abbvie

Northwestern Medicine®
Feinberg School of Medicine

Natural language processing for lupus nephritis computational phenotyping

Yu Deng, PhD, Jennifer A. Pacheco, MS, Anh Chung, MS, Chengsheng Mao, PhD, Chunhua Weng, PhD, Cong Liu, PhD, Adam Cordon, PhD, Jingzhi Yu, BA, Yacob Tedla, PhD, Abel Kho, MD, Rosalind Ramsey-Goldman, MD, Theresa Walunas, PhD, Yuan Luo, PhD

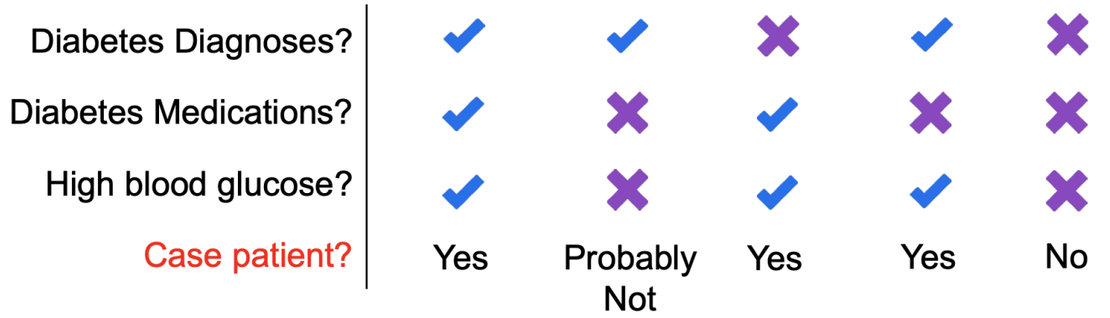
Data & Statistical Sciences

Disclosure

- This publication was neither originated nor managed by AbbVie, and it does not communicate results of AbbVie-sponsored Scientific Research. Thus, it is not in scope of the AbbVie Publication Procedure (PUB-100)
- The authors declare that they have no competing interests.
- This work was supported by National Institute of Arthritis and Musculoskeletal and Skin Diseases (NIAMS5R21AR072263), NIH/NIAMS R61 AR076824-01, NIH/NIAMS 1K08 AR072757-01 and Rheumatology Research Foundation K Supplement Award. This project is part of the Phase III of the eMERGE Network was initiated and funded by the NHGRI through the following grants: U01HG008672 (Vanderbilt University Medical Center); U01HG008680 (Columbia University Health Sciences); and U01HG008673 (Northwestern University). The funding body does not take part in the design of the study and collection, analysis, and interpretation of data and writing the manuscript.

Computational phenotyping in electronic health records

- Computational phenotyping aims to automatically mine or predict clinically significant, or scientifically meaningful, phenotypes from messy EHR data, (Zeng et al., 2018); facilitate genome-wide association study, pragmatic clinical trials, and others
- Challenges in EHR phenotyping
- Need sophisticated Machine learning models or structure data, clinical notes



