LUPUS SCIENCE & MEDICINE™

Development and validation of a risk scoring system to identify patients with lupus nephritis in electronic health record data

Zara Izadi ,[1] Milena Gianfrancesco,[1] Christine Anastasiou,[1] Gabriela Schmajuk,[1,2] Jinoos Yazdany [1]

To cite: Izadi Z, Gianfrancesco M, Anastasiou C, *et al*. Development and validation of a risk scoring system to identify patients with lupus nephritis in electronic health record data. *Lupus Science & Medicine* 2024;**11**:e001170. doi:10.1136/lupus-2024-001170

► Additional supplemental material is published online only. To view, please visit the journal online (https://doi.org/10.1136/lupus-2024-001170).

Received 3 February 2024
Accepted 20 April 2024

[1]University of California San Francisco, San Francisco, California, USA
[2]San Francisco VA Medical Center, San Francisco, California, USA

Correspondence to
Dr Zara Izadi; zaraizadiucsf@gmail.com

ABSTRACT

Objective Accurate identification of lupus nephritis (LN) cases is essential for patient management, research and public health initiatives. However, LN diagnosis codes in electronic health records (EHRs) are underused, hindering efficient identification. We investigated the current performance of International Classification of Diseases (ICD) codes, 9th and 10th editions (ICD9/10), for identifying prevalent LN, and developed scoring systems to increase identification of LN that are adaptable to settings with and without LN ICD codes.

Methods Training and test sets derived from EHR data from a large health system. An external set comprised data from the EHR of a second large health system. Adults with ICD9/10 codes for SLE were included. LN cases were ascertained through manual chart reviews conducted by rheumatologists. Two definitions of LN were used: strict (definite LN) and inclusive (definite, potential or diagnostic uncertainty). Gradient boosting models including structured EHR fields were used for predictor selection. Two logistic regression-based scoring systems were developed ('LN-Code' included LN ICD codes and 'LN-No Code' did not), calibrated and validated using standard performance metrics.

Results A total of 4152 patients from University of California San Francisco Medical Center and 370 patients from Zuckerberg San Francisco General Hospital and Trauma Center met the eligibility criteria. Mean age was 50 years, 87% were female. LN diagnosis codes demonstrated low sensitivity (43–73%) but high specificity (92–97%). LN-Code achieved an area under the curve (AUC) of 0.93 and a sensitivity of 0.88 for identifying LN using the inclusive definition. LN-No Code reached an AUC of 0.91 and a sensitivity of 0.95 (0.97 for the strict definition). Both scoring systems had good external validity, calibration and performance across racial and ethnic groups.

Conclusions This study quantified the underutilisation of LN diagnosis codes in EHRs and introduced two adaptable scoring systems to enhance LN identification. Further validation in diverse healthcare settings is essential to ensure their broader applicability.

WHAT IS ALREADY KNOWN ON THIS TOPIC

⇒ Diagnosis codes for lupus nephritis (LN) are significantly underused in real-world data sources, making identification of this patient population challenging. Research studies focused on creating algorithms to identify LN in real-world data are limited. Currently, there are no externally validated algorithms to identify adult patients with LN from electronic health record (EHR) or claims data.

WHAT THIS STUDY ADDS

⇒ The present study quantifies the limitations of using International Classification of Diseases (ICD) codes for LN diagnosis and introduces novel scoring systems tailored to different data availability scenarios and desired levels of accuracy, for prevalent LN identification from EHRs, offering practical solutions for healthcare providers and researchers. The study addresses the external validity, calibration and performance of these scoring systems across different racial and ethnic groups, facilitating assessments of their applicability in diverse patient populations.

HOW THIS STUDY MIGHT AFFECT RESEARCH, PRACTICE OR POLICY

⇒ LN identification in real-world data is feasible and can be achieved with good accuracy, with important applications for patient management, research and public health initiatives. While text-mining methods and language models have the potential to perform exceptionally for prevalent LN identification from clinical notes, the implementation of these methods in real-world settings is challenging due to the complexities of unstructured clinical text and the need for advanced natural language processing. Therefore, adaptable scoring systems that use data from structured EHR fields, while upholding predictive performance, offer a practical approach to LN identification for both clinical and research applications in diverse healthcare settings.

