stability evaluation



Model stability: calculate Jaccard score

- For every cluster in the original clustering, find the most similar cluster in the new clustering, calculate Jaccard coefficient [4]
- Jaccard coefficient: Intersection over Union
- For 1000 repetitions, get the average Jaccard coefficient for each cluster

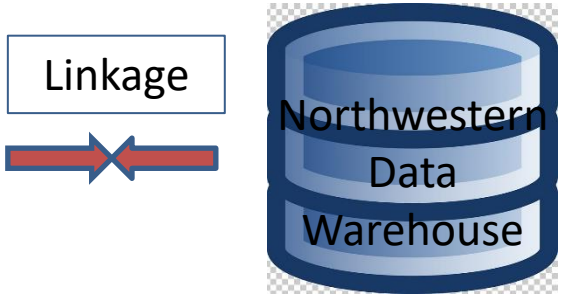


$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}.$$

Results: Patient cohort

856 SLE patients in CLD

Linkage



846 SLE patients

Obtain death data from death master file: available till 2014-03-21

Exclude patients having SLE after 2014-03-21

Exclude patients with missing data

826 SLE patients

END

826 SLE patients

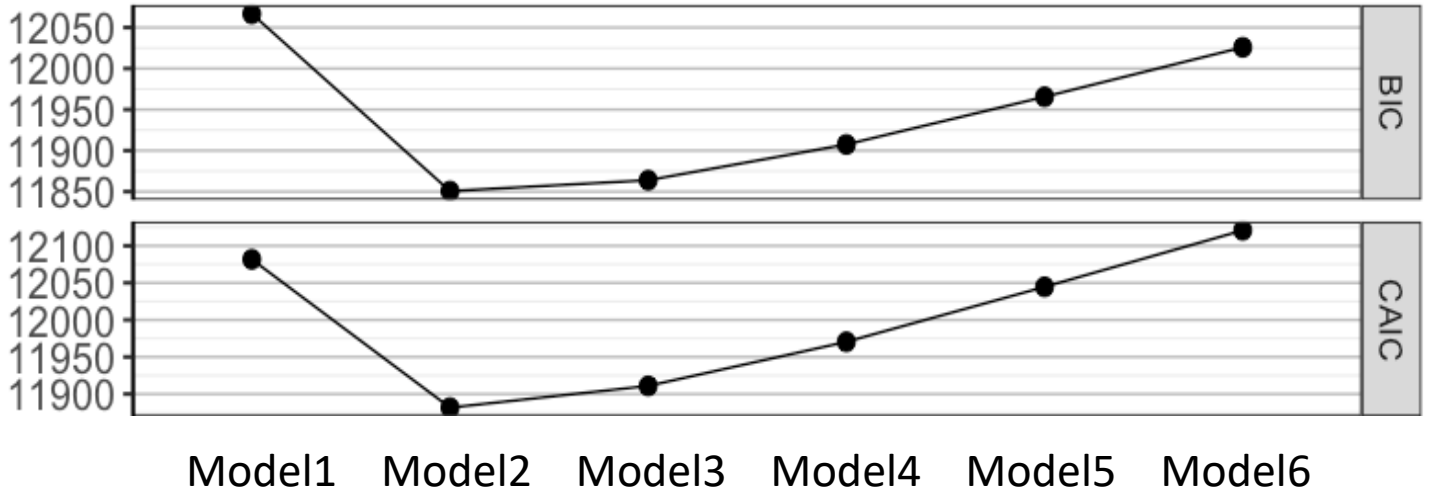
Patient cohort characteristics

- 826 patients in both NMEDW and CLD
- The average onset age of SLE is 30
- As of 2014-03-21, 68/826 patients had died
- The mortality rate was 8.2%

Variable	Percentage (%)
Male	8.23
Caucasian	54.48
African American(AA)	27.00
Young onset (1-16)	10.77
Adult onset (17-50)	83.66
Late onset (50-90)	5.57
Photosensitivity	65.38
Arthritis	87.65
Serositis	63.56
Renal disease	34.75
Neurological disease	7.51
Hematological disorder	52.78
Immnological disorder	73.12
Oral ulcer	44.67
Rash	63.00
Antinuclear antibodies	94.79

