BMC Medical Informatics and Decision Making

RESEARCH **Open Access**

Natural language processing to identify lupus nephritis phenotype in electronic health records

Yu Deng[1], Jennifer A. Pacheco[2], Anika Ghosh[1], Anh Chung[1,6], Chengsheng Mao[1], Joshua C. Smith[3], Juan Zhao[3], Wei-Qi Wei[3], April Barnado[4], Chad Dorn[3], Chunhua Weng[5], Cong Liu[5], Adam Cordon[2], Jingzhi Yu[1], Yacob Tedla[1], Abel Kho[1], Rosalind Ramsey-Goldman[6], Theresa Walunas[1*†] and Yuan Luo[1*†]

From International Conference on Intelligent Biology and Medicine (ICIBM 2021)
Philadelphia, PA, USA. 8-10 August 2021. https://icibm2021.iaibm.org/

Abstract

Background Systemic lupus erythematosus (SLE) is a rare autoimmune disorder characterized by an unpredictable course of flares and remission with diverse manifestations. Lupus nephritis, one of the major disease manifestations of SLE for organ damage and mortality, is a key component of lupus classification criteria. Accurately identifying lupus nephritis in electronic health records (EHRs) would therefore benefit large cohort observational studies and clinical trials where characterization of the patient population is critical for recruitment, study design, and analysis. Lupus nephritis can be recognized through procedure codes and structured data, such as laboratory tests. However, other critical information documenting lupus nephritis, such as histologic reports from kidney biopsies and prior medical history narratives, require sophisticated text processing to mine information from pathology reports and clinical notes. In this study, we developed algorithms to identify lupus nephritis with and without natural language processing (NLP) using EHR data from the Northwestern Medicine Enterprise Data Warehouse (NMEDW).

Methods We developed five algorithms: a rule-based algorithm using only structured data (baseline algorithm) and four algorithms using different NLP models. The first NLP model applied simple regular expression for keywords search combined with structured data. The other three NLP models were based on regularized logistic regression and used different sets of features including positive mention of concept unique identifiers (CUIs), number of appearances of CUIs, and a mixture of three components (i.e. a curated list of CUIs, regular expression concepts, structured data) respectively. The baseline algorithm and the best performing NLP algorithm were externally validated on a dataset from Vanderbilt University Medical Center (VUMC).

†Theresa Walunas and Yuan Luo contributed equally to this work.

*Correspondence:
Theresa Walunas
t-walunas@northwestern.edu
Yuan Luo
yuan.luo@northwestern.edu
Full list of author information is available at the end of the article

Results Our best performing NLP model incorporated features from both structured data, regular expression concepts, and mapped concept unique identifiers (CUIs) and showed improved F measure in both the NMEDW (0.41 vs 0.79) and VUMC (0.52 vs 0.93) datasets compared to the baseline lupus nephritis algorithm.

Conclusion Our NLP MetaMap mixed model improved the F-measure greatly compared to the structured data only algorithm in both internal and external validation datasets. The NLP algorithms can serve as powerful tools to accurately identify lupus nephritis phenotype in EHR for clinical research and better targeted therapies.

Keywords Natural language processing, Electronic health records, Computational phenotyping, Lupus nephritis

Introduction

Systemic Lupus Erythematosus (SLE) is an autoimmune disease that has diverse manifestations, resulting in significant morbidity and mortality [1, 2]. While many autoimmune diseases, such as rheumatoid arthritis, have benefitted from new classes of medications, SLE has seen few advancements in therapy in the last 50 years [3]. It has been hypothesized that the heterogeneity of SLE presentations may make it challenging to understand therapeutic responses across the full scope of SLE presentations and that observational cohort studies and clinical trials would benefit from targeting subpopulations with similar disease presentations [4]. Recently, the Food and Drug Administration has approved two new medications for use in managing lupus nephritis, increasing the urgency of identifying lupus nephritis in people with SLE to ensure the new therapeutics can be targeted to these patients to help reduce kidney damage and improve long term outcomes [5]. Classification criteria for SLE describe a broad range of evidence-based clinical and laboratory descriptors. There are three criteria currently in use: 1) the set developed in 1982 and revised in 1997 by the American College of Rheumatology (ACR) [6], 2) the set developed by the System Lupus International Collaborating Clinics in 2012 (SLICC) [2], and 3) the set developed by the European League Against Rheumatism / American College of Rheumatology (EULAR/ACR) criteria set [7]. Lupus nephritis is one of the most common and severe manifestations of SLE. Approximately 40% of SLE patients develop lupus nephritis [8] and it is included in all three classification criteria sets. In both the SLICC and EULAR/ACR criteria, one way to be classified as "definite lupus" is having a positive anti-nuclear antibody/anti-dsDNA screen in the presence of renal biopsy-proven lupus nephritis [2, 7]. Thus, lupus nephritis is a critical attribute to describe for clinical and research applications and the identification of SLE subpopulations, but often it requires time consuming chart adjudication to identify patients who satisfy this criterion.