INTRODUCTION

Real-world data from electronic health records (EHRs) and medical claims have the potential to provide valuable information in the study of lupus nephritis (LN), including its epidemiology, the quality of care provided to patients, renal outcomes and even missed

diagnoses if these cases could be correctly identified. In clinical practice, accurate identification of prevalent cases of LN is essential for timely patient monitoring, treatment and informing public health initiatives for the management of LN.

Several high-quality population-based studies estimate the prevalence of LN to be between 20% and 65%[1–4]; however, diagnosis codes for LN are significantly underused in real-world data sources, making identification of this patient population challenging. A recent analysis of the national Rheumatology Informatics System for Effectiveness Registry, a large EHR data repository, demonstrated that among 13 416 patients with SLE identified using International Classification of Diseases (ICD9 and 10) codes, only 689 (5.1%) patients had codes indicating LN.[5] Given the known prevalence of LN, these data suggest significant undercoding of this manifestation in rheumatology practices.

Research studies focused on creating algorithms to identify LN in real-world data are limited. One prior study published over 12 years ago showed that a combination of lupus and renal ICD9 codes and nephrologist encounter claims had a positive predictive value (PPV) of 88% for the identification of patients with LN from a single medical centre.[6] However, this study was small, predated ICD10 coding, did not use EHR data and did not include an external validation in other health systems. Another study from 2013 showed significant variability in the number of LN cases identified from Impact, a commercial insurance claims database, depending on the approach used for LN identification; this study did not include clinical validation.[7] In a recent study, an algorithm including both utilisation and diagnostic criteria was validated using data from PEDSnet, a national healthcare systems network, achieving 90% sensitivity and 93% specificity, for identifying paediatric patients with LN.[8] Currently, there are no externally validated algorithms to identify adult patients with LN from EHR or claims data.

The gold standard for diagnosis of LN is demonstration of characteristic inflammatory findings on renal biopsy.[9] However, data suggest that biopsies are not always available either because the diagnosis was made clinically or because pathology reports may not be available in EHR data, especially if biopsies were done remotely or at outside facilities. Therefore, identification of LN in EHR data often requires detailed chart reviews to confirm physician reasoning or to find notes referencing remote procedures. Algorithms developed using data from structured fields in the EHR (including laboratory results, medications, diagnostic codes, referral orders) have the potential to address undercoding of LN diagnoses and to identify incident and prevalent cases of LN; such algorithms could reduce the need for chart review, making large-scale research more efficient and helping clinicians with population health management.