LCA on our data: two clusters generated

- Performed LCA on the features mentioned above
- we repeated 30 times to automate the search for the global maximum of the log-likelihood function
- When the number of clusters was 2, the BIC is the lowest
- Average Jaccard Coefficient (1000 times bootstrapping) for clusters 1 and 2 were 0.876 and 0.906, respectively



Between cluster characteristics comparison

Parameter	Group1	Group2	PValue
Total Number	468(56.71%)	358(43.29%)	NA
YoungOnset	63(13.42%)	26(7.32%)	0.04216
AdultOnset	382(81.63%)	309(86.31%)	0.04216
LateOnset	23(4.95%)	23(6.38%)	0.04216
Race (Caucasian)	189(40.40%)	261(72.93%)	< 2.2e-16
Race(AA)	157(33.57%)	66(18.39%)	< 2.2e-16
Race (others)	122(26.03%)	31(8.68%)	< 2.2e-16
SEX (male)	63(13.37%)	13(3.75%)	2.467e-06

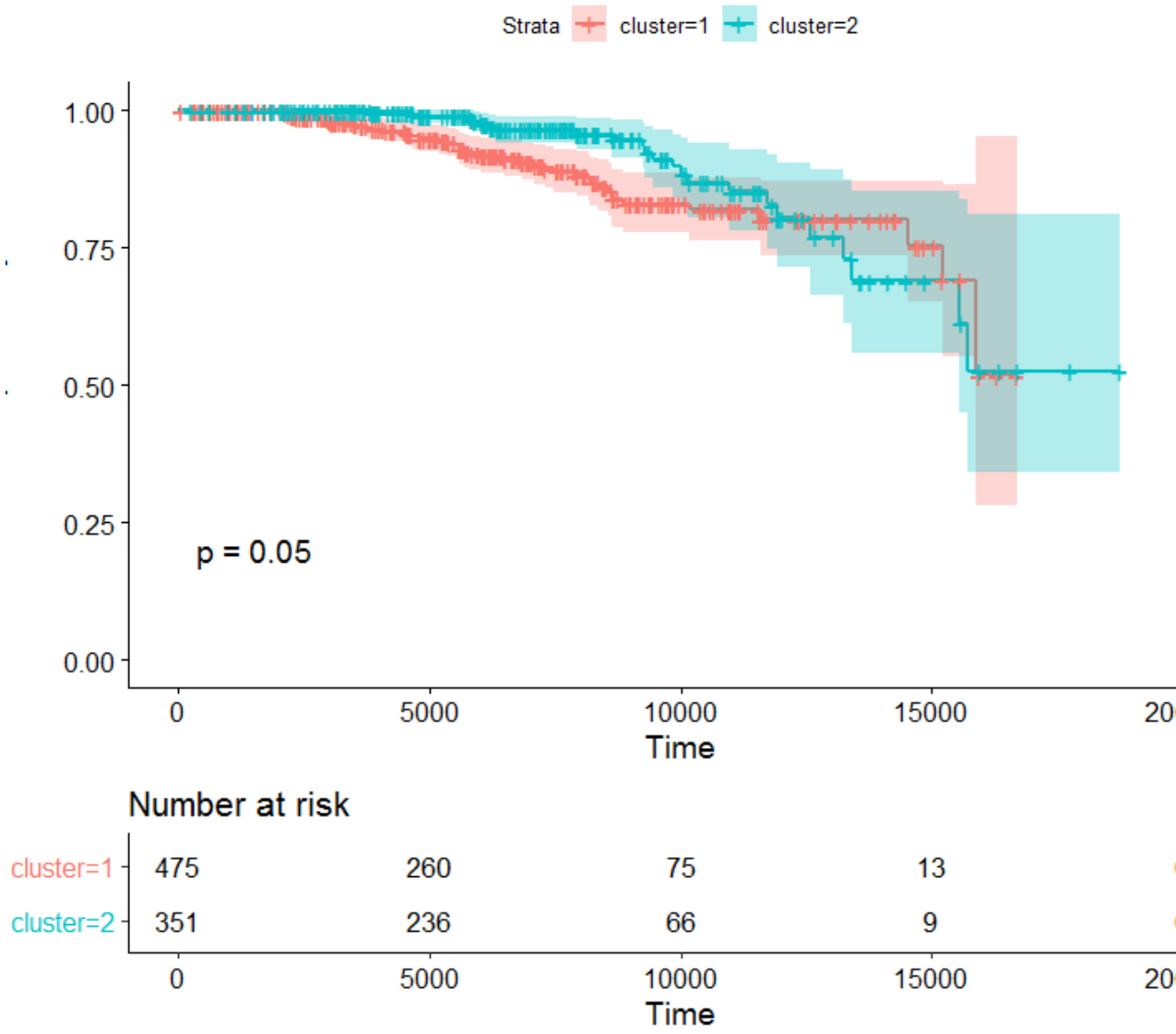
Parameter	Group1	Group2	PValue
Total Number	468(56.71%)	358(43.29%)	NA
Rash	232(49.66%)	287(80.37%)	< 2.2e-16
Photosensitivity	(45.21%)	329(91.79%)	< 2.2e - 16
Arthritis	393(83.98%)	331(92.46%)	0.0006997
Serositis	181(38.78%)	119(33.37%)	0.0499
Renal	246(52.46%)	41(11.54%)	< 2.2e-16
Neural	47(10.07%)	15(4.15%)	0.0005999
HEME	271(58.00%)	165(45.95%)	4.974e-05
ANA	453(96.71%)	330(92.28%)	0.003458
IMMUN	434(92.74%)	170(47.43%)	< 2.2e-16
Ulcer	139(29.67%)	230(64.33%)	< 2.2e-16