Electronic health records (EHRs) are a readily available data source that includes a record of clinical care and procedures, diagnoses, laboratory test results, medication orders, and clinical notes for describing disease manifestations in persons with SLE. EHRs have been demonstrated useful in genome association studies, drug comparative effectiveness studies [9, 10], and others. However, a large amount of information in the EHR, such as histology notes for kidney biopsies, is generally only located in text-based notes from which it is challenging to extract information using simple rule-based identification algorithms and text string searches [11, 12]. Several prior studies developed algorithms to identify lupus nephritis using administrative claims data [13]. Chibnik et al. identified lupus nephritis in claims data and reached a positive predictive value (PPV) of 88% but sensitivity and specificity were not mentioned [14]. Li et al. used various combinations of International Classification of Diseases (ICD) codes to identify lupus nephritis [15]. Their algorithm achieved good sensitivity and specificity but a low positive predictive value (PPV) of 63.4%. Most of these studies only used structured data (i.e. ICD codes, laboratory test value), and the algorithms were often not validated in an external dataset [14, 15]. Thus, correctly identifying lupus nephritis from EHR for large cohort studies, in addition to identifying critical procedures, diagnoses and lab results, also requires the development of natural language processing (NLP) tools that can utilize histology reports and clinical notes. Previously, studies with other structured data-based concepts (e.g. multiple sclerosis, rheumatoid arthritis) have demonstrated that NLP can significantly improve rate of identification [11, 16].

In this study, we focus on the identification of lupus nephritis in the SLICC criteria in EHR data using NLP technologies to mine clinical notes and pathology reports. To do this, we compared algorithms for the identification of lupus nephritis based on structured data alone to four different NLP models to determine whether NLP could improve identification of persons with lupus nephritis. Our approach facilitates accurate identification of lupus nephritis in the EHR, enabling researchers to better understand patients' SLE characteristics and serving as a foundation for lupus nephritis-related large cohort observational studies and clinical trials. We trained and evaluated the performance of all four algorithms in a dataset from Northwestern Medicine

Electronic Data Warehouse (NMEDW) and then further validated the performance in an external dataset from Vanderbilt University Medical Center (VUMC).

Methods

Data source

The Chicago Lupus Database (CLD), established in 1991, is a registry database specifically designed for lupus related studies. It is a physician validated registry of 1,052 patients with possible or definite lupus according to the 1982 American College of Rheumatology classification criteria revised in 1997 [17, 18]. The patients in the CLD met at least three ACR criteria (step 1 in Fig. 1). Among the 1052 patients in the CLD, 878 patients had definite lupus according to the Systemic Lupus International Collaborating Clinics (SLICC) classification criteria (step 2 in Fig. 1) [2]. Among these patients, 178 have lupus nephritis according to the definition in SLICC. The presence or absence of lupus nephritis in patients in the CLD is verified by the physician chart review.

The Northwestern Medicine Electronic Data Warehouse (NMEDW) is the primary data repository for all the medical records of patients who receive care within the Northwestern Medicine system [19]. Established in 2007, the NMEDW contains records for over 3.8 million patients, with most EHR data going back to at least 2002, and with some billing claims data going back to 1998 or earlier. By linking patients in the CLD to patient records in the NMEDW through their medical record numbers, we identified 818 definite SLE patients based on SLICC criteria who were both in the CLD and the NMEDW (see step 3 in Fig. 1). To ensure our patient cohort has sufficient depth of data in both data sources, we excluded any patients who had less than four clinical encounters documented in the NMEDW [20, 21], reducing the final case cohort size to 472 (see step 4 in Fig. 1). All inpatient and outpatient notes from transplant, nephrology, and rheumatology departments were retrieved. The retrieved clinical narratives included pathology reports, progress notes, consult notes, and discharge notes.

Algorithm development

In this study, we focus on identifying renal criterion/lupus nephritis in the SLICC classification which is defined as "having a urine protein/creatinine ratio (or 24-h urine protein collection) equivalent to 500 mg of protein per 24-h period, or red blood cell casts in the urine" [2]. The renal criterion/lupus nephritis in the SLICC classification includes both biopsy-proven and non-biopsy-proven nephritis. To set up the gold standard label for lupus nephritis, the physicians in our team who are expert clinicians on lupus, performed chart review using data from the CLD which has more in-depth information on lupus related information compared to EHR data. The physicians also excluded other causes of glomerular disease when adjudicating the diagnosis of lupus nephritis.

We developed five algorithms (see Table 1 for the overview of the five algorithms) to identify lupus nephritis from SLE patients' EHR data including a baseline algorithm that used only structured data and four NLP models that used structured data and clinical notes. In the baseline algorithm, a patient is classified as lupus nephritis based on ICD9/10 diagnosis codes and laboratory test results. The details of the structured data used in the baseline algorithm are shown in Additional file 1: Table S1. For the NLP models, following the steps in Zeng et al. [22–24], we extracted different feature sets for model implementation including concept unique identifier (CUI) features and regular expression (regex) matches from the notes. For the CUI features, we first preprocessed the notes by removing duplicated records and tokenizing sentences. We then applied MetaMap to annotate medical concepts in each sentence [25]. MetaMap is an NLP application that maps biomedical text to the Unified Medical Language System (UMLS) Metathesaurus and assigns a CUI to each word or term [26]. Any CUIs recognized as being negated by MetaMap (i.e., "no glomerulonephritis") were excluded. For regex features, five concepts were used as features, including nephritis class II, nephritis class III, nephritis class IV, nephritis

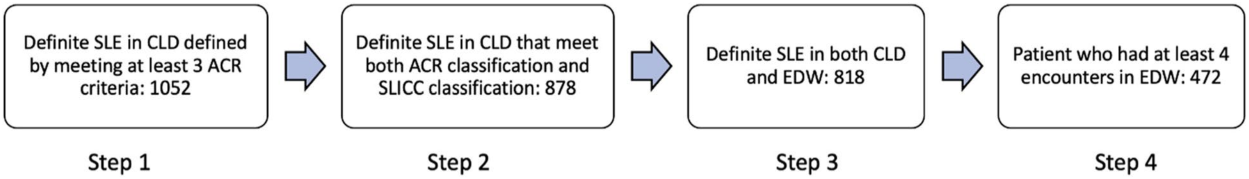
Definite SLE in CLD defined by meeting at least 3 ACR criteria: 1052
Definite SLE in CLD that meet both ACR classification and SLICC classification: 878
Definite SLE in both CLD and EDW: 818
Patient who had at least 4 encounters in EDW: 472
Step 1
Step 2
Step 3
Step 4