Lupus nephritis computational phenotyping

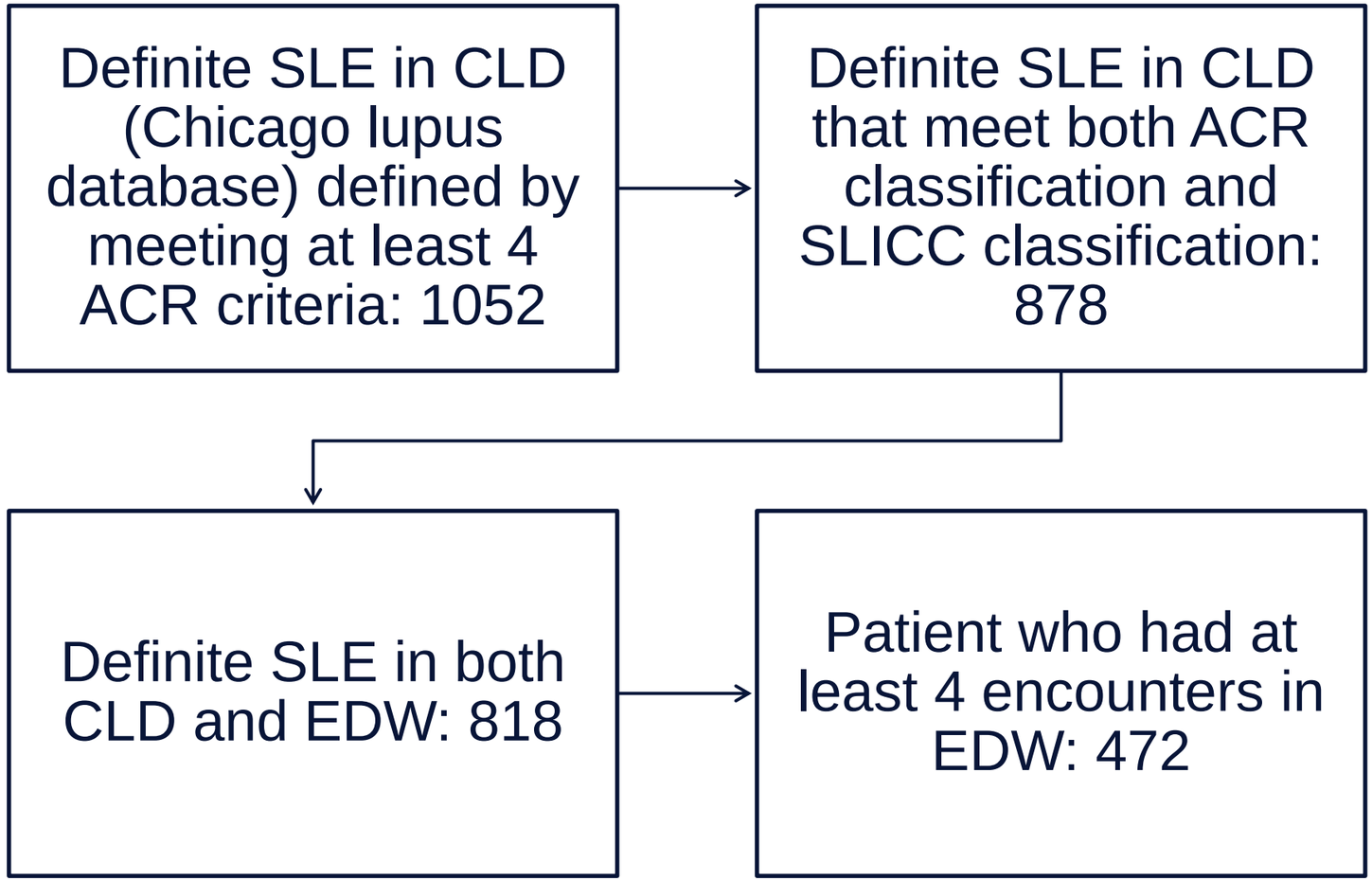
- Lupus nephritis is one of the major risk factors for systemic lupus erythematosus (SLE) mortality.
- Chibnik et al. ICD codes; claims data; positive predictive value (PPV) 88%; sensitivity and specificity not mentioned
- Li et al. used ICD codes; good sensitivity and specificity; low PPV, 63.4%.
-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