Heatmap plot for two clusters

- Cluster 1 was enriched in patients with organ involvement and fewer patients with skin manifestations (49.66% rash, 45.21% photosensitivity, 29.67% oral ulcer)
- Cluster 2 was enriched in patients with skin manifestations (80.35% rash, 91.77% photosensitivity, 64.31% oral ulcer) but less organ involvement
- cluster 1 also consists of more non-white (59.60% vs 27.07%), male (12.14% vs 3.12%) and younger onset patients (13.42% vs 7.32%)

Survival curve

- The 475 SLE patients in cluster1, we had 51 death event, mortality rate was 10.7%
- Cluster2, out of 351 patients we had 26 death event, mortality rate was 7.4%
- Log-rank test showed that two clusters had significant survival curve with P = 0.05



Conclusion

- LCA identified 2 clusters (primarily organ involvement group and skin involvement group) with distinct clinical manifestations and survival outcomes
- Although results from this study are preliminary, it shows strong potential of unsupervised learning in identifying homogeneous SLE groups

Machine learning to integrate multi-modal healthcare data

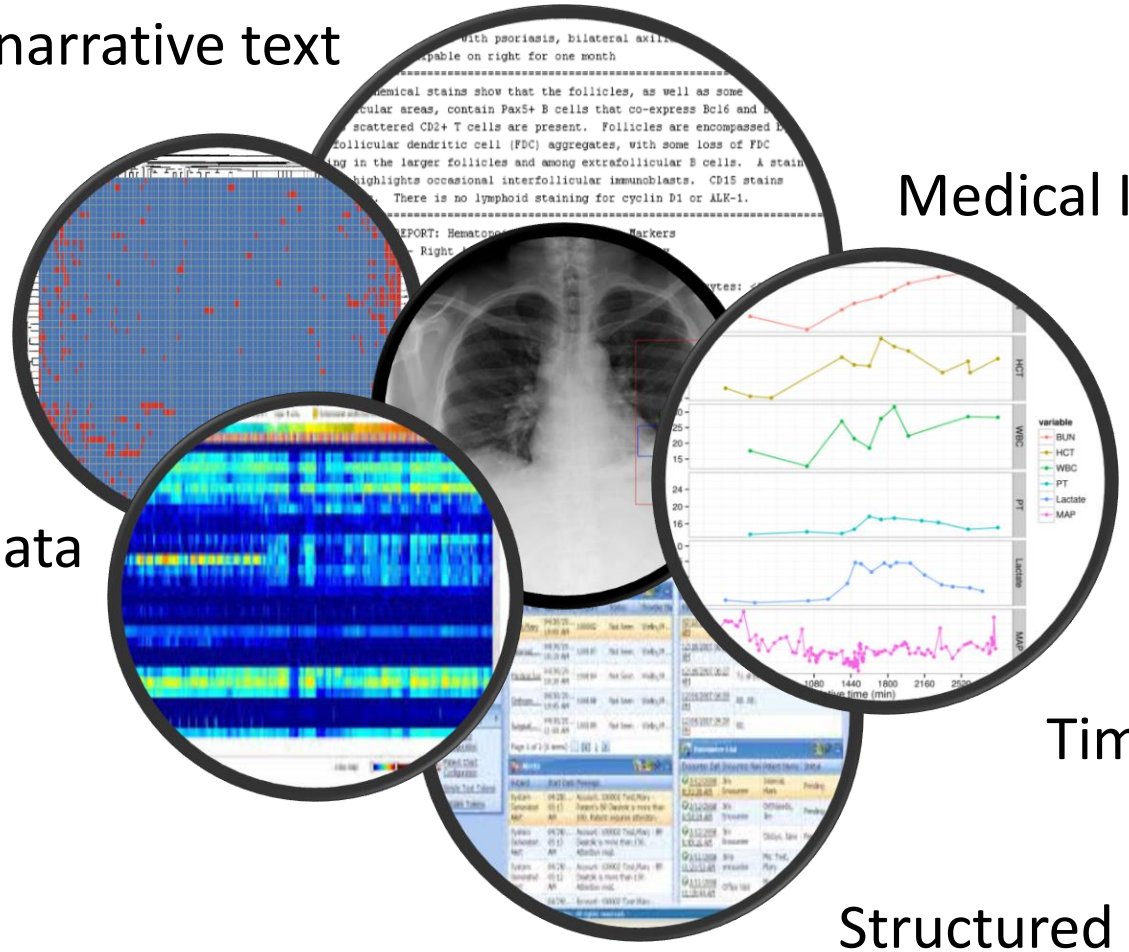
Clinical narrative text

Medical Imaging

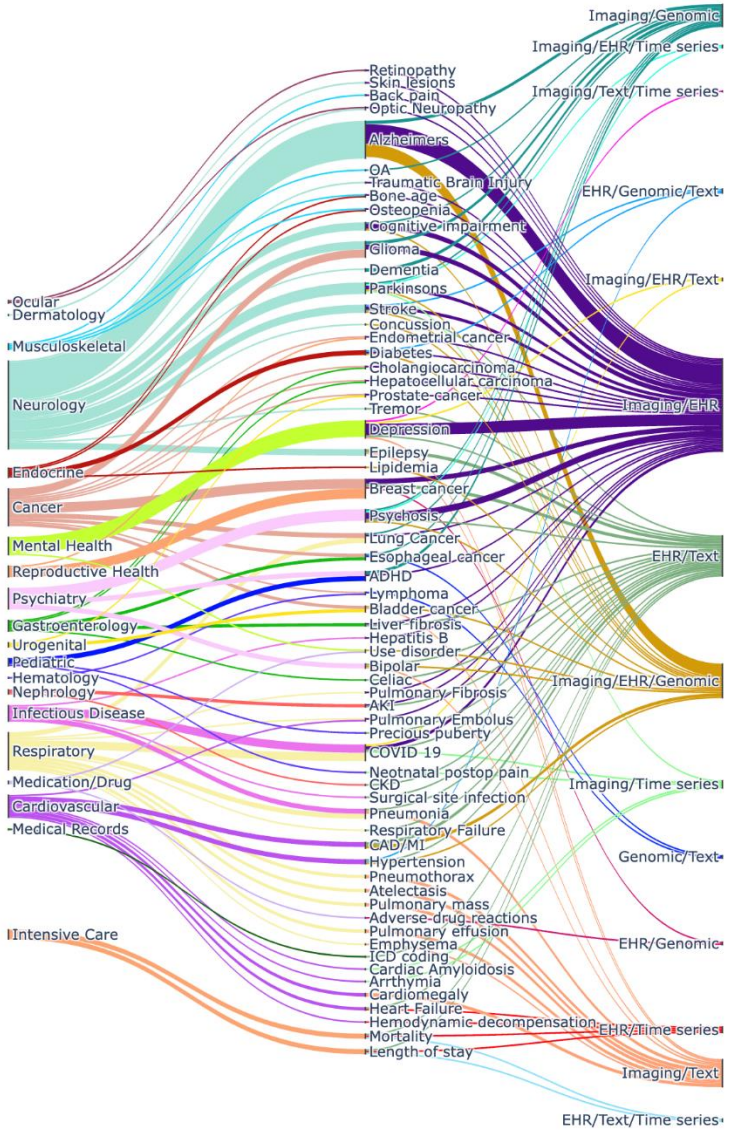
Omics data

Time series

Structured data