Fig. 1 SLE case cohort selection process. We identified 1052 SLE patients who met at least 3 ACR criteria based on physician chart review. Among these 1052 patients, we further identified 878 patients who also met SLICC classification criteria. Among the 878 patients, 818 patients were in NMEDW. We further restricted our study cohort to patients who had at least 4 encounters in the NMEDW which left 472 patients in the final cohort. Abbreviations: ACR criteria, American College of Rheumatology Classification Criteria; CLD, Chicago Lupus Database; SLE, systemic lupus erythematosus; NMEDW, Northwestern Medicine Enterprise Data Warehouse

Table 1 Algorithm description

Algorithm name	Classification model	Description
Baseline algorithm	Rule-based	A patient is confirmed to have lupus nephritis if he/she has proteinuria > 0.5 mg in laboratory test or has ICD 9/10 diagnosis code for lupus nephritis.
Full MetaMap model (binary)	L2-regularized logistic regression	Features are 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.

class V, and proteinuria. We developed regular expression patterns to search for text related to the five concepts (see Additional file 1: Table S1 for the list of regex patterns). We built four NLP models using different feature sets. In the first NLP model, we implemented rule-based algorithm using both regex features and structured data. A patient is classified as lupus nephritis if they have any match for the regex patterns, ICD 9/10 codes, or laboratory test of interest (see Additional file 1: Table S1, S2). For the other three NLP models, we implemented an L2-regularized logistic regression classifier. We chose L2-regularized logistic regression because it can handle high dimensional feature space and multicollinearity problems by penalizing its coefficients in the loss function. In addition, the model is straightforward, and model output is easy to interpretate. We tried both L1 and L2-regularized logistic regression and selected the latter because it generates equivalent if not superior performance compared to L1-regularized logistic regression in our NU dataset. In the first L2-regularized logistics regression-based NLP model– the full MetaMap (binary) model, all positive mentioned MetaMap CUIs were used as binary type features. In the second L2-regularized logistics regression-based NLP model– the full MetaMap (count) model, the number of occurrences for every positive mapped CUIs were used as features. The minimum document frequency was set as 30 and 40 in MetaMap (binary) model and MetaMap (count) model, respectively to avoid feature sparsity. The frequencies were chosen by trying a list of frequencies and the ones generated the highest F measure were selected. In the last L2-regularized logistics regression-based NLP model– the MetaMap mixed model, we used a mixture of lupus nephritis related CUIs, structured data, and regex concepts as features. The CUIs include C0024143, C0268757, C0268758, C4053955, C4053958, C4053959, C4054543 (see Additional file 1: Table S3 for each CUI definition). For the structured data component, a single binary feature is used. A patient is indicated to be positive for the structured data feature if he/she is predicted positive in the baseline algorithm. There were 13 variables in total for the MetaMap mixed model including 7 features from CUIs, 5 lupus nephritis related concepts for regex expression search, and 1 feature from structured data.

Model training and evaluation

We split the data from NMEDW into training (75%) and testing datasets (25%). In the training dataset, to get the optimal hyperparameter, we used grid search on parameter *C*, which is the inverse of regularization strength, ranging from 1e-5 to 1e5 with interval spacing equal to 10. For the L2-regularized logistics regression-based NLP models, we selected "sag" method as our optimizer [27]. We set the class weight as balanced to adjust for disproportionate class frequencies. Parameters that generated the best accuracy were retained. We evaluated our model in the testing set (internal validation) based on sensitivity, specificity, PPV, negative predictive value (NPV), F measure, and area under the curve (AUC). We further explored feature contribution by extracting the top 5 features with the highest positive coefficient in MetaMap (binary), MetaMap (count), and MetaMap mixed model, respectively. We also evaluated feature importance by generating mean absolute Shapley value (SHAP) plots. L2-regularized logistic regression was conducted using 'scikit-learn' library in Python, version 3.7.3. Regular expression was performed using 're' package in Python, version 3.7.3 [27, 28]. Shapley value was generated using 'shap' package in Python, version 3.7.3. SHAP plot was generated using 'matplotlib' package in Python, version 3.7.3.

External validation

We further validated both the baseline algorithm, and the best performing NLP model (based on results from the testing set at Northwestern University site) in an external validation dataset at Vanderbilt University Medical Center (VUMC), a regional, tertiary care center [29, 30]. The VUMC data warehouse contains over 3.2 million subjects with de-identified clinical records from the EHR collected across the past several decades. We