Chibnik et al 2010;
Li et al 2020

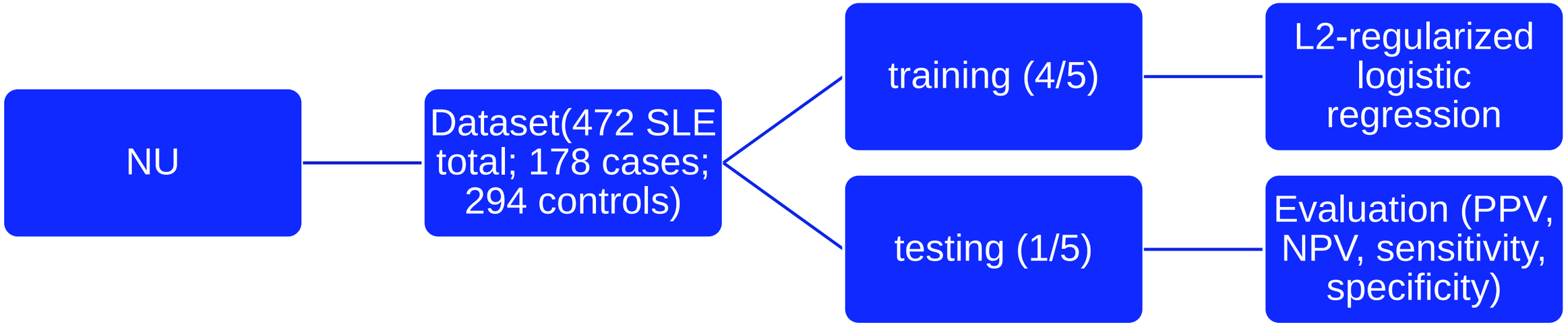
Objectives

- Develop NLP based algorithms to identify lupus nephritis phenotype among SLE patients
- Evaluate the algorithm in an independent dataset

SLE patient cohorts



SLE patient cohorts



Model Descriptions

Models	Algorithm	Features
Benchmark algorithm	Rule-based	A patient is confirmed to have lupus nephritis if he/she has proteinuria>0.5mg in laboratory test or has ICD 9/10 diagnosis code for lupus nephritis.
Full MetaMap model (binary)	L2-regularized logistic regression	Features are all the non-negative mention of MetaMap CUIs. We treated CUIs as binary variables and fitted L2-regularized logistic regression to predict lupus nephritis.
Full MetaMap model (count)	L2-regularized logistic regression	The same as the full MetaMap model (binary) except that MetaMap CUIs are treated as numeric variables representing the count of instances each concept is mentioned in the clinical text.
MetaMap mixed model	L2-regularized logistic regression	There are 13 features in this model including 7 CUI features, 5 RegEx concepts, and 1 feature from structured data.

Model I: Benchmark 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

MetaMap Output

MetaMap: an API developed to map biomedical text to the UMLS Metathesaurus (Aronson et al., 2001)

The UMLS (Unified Medical Language System): a set of files and software that brings together many health and biomedical vocabularies and standards to enable interoperability between computer systems

UMLS Metathesaurus: health/biomedical terms and codes from many vocabularies, including CPT, ICD-10-CM, LOINC, MeSH, RxNorm, and SNOMED CT. Hierarchies, definitions, and other relationships and attributes.

```
00000000|MMI|16.77|Patients|C0030705|[podg]|["*^patient"-tx-5-"Patient"-noun-0,"*^patient"-tx-4-"Patient"-n
00000000|MMI|16.05|Dominican Republic|C0013014|[geoa]|["DR"-tx-1-"Dr"-noun-0]|TX|19/2|Z01.107.084.900.300;Z
00000000|MMI|13.05|Kidney Diseases|C0022658|[dsyn]|["RENAL DISORDER"-tx-4-"renal disorder"-noun-0]|TX|85/14
00000000|MMI|12.90|Physicians|C0031831|[prog]|["DR"-tx-1-"Dr"-noun-0]|TX|19/2|M01.526.485.810;N02.360.810
00000000|MMI|11.49|Oral Ulcer|C0149745|[dsyn]|["ORAL ULCER"-tx-5-"oral ulcer"-noun-0]|TX|132/10|C07.465.672
```