Current status of multi-modal machine learning in healthcare



CURRENT LIMITATIONS IN THE ML DATA FUSION PIPELINE

MODEL TRAINING

- Training multiple models
- Weighting of data interaction
- Voting issues for multiple models

MODEL BUILDING

- Complex and time-consuming multi-modal models, creating a barrier to creation
- Unclear which fusion models are superior

MODEL TESTING

- Lack of comparison with single modality
- Lack of comparison with alternative fusion strategies

Merging complementary and correlated data

DATA FLOW-THROUGH

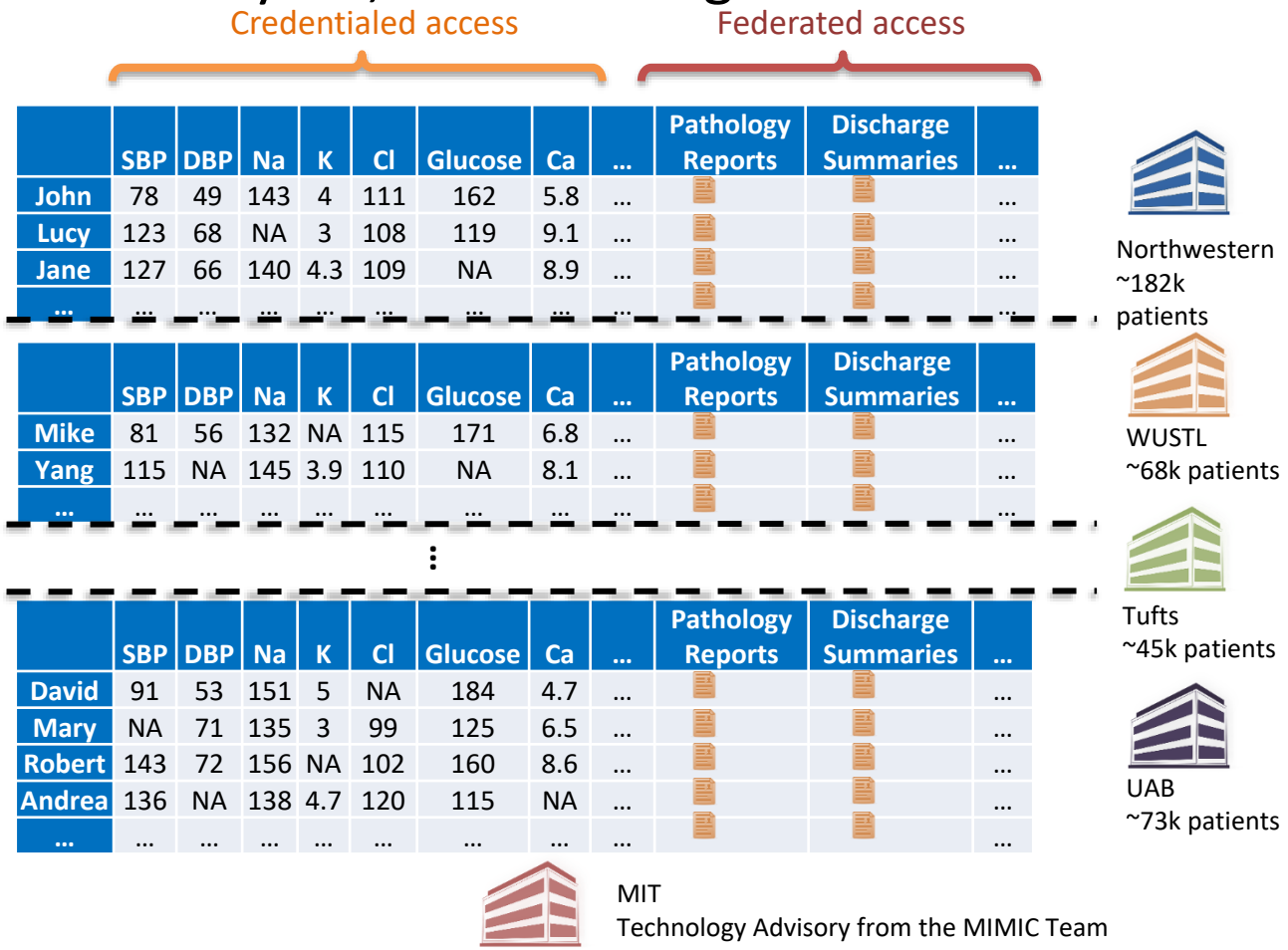
- Digitally recorded data/retrospective
- Missing data
- Single site