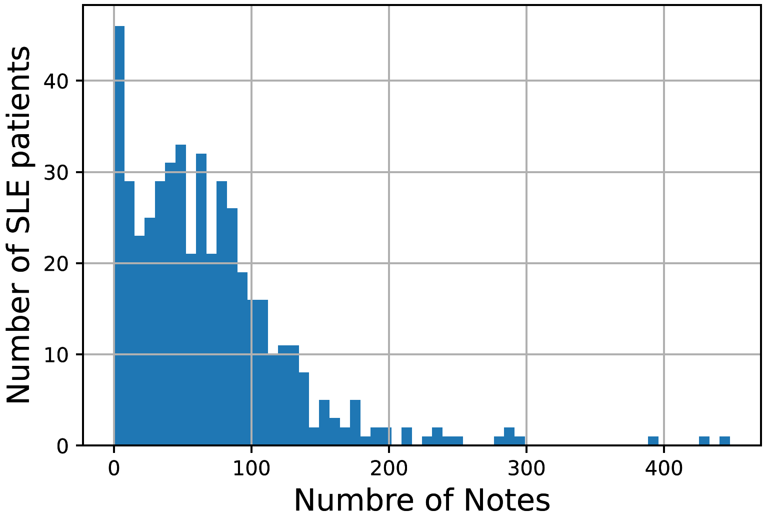

Fig. 2 Histogram of note count per patient for SLE patients

first performed a simple SLE phenotyping algorithm based on SLE ICD9/10 codes to get a SLE cohort (not chart reviewed) on which to run our lupus nephritis algorithm. We then randomly selected 75 patients on which to evaluate our lupus nephritis algorithm. A rheumatologist manually reviewed the chart for these 75 patients. Among these patients, there were 18 patients with definite lupus, 1 with possible SLE, and 56 with no SLE. Among these 75 patients, there were 14 patients with lupus nephritis all of whom had definite lupus and 61 patients without. We evaluated the F measure, sensitivity, specificity, PPV, and NPV for the lupus nephritis baseline algorithm and NLP model with the highest F measure based on the results from the Northwestern University (NU) dataset. F measure evaluates the accuracy of the algorithm, it is calculated as the following:

$$F\ measure = \frac{2 * precision * recall}{precision + recall},$$

Here precision and recall are also known as PPV and sensitivity, respectively.

Results

Among the 472 SLE patients at NU, there were 178 patients (37.7% of the cohort) who developed lupus nephritis. The average number of notes per patient is 68.58 (standard deviation [SD] = 59.37). The distribution of the number of notes for the patient cohort is shown in Fig. 2. Out of the 472 patients, 206 had ICD codes related to lupus nephritis, 4 had red blood cell cast test, and 230 had urine protein test results available.

The performance for the five algorithms is shown in Table 2. All four NLP models have higher sensitivity, specificity, PPV, and NPV compared to the baseline algorithm using structured data alone. All the logistics regression-based NLP models had higher F measure compared to rule-based NLP model using structured data and regex patterns. The full MetaMap (binary)

Table 2 Model performance

Dataset	Algorithm	Sensitivity	Specificity	PPV	NPV	F Measure
NU (testing set)	Baseline	0.43	0.6	0.39	0.64	0.41
NU (testing set)	Regex + structured	0.49	0.93	0.81	0.76	0.61
NU (testing set)	Full MetaMap (binary)	0.63	0.92	0.82	0.81	0.71
NU (testing set)	Full MetaMap (counts)	0.6	0.95	0.88	0.80	0.71
NU (testing set)	MetaMap mixed	0.74	0.92	0.84	0.86	0.79
VUMC	Baseline	0.86	0.67	0.38	0.95	0.52
VUMC	MetaMap mixed	0.93	0.98	0.93	0.98	0.93

For logistic regression-based models, probability of 0.5 is used as the threshold for classification

Abbreviations: *SLE* systemic lupus erythematosus, *NU* Northwestern University, *VUMC* Vanderbilt University Medical Center, *NLP* natural language processing, *PPV* positive predictive value, *NPV* negative predicted value model has higher sensitivity compared to the full MetaMap (count) model, (0.63 vs 0.6), NPV (0.81 vs 0.8), and comparable F measure (0.71 vs 0.71). The MetaMap mixed model has higher sensitivity (0.74) and NPV (0.86) as well as F measure (0.79) compared to the other two models. Similarly, MetaMap mixed model has higher AUC (0.89) compared to full MetaMap (binary) (AUC = 0.85) and full MetaMap (count) (AUC = 0.84) model (see Fig. 3). Therefore, we selected the MetaMap mixed model as the final NLP model to be validated at VUMC in addition to the baseline algorithm. In the VUMC dataset, which included 75 patients, the MetaMap mixed model has higher sensitivity, specificity, PPV, and NPV compared to the baseline algorithm. The F measure improved from 0.52 to 0.93 as shown in Table 2.

In terms of feature importance, the top 5 features with the highest positive coefficient for each classifier are shown in Table 3. C0024143 (lupus nephritis) appears to have high positive coefficient in all three L2-regualrized classifiers. C1962972 (proteinuria finding) are the 4th highest positive coefficient in both MetaMap (binary) and MetaMap (count) model. Our full MetaMap models are able to pick up many important lupus nephritis related concepts such as kidney disease, proteinuria, lupus nephritis as high coefficient features. The SHAP plot shows the top 10 most important features for classification in each model. As shown in Figs. 4, 5 and 6, most of the important features are related to lupus nephritis clinically.