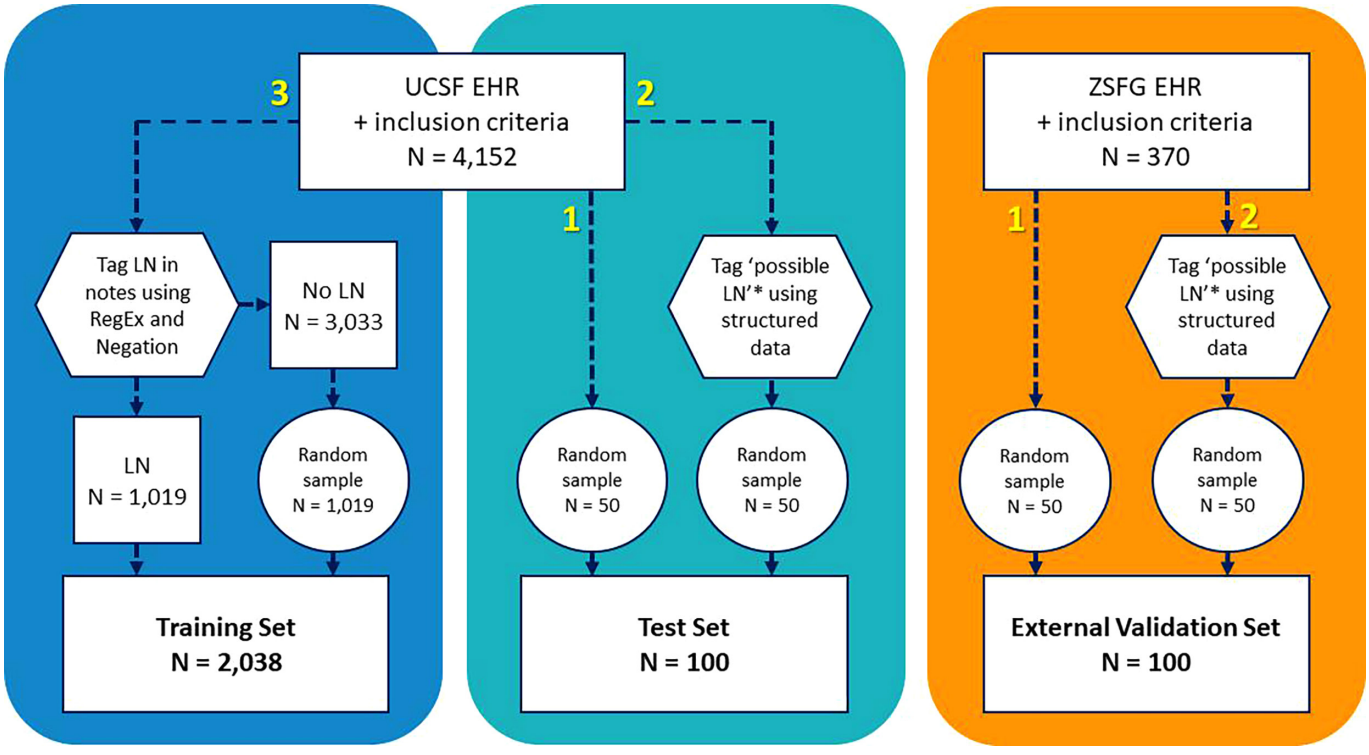

Figure 1 Data partitioning and an overview of methodology for the development of the training, test and external validation sets. Inclusion criteria: adults with SLE, defined as having one or more International Classification of Diseases (ICD) 9/10 codes for SLE, and available notes from June 2012 to January 2022. *Possible LN was defined as either (1) documentation of any ICD codes for chronic kidney disease, or (2) documentation of one or more urine protein-to-creatinine ratios ≥0.5 g/g. EHR, electronic health record; LN, lupus nephritis; UCSF, University of California San Francisco; ZSFG, Zuckerberg San Francisco General.

In this study, we aimed to quantify the performance of ICD9 and 10 codes for identifying prevalent LN using a manual chart review of the EHR as the gold standard. We also aimed to develop simple-to-use scoring systems to identify patients with prevalent LN for use in settings where LN diagnosis codes are available as well as those where LN diagnosis codes are unavailable or heavily underused.

METHODS

Data source

We used structured data and clinical notes from the EHRs of the University of California San Francisco Medical Center (UCSF), and Zuckerberg San Francisco General Hospital and Trauma Center (ZSFG), two large health systems in the San Francisco Bay Area. Eligible patients included adults with SLE, defined as having one or more ICD9/10 codes for SLE,[10] and available notes from June 2012 to January 2022. We followed the Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis statement for prediction model development and validation.[11]

Data partitioning

Data were partitioned into a training set, test set and external validation set. The training set included patients with SLE from UCSF (excluding those in the test set; n=2038). To achieve a balanced distribution of LN, the training set included all eligible patients who were identified as having prevalent LN and an equal number of eligible patients without LN selected at random (figure 1). In the training set, we used regular expressions ('lupus nephritis' and related terms) with negation (of, for example, 'family history of LN', 'no evidence of LN', etc) to loosely tag prevalent LN.

The test set and the external validation set comprised random samples of 100 patients meeting the eligibility criteria from UCSF and ZSFG, respectively—none of the patients in the test set or external validation set were included in the training set. For each health system, we included a random sample of 50 patients and a random sample of 50 patients with 'possible' LN defined as either (1) documentation of any ICD codes for acute or chronic kidney disease, or (2) documentation of any urine protein-to-creatinine ratio (UPCr) >0.5 g/g; these 'possible' cases were included to ensure good representation of LN in the samples.