TRANSLATIONAL SUPPORT

- Lack of FDA approved tools (0%)
- Ease of use for clinical partnerships
- Clinical relevance is unclear

Creating flagship dataset

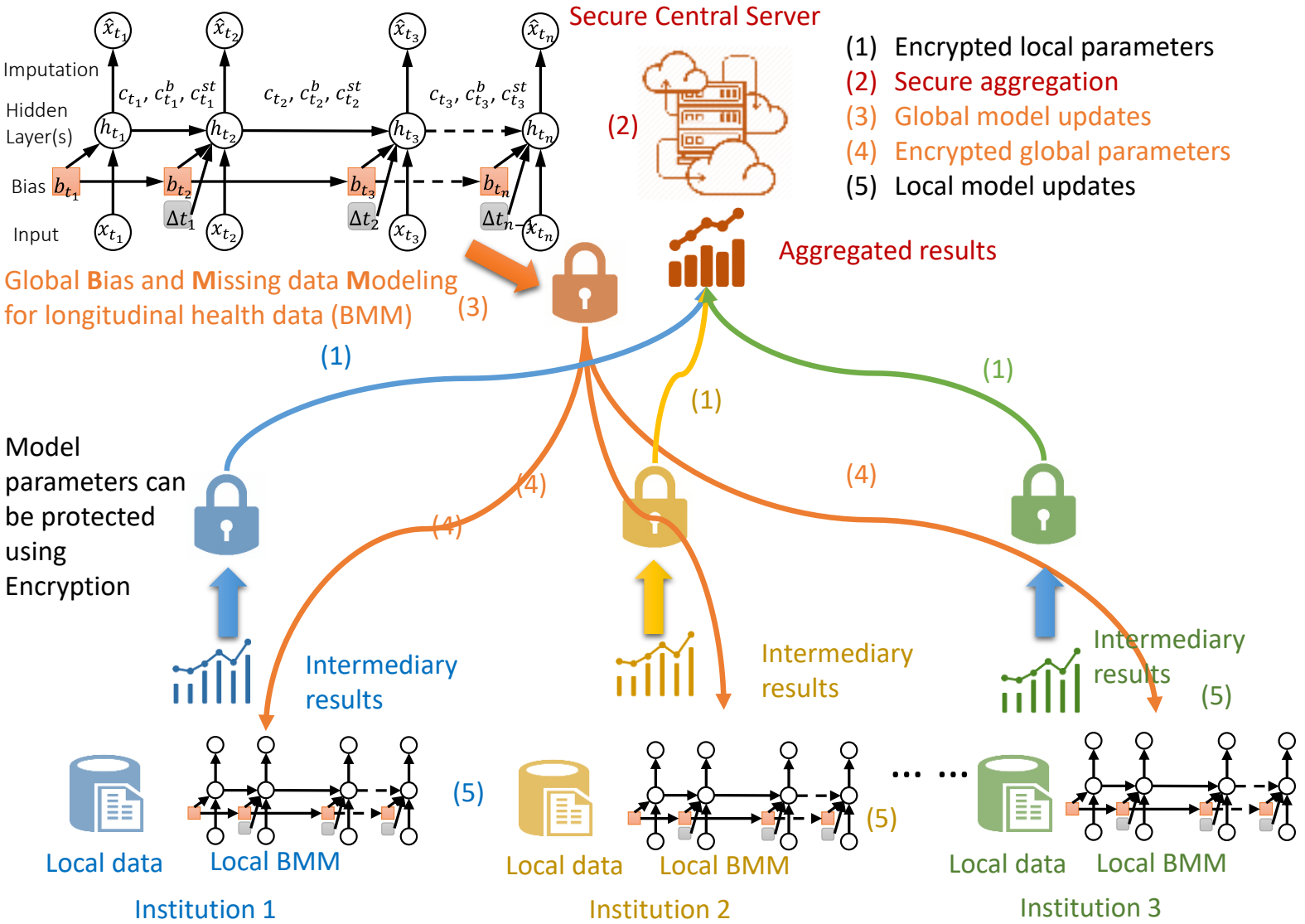
- **C**ollaborative **R**esource for **I**ntensive care **T**ranslational science, **I**nformatics, **C**omprehensive **A**nalytics, and **L**earning