Sample Results:

```
24119710|MMI|637.30|Isopoda|C0598806|[euka]|["Isopod"-ab-1-"isopod"-adj-0,"Isopoda"-ti-1-"Isopoda"-noun-0]|TI;AB|228/6;136/7|B01.050.500.131.365.400
24119710|MMI|585.19|Subphylum Crustacea|C0010395|[euka]|["Crustacea"-ti-1-"Crustacea"-noun-0]|TI|125/9|B01.050.500.131.365
24119710|MMI|170.37|Effect|C1280500|[qlco]|["effects"-ti-1-"Effects"-noun-0]|TI|21/7|
24119710|MMI|3.44|Various patch test substance|C0440102|[irda]|["Various"-ab-1-"various"-adj-0]|AB|322/7|
 5538822|MMI|2.05|Basal Cell|C0596155|[cell]|["basal cells"-ab-41-"basal cells"-noun-0]|AB|7059/5,7073/5|
 5099124|MMI|2.51|Dominant|C1527180|[ftcn]|["Dominant"-ab-7-"dominant"-adj-0]|AB|[1351/8],[1437/8]|
23074497|MMI|7.59|Transplant Recipients|C0376387|[podg]|["Transplant Recipients"-ab-27-"recipients transplants"-noun-0]|AB|[4061/10,4075/11],
                                                                                                                    [4166/10,4180/11]|M01.925
```

Bar "|" Separated Fields Included where Information is Available:

(1) **ID** – Unique identifier used to identify text being processed. If no identifier is found in the text, 00000000 will be displayed. For MEDLINE ASCII formatted input, the ID is considered the PMID.

(2) **MMI** – Always MMI. **Please Note**: This would previously have been shown as MM instead of MMI. ***[Changed in MetaMap 2014]***

(3) **Score** – MetaMap Indexing (MMI) score with a maximum score of 1000.00. The higher the score, the greater the relevance of the UMLS concept according to the MMI algorithm. The MMI results are presented in highest-to-lowest relevance order.

(4) **UMLS Concept Preferred Name** – The UMLS preferred name for the UMLS concept identified in the text.

(5) **UMLS Concept Unique Identifier (CUI)** – The CUI for the identified UMLS concept.

(6) **Semantic Type List** – Comma-separated list of Semantic Type abbreviations for the identified UMLS concept. More information on ***Semantic Types and Semantic Groups*** can be found at (https://metamap.nlm.nih.gov/SemanticTypesAndGroups.shtml).

(7) **Trigger Information** – Comma separated sextuple showing what triggered MMI to identify this UMLS concept. Ordering tends to go from last to first found. The sextuple consists of "UMLS concept-loc-locPos-text-Part of Speech-Negation Flag".

- **UMLS Concept (Preferred or Synonym Text)**
- **loc** – Location in the text if identifiable**.** ti – Title, ab – Abstract, and tx – Free Text
- **locPos** – Number of the utterance within the loc starting with one (1). For example, "ti-1" denotes the first utterance in the title.
- **text** – The actual text mapped to this UMLS concept identification.
- **Part of Speech** – Determined by the MedPost Tagger or Lexical Lookup ***[New field in MetaMap 2014]***
- **Negation Flag** – 1 if the text is considered Negated by MetaMap and 0 if it is not considered Negated. ***[New field in MetaMap 2014]***

(8) **Location** – Summarizes where UMLS concept was found. TI – Title, AB – Abstract, TX – Free Text, TI;AB – Title and Abstract

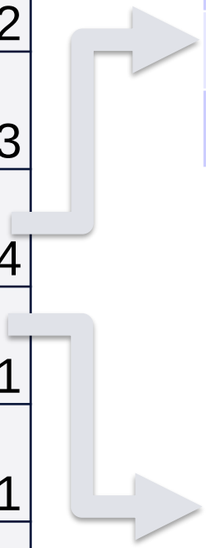
(9) **Positional Information** – Semicolon-separated list of positional-information terns, showing StartPos, slash (/), and Length of each trigger identified in the Trigger Information field. The Positional Information field will contain as many semicolon-separated chunks of positional information as there are sextuples of trigger information in the Trigger Information field. StartPos begins at position zero (0) of the input text. Each semicolon-separated chunk of positional information can take one of four forms:

a) The simplest form: StartPos/Length, e.g., 228/6, 136/7, 125/9, 21/7, 322/7 in the first four examples (all PMID 24119710), meaning that the concept in question was identified in a single span in the text beginning at position 228 and spanning 6 characters, or beginning at position 136 and spanning 7 characters, etc. Note that the first example contains two sextuples in the Trigger Information field, and therefore two semicolon-separated chunks of positional information.

b) Multiple comma-separated StartPos/Length pairs, e.g., 7059/5,7073/5 in the fifth example (PMID 5538822), indicating disjoint text strings mapped to one concept; an example is "basal" and "cells" in the phrase "basal layer cells".