Outcome ascertainment

Prevalent LN, which included current or previous disease, was defined as documentation of LN diagnoses within EHR notes. A structured manual chart review was used to identify patients with prevalent LN in the test set and the external validation set. The chart review was performed by two rheumatologists; a third rheumatologist adjudicated any discordant results. Patients were classified as having 'no LN', 'definite LN' (an available biopsy report, or reference to a biopsy report demonstrating WHO[12] or International Society of Nephrology/Renal Pathology Society[13] LN class III, IV or V), 'potential LN' (no biopsy report or reference, but physician-diagnosed LN based on clinical presentation) and 'diagnostic uncertainty' (physician states LN is possible but not certain). We chose to focus on class III, IV and V LN since these forms require treatment and can be associated with renal decline if not addressed. For the purposes of this study, LN was defined using a 'strict' (definite LN vs all other categories) and an 'inclusive' (definite LN, potential LN or diagnostic uncertainty vs no LN) definition.

Data preparation

Dichotomous (yes, no) variables were derived from structured fields of EHRs of both health systems to indicate any documentation of LN diagnoses (ICD10: M32.14, M32.15, or ICD9: 710.0 plus 580–583), or acute or chronic kidney disease (ICD10: N00–N08, N17–N19 and R80, or ICD9: 580–586 and 791.0), or proteinuria (ICD10: N00–N08, N17–N19, R80, or ICD9: 580–586, 791.0). Similarly, dichotomous variables were derived to indicate any documentation of face-to-face nephrology encounters, dialysis procedure codes, diagnoses of hypertension, as well as abnormal laboratory test results and medications potentially related to the treatment of LN. Abnormal laboratory results included urinalysis tests positive for protein or red blood cells, UPCr or urine albumin–creatinine ratios >0.5 g/g or timed urine protein measurements >0.5 g/24 hours, two consecutive readings of estimated glomerular filtration rate <60 mL/min, detected anti-double-stranded DNA, low C3 or C4 levels, and low serum albumin levels. Medications included any use of intravenous methylprednisolone, cyclophosphamide (oral or intravenous), mycophenolate or mycophenolic acid, belimumab (intravenous or subcutaneous), voclosporin, tacrolimus, azathioprine, rituximab, obinutuzumab, hydroxychloroquine, and oral prednisone or equivalent oral glucocorticoid. Since individuals with LN often require moderate or high doses of steroids for treatment, use of prednisone-equivalent glucocorticoid doses >10 mg/day, >20 mg/day and >40 mg/day was also separately coded as dichotomous variables. Demographic variables included age (as a continuous variable), race and ethnicity, and sex. Number of face-to-face nephrology encounters, maximum UPCr ratio and maximum prednisone-equivalent glucocorticoid doses were also included as continuous variables. Patients with no documented UPCr ratios or prednisone-equivalent glucocorticoid doses had their maximum UPCr ratio and maximum prednisone-equivalent glucocorticoid dose imputed to 0.

Predictor selection

We used machine learning for predictor selection since the approach is suited to data with high dimensionality, as well as an ensemble classifier that has been shown to perform well with clinical data.[14–16] A series of gradient boosting models (GBMs) including various predictor subsets were used for predictor selection. The most

comprehensive model included 40 predictors that covered diagnosis codes, nephrology encounters, procedure codes, comorbidities, laboratory test results, medications and demographics (input variables described above). GBMs were trained on prevalent LN using three repeats of 10-fold cross-validation for hyperparameter optimisation. All analyses were performed in R V.3.6.1, using the Classification and Regression Training package.[17]

Development of the LN scoring systems

We chose to develop scoring systems that would be adaptable to the desired accuracy in applying them.[18 19] The scoring systems were based on multivariable logistic regression models that included predictors ranked by order of importance by GBM.[20] To maximise model sensitivity (recall), predictors were added to the logistic regression model in order of importance, until there were no further incremental gains in sensitivity. To improve regression fit, we assessed linearity in the relationship between continuous predictors and the outcome and used categorisation or truncation (of values exceeding 95th percentile) when necessary. In addition, all two-way interactions were tested and incorporated in the model if statistically significant ($p<0.05$).