CRITICAL

NIH National Center for Advancing Translational Sciences

Modeling missingness and bias in distributed setting



Growing the tree: AI4H Clinic

- The event is open to clinicians who want to discuss a clinical problem or challenge they face that might be addressable through Artificial Intelligence
- Integrated with classroom teaching
  - Students, guided by instructor, brainstorm on solutions to the clinical problem
  - Students also get the opportunity to work on projects

David Liebovitz, MD,
Internal Medicine

Ceylan Z Cankurtaran,
MD, Radiology

James Adams, MD,
Emergency Medicine

Sadiya Khan, MD, MSc,
Cardiology

Scott Dresden, MD,
Emergency Medicine

Leena Mithal, MD,
Pediatrics

Srikanth Divi, MD,
Orthopaedic Surgery

Abel N Kho, MD
Internal Medicine

The pandemic as a stress test for machine learning in healthcare

REVIEW ARTICLE | FOCUS
https://doi.org/10.1038/s41591-018-0300-7
nature medicine

High-performance medicine: the convergence of human and artificial intelligence

Eric J. Topol

nature machine intelligence
ANALYSIS
https://doi.org/10.1038/s42256-021-00307-0
Check for updates

OPEN

Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans

Michael Roberts[1,2], Derek Driggs[1], Matthew Thorpe[3], Julian Gilbey[1], Michael Yeung[4], Stephan Ursprung[4,5], Angelica I. Aviles-Rivero[1], Christian Etmann[1], Cathal McCague[4,5], Lucian Beer[4], Jonathan R. Weir-McCall[4,6], Zhongzhao Teng[4], Effrossyni Gkrania-Klotsas[7], AIX-COVNET*, James H. F. Rudd[8,36], Evis Sala[4,5,36] and Carola-Bibiane Schönlieb[1,36]