Model II, III

CUI	Num_occr	filename
C0039225	1	p1_note1
C0013058	1	P1_note2
C0013058	2	P1_note3
C0015260	1	P1_note4
C3494508	1	P2_note1
C0039225	2	P3_note1
C0813211	1	P3_note2



Model II: Full MetaMap model (binary)

	C0039225	C0013058	C0015260	C0813211
p1	1	1	1	0
p2	0	1	0	0
p3	1	0	0	1

Model III: Full MetaMap model (counts)

	C0039225	C0013058	C0015260	C0813211
p1	1	3	1	0
p2	0	1	0	0
p3	2	0	0	1

Model Descriptions

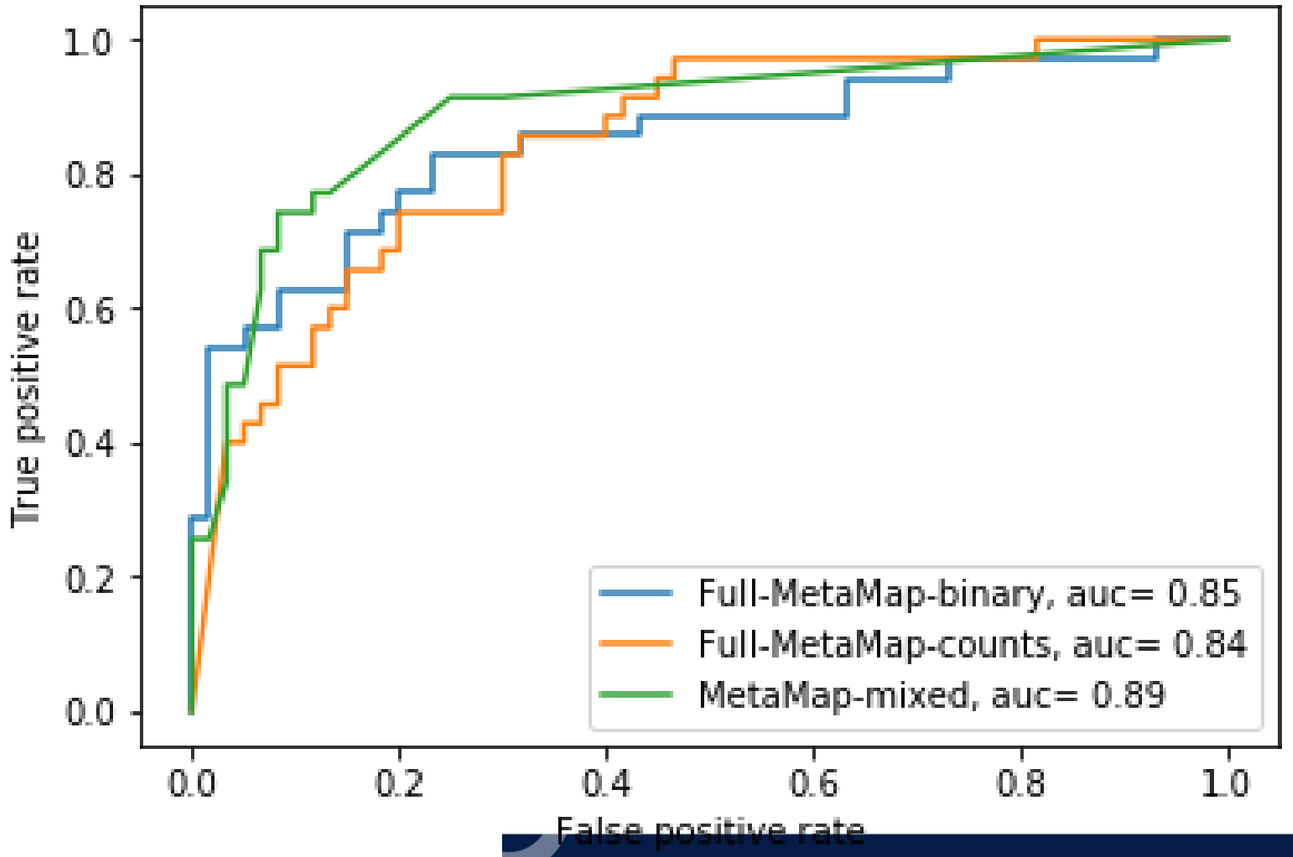
Models	Algorithm	Features
Benchmark algorithm	Rule-based	A patient is confirmed to have lupus nephritis if he/she has proteinuria>0.5mg in laboratory test or has ICD 9/10 diagnosis code for lupus nephritis.
Full MetaMap model (binary)	L2-regularized logistic regression	Features are all the non-negative mention of MetaMap CUIs. We treated CUIs as binary variables and fitted L2-regularized logistic regression to predict lupus nephritis.
Full MetaMap model (count)	L2-regularized logistic regression	The same as the full MetaMap model (binary) except that MetaMap CUIs are treated as numeric variables representing the count of instances each concept is mentioned in the clinical text.
MetaMap mixed model	L2-regularized logistic regression	There are 13 features in this model including 7 CUI features, 5 RegEx concepts, and 1 feature from structured data.