Discussion

In this study, we developed five algorithms to identify lupus nephritis: a baseline algorithm using structured data only, a rule-based model using regex and structured data, a full MetaMap model with binary features, a full MetaMap model with count features, and a MetaMap mixed model. In the NU testing dataset, the MetaMap mixed model outperformed (F measure = 0.79) both the baseline algorithm (F measure = 0.41) and the other two

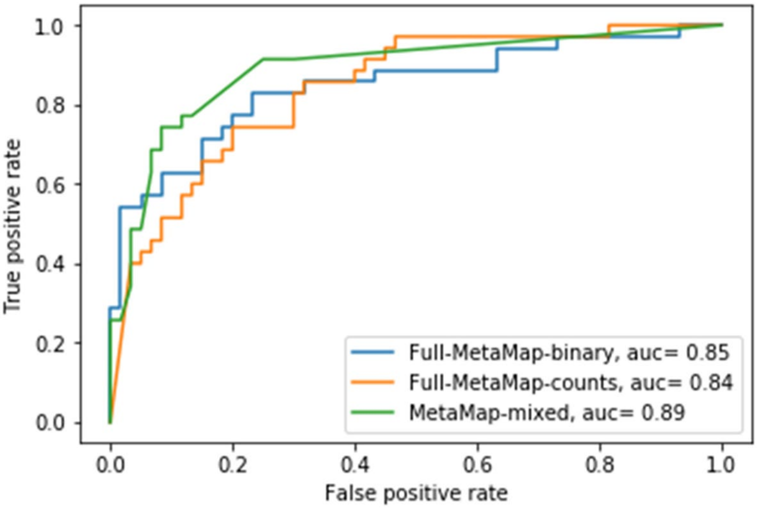

Fig. 3 Area under the curve (AUC) for Full MetaMap (binary), Full MetaMap (counts), and MetaMap mixed model in NU testing set

Table 3 Top 5 positive coefficient for each classifier

Coefficient ranking	Full MetaMap (binary)		Full MetaMap (count)		MetaMap Mixed	
	Feature, definition	Coefficient	Feature, definition	Coefficient	Feature, definition	Coefficient
1	C0027697, nephritis	0.04	C0022646, kidney	5.01	C0024143, lupus nephritis	1.26
2	C0024143, lupus nephritis	0.04	C0024143, lupus nephritis	4.79	'RENAL'	0.64
3	C0022658, kidney disease	0.03	C0033687, proteinuria	3.58	'nephritis class IV'	0.54
4	C1962972, proteinuria finding	0.03	C1962972, proteinuria finding	3.53	'proteinuria > 0.5 gm'	0.45
5	C003368, sultroponium	0.03	C1707664, Delayed Release Dosage Form	3.52	C4053955, SLE class IV	0.25

Features are ranked by the value of their associated coefficients

RENAL renal indictor from structured data only, *gm* gram

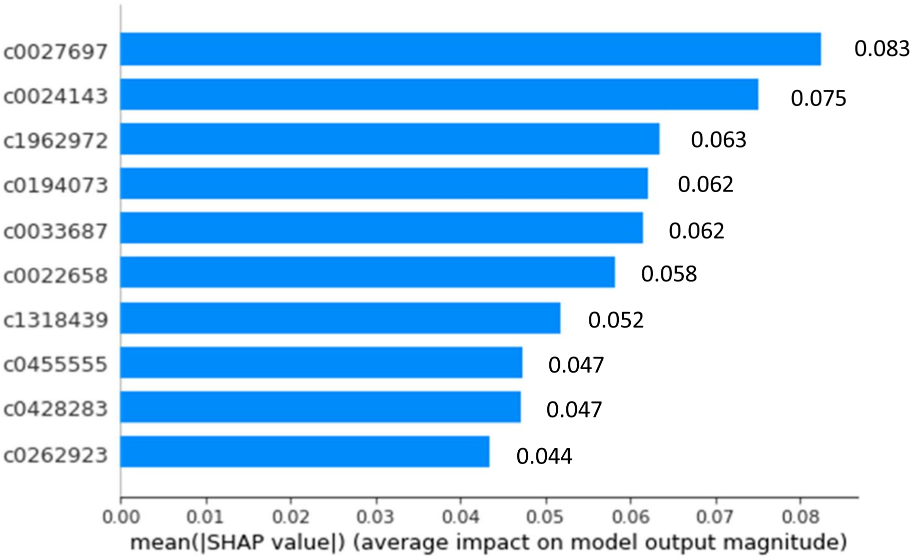

Fig. 4 SHAP plot for full MetaMap (binary) model with SHAP feature importance measured as the mean absolute Shapley values. The Features are ordered according to their importance. The SHAP bar plot shows global importance of each feature which is taken to be the mean absolute SHAP value for that feature over all the given samples. The most important feature in this plot is C0027697 which has a global importance of 0.083 compared to an average feature global importance of 0.006. C0027697: nephritis; C0024143: lupus nephritis; C0194073: kidney biopsy; C0033687: proteinuria; C0022658: kidney diseases; C1318439: urine creatinine measurement; C045555: H/O: nephritis; C0428283: urine creatinine level finding; C0262923: Urine protein test