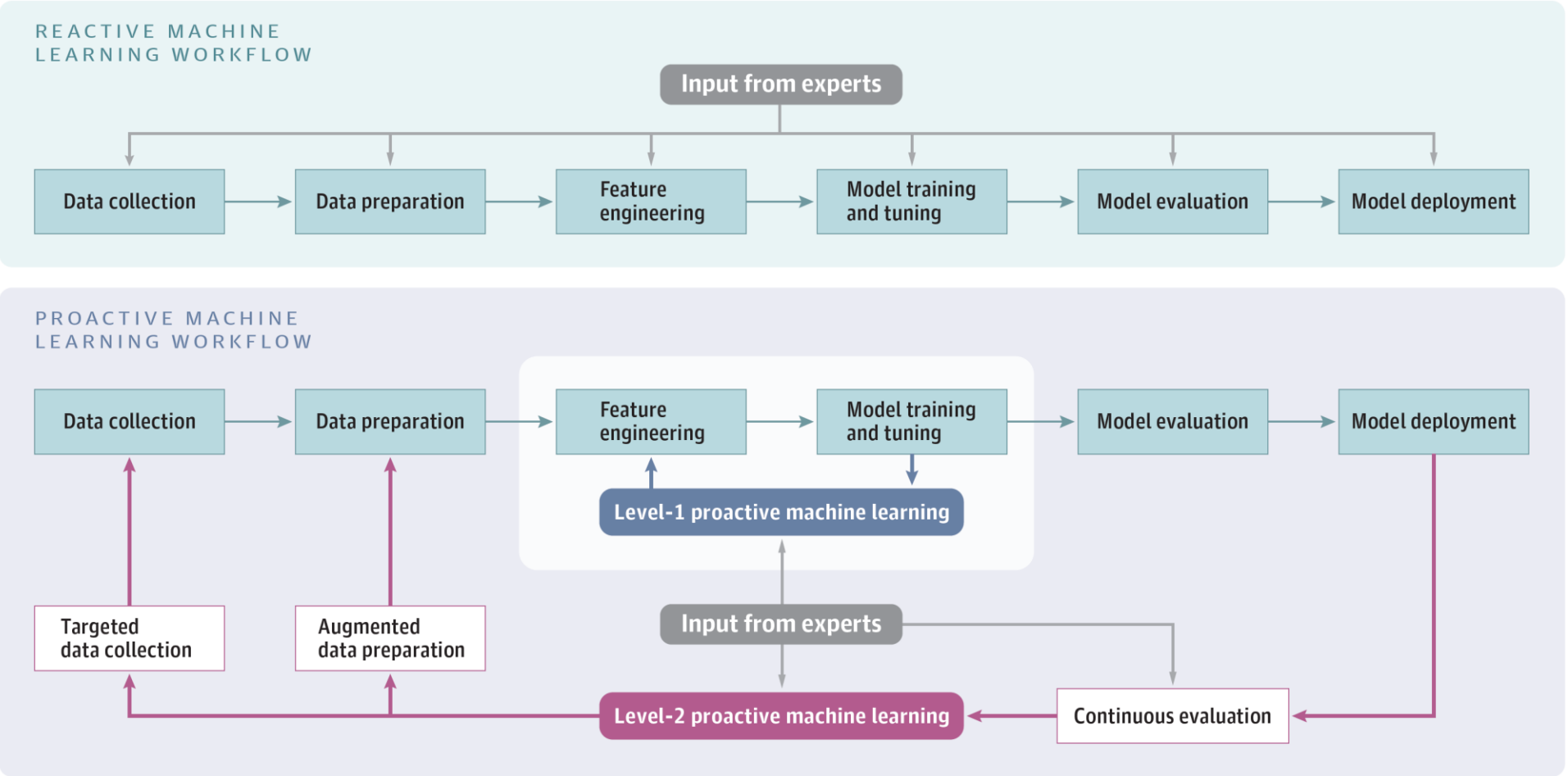
Article

Efficient and targeted COVID-19 border testing via reinforcement learning

https://doi.org/10.1038/s41586-021-04014-z
Received: 16 March 2021
Accepted: 10 September 2021

Hamsa Bastani[1,7], Kimon Drakopoulos[2,7], Vishal Gupta[2,7], Ioannis Vlachogiannis[3], Christos Hadjicristodoulou[4], Pagona Lagiou[5], Gkikas Magiorkinis[5], Dimitrios Paraskevis[5] & Sotirios Tsiodras[6]

Moving from reactive to proactive machine learning



Luo Y, Wunderink RG, Lloyd-Jones D. Proactive vs Reactive Machine Learning in Health Care: Lessons From the COVID-19 Pandemic. *JAMA. 2022.*

References

- [1]M.C. Hochberg, Updating the American College of Rheumatology revised criteria for the classification of systemic lupus erythematosus, *Arthritis Rheum***40**(1997),1725.
- [2]E.M. Tan, A.S. Cohen, and R.J. Winchester, The 1982 revised criteria for the classification of systemic lupus erythematosus, *Arthritis Rheum* **25** (1982), 1271-1277.
- [3]D.L. Linzer, J, poLCA: An R Package for Polytomous Variable Latent Class Analysis *Journal of Statistical Software* **42** (2011), 29.
- [4] Daniel Oberski,2015, latent class analysis, lecture notes, Dept of Methodology and statistics, Tilburg university, delivered 2015

Thank you!

- Collaboration welcome
- yuan.luo@northwestern.edu
- @yuanhypnosluo
- We are hiring, multiple postdocs position available
- https://labs.feinberg.northwestern.edu/ylab/

NUCATS Clinical and Translational Sciences Institute

Northwestern

Northwestern Medicine

eMERGE Network electronic medical records & genomics