Results

Algorithm	Sensitivity	Specificity	PPV	NPV	F Measure
Baseline	0.43	0.6	0.39	0.64	0.41
Full MetaMap (binary)	0.63	0.92	0.82	0.81	0.71
Full MetaMap (counts)	0.6	**0.95**	**0.88**	0.8	0.71
MetaMap mixed	**0.74**	0.92	0.84	**0.86**	**0.79**

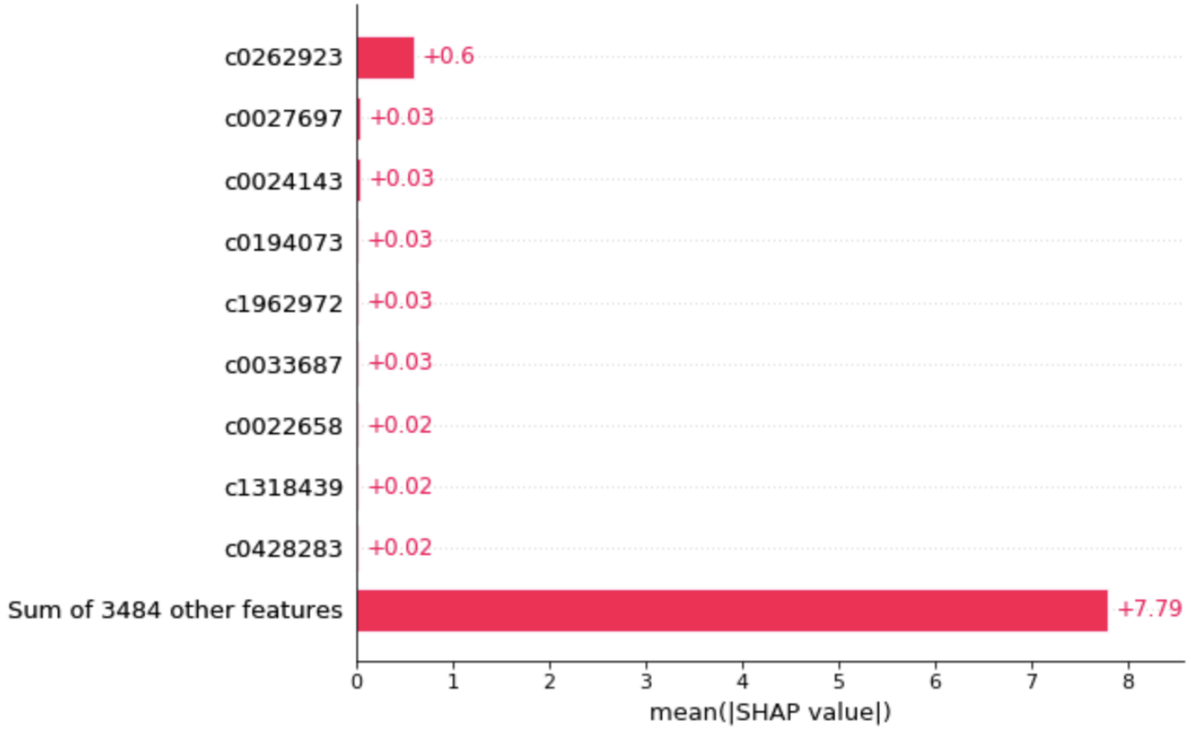
Results

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



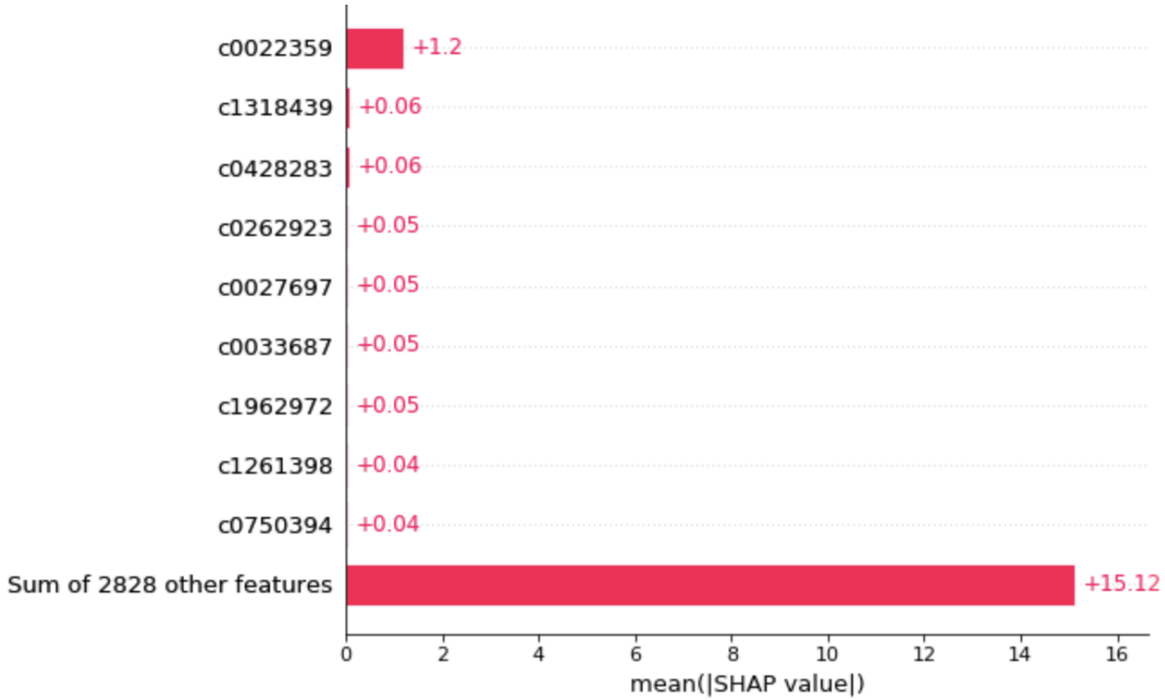
SHAP plot for Full MetaMap (binary) model with SHAP feature importance measured as the mean absolute Shapley values.

- C0262923: Urine protein test
- C0027697: Inflammation of renal tissue;
- C0024143: Glomerulonephritis in the context of systemic lupus erythematosus;
- C0194073: Kidney Biopsy
- C1962972: Proteinuria, CTCAE 3.0;
- C0033687: Proteinuria;



SHAP plot for MetaMap (counts) model. SHAP feature importance measured as the mean absolute Shapley values.

- C1318439: A quantitative measurement of the total amount of creatinine present in a sample of urine;
- C0428283: Urine creatinine level finding;
- C0262923: Urine protein test;
- C0027697: Nephritis;



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 .