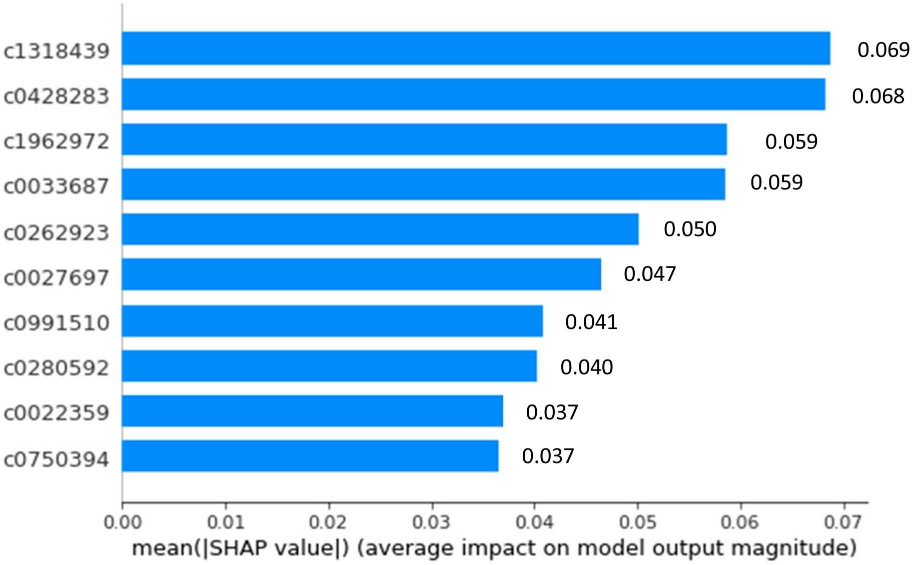

Fig. 5 SHAP plot for full MetaMap (count) model with SHAP feature importance measured as the mean absolute Shapley values. The Features are ordered according to their importance. The SHAP bar plot shows global importance of each feature which is taken to be the mean absolute SHAP value for that feature over all the given samples. The most important feature in this plot is C1318439 which has a global importance of 0.069 compared to an average feature global importance of 0.005. C1318439: Urine creatinine measurement; C0428283: Urine creatinine level finding; C1962972: Proteinuria, CTCAE 3.0; C0033687: Proteinuria; C0262923: Urine protein test; C0027697: Nephritis; C0991510: Foam drug form; C0280592: doxorubicin/fluorouracil/mitomycin/vincristine protocol; C0022359: jaw. C0750394: white blood cell count decreased

NLP models (F measure = 0.71, 0.71 respectively). In the VUMC validation dataset, the MetaMap mixed model significantly improved the F measure over the baseline algorithm (0.93 versus 0.52).

Error analysis

In the MetaMap mixed model, we investigated 10 SLE patients in the training set that were wrongly classified by L2-regularized logistic regression. One patient was

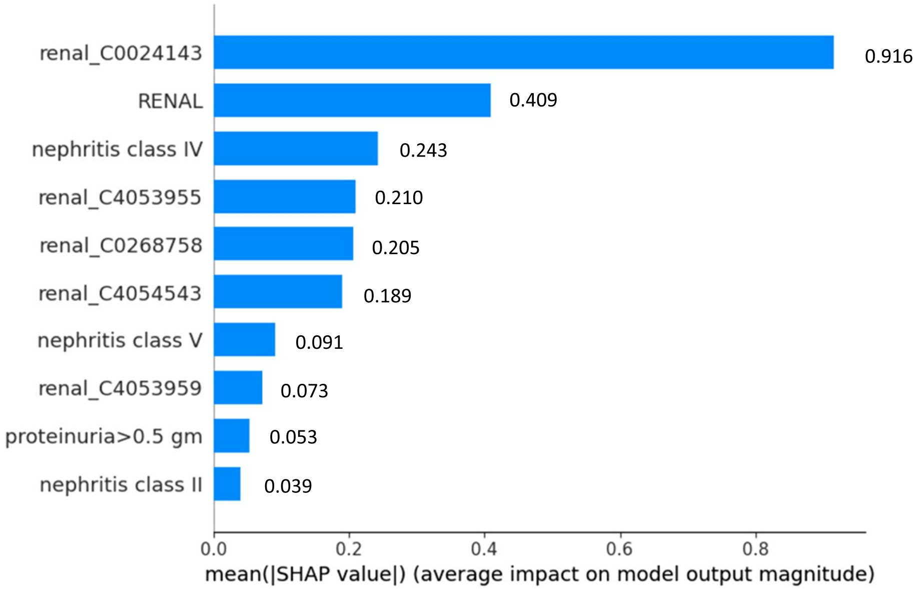

Fig. 6 SHAP plot for MetaMap mixed model with SHAP feature importance measured as the mean absolute Shapley values. The Features are ordered according to their importance. The SHAP bar plot shows global importance of each feature which is taken to be the mean absolute SHAP value for that feature over all the given samples. The most important feature in this plot is renal_C002413 which has a global importance of 0.916 compared to an average feature global importance of 0.192. 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

wrongly predicted as negative for lupus nephritis with a 0.49 probability of having lupus nephritis. In the feature set the algorithm identified, the patient was positive for CUI C002413 (glomerulonephritis in the context of systemic lupus erythematosus) and was negative for all the other features. It was mentioned in the notes that the patient had 'stage 2 LN'. Lupus nephritis class II is one of the features used in our algorithm. However, our regex did not include this specific variation of wording for lupus nephritis class II. This pattern could be incorporated in the NLP in the future to improve algorithm performance.

In another example, a 26-year-old female was wrongly predicted as positive for lupus nephritis with a probability of 0.53 of having lupus nephritis. In the feature set the algorithm identified, the patient was positive for C0024143 (glomerulonephritis in the context of systemic lupus erythematosus) and proteinuria features both of which were positively associated with lupus nephritis. Our algorithm showed that the patient had matched for 'proteinuria > 0.5' in the notes which was in the context of 'negative renal disorder: either persistent proteinuria (> 0.5 g/day or + ++) or cellular casts'. Our regex pattern was not able to capture the negation at the beginning of the sentence. Therefore, it falsely predicted the patient as positive for lupus nephritis.

All NLP models outperformed the baseline algorithm in the NU testing (internal validation) dataset. In the baseline model, 20/35 lupus nephritis patients were wrongly classified as non-lupus nephritis patients, while the MetaMap mixed model reduced the misclassified cases to 9/35. The baseline algorithm relies solely on ICD 9/10 diagnosis and laboratory test results. In the baseline rule-based algorithm, laboratory tests missing from the EHR largely influenced the performance. In the NLP MetaMap mixed model, using features from multiple modalities (EHR notes-derived regex, CUIs features, laboratory tests, and ICD codes) that complement each other, and a penalized logistic regression model improved the accuracy and generalizability of the model. As part of the future work, we plan to apply advanced imputation methods [31, 32] to fill in missing laboratory tests in order to further improve the phenotyping performance.