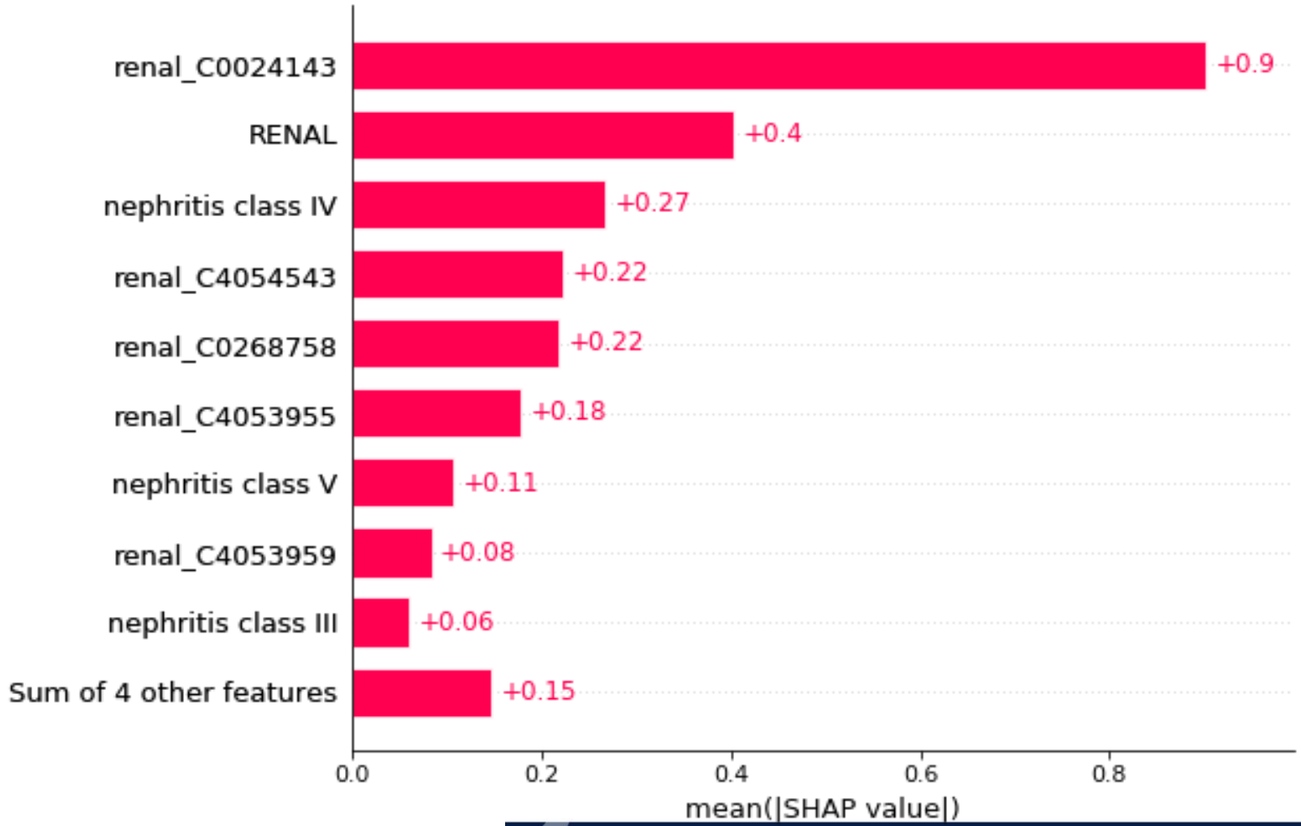
RENAL: renal indictor from structured data;

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) .

References

- Zeng, Zexian, et al. "Natural language processing for EHR-based computational phenotyping." *IEEE/ACM transactions on computational biology and bioinformatics* 16.1 (2018): 139-153.
- Chibnik, Lori B., E. M. Massarotti, and Karen H. Costenbader. "Identification and validation of lupus nephritis cases using administrative data." *Lupus* 19.6 (2010): 741-743.
- Li, Tingting, et al. "Development and validation of lupus nephritis case definitions using United States veterans affairs electronic health records." *Lupus* 30.3 (2021): 518-526.
- Aronson, Alan R. "Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program." *Proceedings of the AMIA Symposium*. American Medical Informatics Association, 2001.