Limitations

Our study has certain limitations. Firstly, even though we had physician adjudication to set up gold standard label, this could still be imperfect in the cases of lack of patient biopsy or other information to identify the true label of lupus nephritis. Although, the impact of such was minimized by using the CLD registry as the data source which has more in-depth lupus related information compared to EHR data and can help more accurately set up the gold standard label. Secondly, we only had 75 patients in the VUMC validation dataset.

This is due to limited resources for chart review. The prevalence of lupus nephritis is high among our lupus population in the external VUMC dataset. This is likely contributing to the fact that VUMC is a tertiary care center which has sicker patients and the small sample size which may increase the chance of sample bias. Future study is needed to further validate our algorithm performance in a larger external dataset.

Conclusion

In conclusion, we developed five algorithms, a structured data only algorithm and four NLP models, to identify lupus nephritis phenotypes. We evaluated the algorithms in an internal and an external validation dataset. All four NLP models outperformed the baseline algorithm in the internal validation dataset. In the external validation dataset, our NLP MetaMap mixed model improved the F-measure greatly compared to the structured data only algorithm. Our NLP algorithms can serve as powerful tools to accurately identify lupus nephritis phenotype in EHR for clinical research and better targeted therapies.

Abbreviations

SLE	Systemic lupus erythematosus
EHRs	Electronic health records
NLP	Natural language processing
NMEDW	Northwestern Medicine Enterprise Data Warehouse
VUMC	Vanderbilt University Medical Center
CUI	Concept unique identifier
ACR	American College of Rheumatology
SLICC	System Lupus International Collaborating Clinics
EULAR/ACR	European League Against Rheumatism / American College of Rheumatology
ICD	International Classification of Disease
PPV	Positive predictive value
NPV	Negative predictive value
CLD	Chicago Lupus Database
Regex	Regular expression
UMLS	Unified Medical Language System (UMLS) Metathesaurus
NU	Northwestern University

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12911-024-02420-7.

Additional file 1: Table S1. ICD-9/10 codes and LOINC codes used for baseline algorithm. **Table S2.** Regex concepts and their associated searching keywords for lupus nephritis. **Table S3.** CUIs and their definition.

Acknowledgements

Not applicable.

About this supplement

This article has been published as part of BMC Medical Informatics and Decision Making Volume 22 Supplement 2, 2022: Selected articles from the International Conference on Intelligent Biology and Medicine (ICIBM 2021): medical informatics and decision making. The full contents of the supplement are available online at https://bmcmedinformdecismak.biomedcentral.com/articles/supplements/volume-22-supplement-2.

Authors' contributions

YD led the study, performed all data analyses, and wrote the manuscript. YL and TW designed the study. YL and TW supervised the project. RRG provided clinical expertise on interpreting the data. JAP coordinated the project, assisted with the design of the algorithms, implemented the algorithm without NLP, and provided expertise in EHR data analysis. RRG, TW, AC, AB, WW, and CD made substantial contribution to the data acquisition. AG assisted to develop structured data only and regex-based algorithms. All the other authors read, edited, and approved the final manuscript.

Funding

This work was supported by National Institute of Arthritis and Musculoskeletal and Skin Diseases (NIAMS 5R21AR072263), 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 publication fee will be charged through the eMERGE funding (U01HG008673). The funding body does not take part in the design of the study and collection, analysis, and interpretation of data and writing the manuscript.

Availability of data and materials

The datasets generated and analyzed during the current study are not publicly available due to protected patient information but are available from the corresponding author on reasonable request.

Declarations

Ethics approval and consent to participate

This study was a retrospective study of existing records. Ethics approval was provided by Northwestern University Institutional Review Board and Vanderbilt University Institutional Review Board.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]Center for Health Information Partnerships, Feinberg School of Medicine, Northwestern University, Chicago, USA. [2]Center for Genetic Medicine, Feinberg School of Medicine, Northwestern University, Chicago, USA. [3]Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, USA. [4]Department of Medicine, Vanderbilt University Medical Center, Nashville, USA. [5]Department of Biomedical Informatics, Columbia University, New York City, USA. [6]Department of Medicine/Rheumatology, Feinberg School of Medicine, Northwestern University, Chicago, USA.

Received: 9 April 2021 Accepted: 9 January 2024
Published online: 03 March 2024

References

1. Almaani S, Meara A, Rovin BH. Update on lupus nephritis. Clin J Am Soc Nephrol. 2017;12(5):825–35. https://doi.org/10.2215/CJN.05780616.
2. Petri M, et al. Derivation and validation of the systemic lupus international collaborating clinics classification criteria for systemic lupus erythematosus. Arthritis Rheum. 2012;64(8):2677–86. https://doi.org/10.1002/art.34473.
3. Dörner T, Furie R. Novel paradigms in systemic lupus erythematosus. Lancet. 2019;393(10188):2344–58. https://doi.org/10.1016/S0140-6736(19)30546-X.
4. Murphy G, Isenberg DA. New therapies for systemic lupus erythematosus — past imperfect, future tense. Nat Rev Rheumatol. 2019;15(7):403–12. https://doi.org/10.1038/s41584-019-0235-5.
5. FDA approves first oral therapy for lupus nephritis. https://www.hcplive.com/view/fda-approves-first-oral-therapy-voclosporin-for-lupus-nephritis. Accessed 23 Jan 2024.

6. Hochberg MC. Updating the American College of Rheumatology revised criteria for the classification of systemic lupus erythematosus. Arthritis Rheum. 1997;40(9):1725. https://doi.org/10.1002/art.1780400928.
7. Aringer M, et al. 2019 European league against rheumatism/American college of rheumatology classification criteria for systemic lupus erythematosus. Ann Rheum Dis. 2019;78(9):1151–9. https://doi.org/10.1136/annrheumdis-2018-214819.
8. Hoover PJ, Costenbader KH. Insights into the epidemiology and management of lupus nephritis from the US rheumatologist's perspective. Kidney Int. 2016;90(3):487–92. https://doi.org/10.1016/j.kint.2016.03.042.
9. Deng Y, Ghamsari F, Lu A, Yu J, Zhao L, Kho AN. Use of real-world evidence data to evaluate the comparative effectiveness of second-line type 2 diabetes medications on chronic kidney disease. J Clin Transl Endocrinol. 2022;30:100309.
10. Deng Y. Advancing computational methods to derive insights from real-world health data. Doctor, Northwestern University, ProQuest Dissertations and Theses database. 2022.
11. Zeng Z, Deng Y, Li X, Naumann T, Luo Y. Natural language processing for EHR-based computational phenotyping. IEEE/ACM Trans Comput Biol Bioinform. 2019;16(1):139–53. https://doi.org/10.1109/TCBB.2018.2849968.
12. Luo Y, Uzuner O, Szolovits P. Bridging semantics and syntax with graph algorithms-state-of-the-art of extracting biomedical relations. Brief Bioinform. 2017;18(4):722. https://doi.org/10.1093/bib/bbx048.
13. Moores KG, Sathe NA. A systematic review of validated methods for identifying systemic lupus erythematosus (SLE) using administrative or claims data. Vaccine. 2013;31(Suppl 10):K62-73. https://doi.org/10.1016/j.vaccine.2013.06.104.
14. Chibnik LB, Massarotti EM, Costenbader KH. Identification and validation of lupus nephritis cases using administrative data. Lupus. 2010;19(6):741–3. https://doi.org/10.1177/0961203309356289.
15. Li T, et al. Development and validation of lupus nephritis case definitions using United States veterans affairs electronic health records. Lupus. 2021;30(3):518–26. https://doi.org/10.1177/0961203320973267.
16. Liao KP, Cai T, Savova GK, Murphy SN, Karlson EW, Ananthakrishnan AN, et al. Development of phenotype algorithms using electronic medical records and incorporating natural language processing. BMJ. 2015;350. https://www.bmj.com/content/350/bmj.h1885.full.
17. Chicago Lupus Database: Systemic Lupus Research Studies: Feinberg School of Medicine: Northwestern University. https://www.lupus.northwestern.edu/research/cld.html. Accessed 23 Jan 2024.
18. Hochberg MC. Updating the American College of Rheumatology revised criteria for the classification of systemic lupus erythematosus. Arthritis and Rheumatism. 1997;40(9):1725.
19. Northwestern Medicine Enterprise Data Warehouse (NMEDW): Research: Feinberg School of Medicine: Northwestern University. https://www.feinberg.northwestern.edu/research/cores/units/edw.html. Accessed 23 Jan 2024.
20. Rasmussen LV, et al. Design patterns for the development of electronic health record-driven phenotype extraction algorithms. J Biomed Inform. 2014;51:280–6. https://doi.org/10.1016/j.jbi.2014.06.007.
21. Zhong Y, Rasmussen L, Deng Y, Pacheco J, Smith M, Starren J, et al. Characterizing design patterns of EHR-driven phenotype extraction algorithms. In: 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE; 2018. p. 1143–6. https://ieeexplore.ieee.org/abstract/document/8621240/.
22. Zeng Z, et al. Using natural language processing and machine learning to identify breast cancer local recurrence. BMC Bioinformatics. 2018;19(17):65–74.
23. Zeng Z et al. Identifying breast cancer distant recurrences from electronic health records using machine learning. J Healthc Inform Res. 2019:1–17. https://doi.org/10.1007/s41666-019-00046-3.
24. Zeng Z, et al. Contralateral breast cancer event detection using natural language processing. In: AMIA Annual symposium proceedings. American Medical Informatics Association; 2017. p. 1885–92. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5977664/.
25. MetaMap - a tool for recognizing UMLS concepts in text. https://www.nlm.nih.gov/research/umls/implementation_resources/metamap.html. Accessed 23 Jan 2024.
26. Unified Medical Language System (UMLS). https://www.nlm.nih.gov/research/umls/index.html. Accessed 23 Jan 2024.
27. sklearn.linear_model.Ridge — scikit-learn 0.23.2 documentation.
28. re — Regular expression operations — Python 3.9.2rc1 documentation.
29. Vanderbilt University Medical Center. https://www.vumc.org/main/home. Accessed 23 Jan 2024.
30. Research Data Warehousing | Department of Biomedical Informatics. https://www.vumc.org/dbmi/research-data-warehousing. Accessed 23 Jan 2024.
31. Luo Y, Szolovits P, Dighe AS, Baron JM. 3D-MICE: integration of cross-sectional and longitudinal imputation for multi-analyte longitudinal clinical data. J Am Med Inform Assoc. 2017;25(6):645–53. https://doi.org/10.1093/jamia/ocx133.
32. Luo Y. Evaluating the state of the art in missing data imputation for clinical data. Brief Bioinform. 2021. https://doi.org/10.1093/bib/bbab489.

Publisher's Note