To facilitate the clinical and research application of our findings, we developed two point-based scoring systems (LN-Code and LN-No Code) in which points were assigned to each predictor by multiplying each β coefficient (log OR) from the logistic regression model by a constant arbitrary number and rounding (to the nearest integer for points 1–5 and to the nearest fifth integer for points >5) to facilitate total risk score calculation. A total risk score was assigned to each patient by summing the points for each predictor in the scoring system. Mean predicted probabilities of LN corresponding to total risk scores were reported; this approach was intended to facilitate application of the scores depending on data availability and the desired predicted probability for the application.

Performance evaluation

For the test set and the external validation set, separately, we determined the sensitivity (recall), specificity, PPV, negative predictive value (NPV), and accuracy of ≥1 ICD9/10 code and ≥2 ICD9/10 codes ≥30 days apart, for identifying patients with a strict or an inclusive definition of LN. The above metrics and area under curve (AUC) were used to evaluate the predictive performance of multivariable logistic regression models used for the development of the scoring systems. In addition, we assessed calibration of the regression models by comparing the mean predicted LN probabilities with the mean observed probabilities within every decile of predicted risk in the test set and the external validation set. Predictive performance of the regression models was additionally evaluated across races and ethnicities for identifying patients with a strict definition of LN, using pooled data from the test set and the external validation set. This was performed to ensure that the algorithm performed in an equitable manner by race and ethnicity.

Table 1 Demographic characteristics of the study populations

Patient demographics		University of California San Francisco N=4152	Zuckerberg San Francisco General N=370
Age*, mean (SD)		50.4 (18.6)	49.5 (15.5)
Sex, N (%)	Female	3629 (87.4)	313 (84.6)
	Male	520 (12.5)	56 (15.1)
	Other or unknown	3 (0.1)	1 (0.3)
Race/ethnicity, N (%)	Asian	688 (16.6)	94 (25.4)
	Hispanic or Latino	802 (19.3)	122 (33.0)
	Non-Hispanic black	473 (11.4)	76 (20.5)
	Non-Hispanic white	1534 (37.0)	35 (9.5)
	Other or mixed	413 (10.0)	41 (11.1)
	Unknown	242 (5.8)	2 (0.5)

*Current age, at data extraction.

RESULTS

The random samples for manual chart review (forming the test set and external validation set) were selected from a total of 4152 patients from UCSF and 370 patients from ZSFG who met the eligibility criteria for inclusion in the study. Mean (SD) age of patients meeting the eligibility criteria was 50.4 (18.6) years and 49.5 (15.5) years, at UCSF and ZSFG, respectively. As expected, the majority of patients were female (87.4% at UCSF, 84.6% at ZSFG). The most prevalent race or ethnicity was non-Hispanic white (37.0%) at UCSF, and Hispanic or Latino (33.0%) at ZSFG (table 1).

Performance of LN ICD codes

A majority of chart-reviewed cases had definite LN and were identified with the ICD10 codes M32.14 or M32.15, or ICD9 code 710.0 in combination with ICD9 codes 583.81, 581.81 or 583.89. The total number of chart-reviewed cases meeting the strict and inclusive definitions of LN was 34 and 41 (out of 100) from UCSF (test set), and 40 and 46 (out of 100) from ZSFG (external validation set). In order to quantify their underutilisation, we evaluated the sensitivity of LN ICD codes. The sensitivity of ≥1 LN ICD9/10 code based on strict and inclusive definitions of LN was 56% and 51% in the test set, and 73% and 65% in the external validation set. The specificity, PPV and NPV of ≥1 LN ICD9/10 code based on a strict definition of LN was 94%, 83% and 81% in the

Table 2 Diagnostic performance of lupus nephritis (LN) International Classification of Diseases (ICD) codes

Test set	Diagnostic criteria	LN definition	Sensitivity, %	Specificity, %	PPV, %	NPV, %	Accuracy, %
UCSF	≥1 ICD code	Strict	56	94	83	81	81
		Inclusive	51	97	91	74	78
	≥2 ICD codes, ≥30 days apart	Strict	50	96	85	79	80
		Inclusive	43	95	85	71	74
ZSFG (external set)	≥1 ICD code	Strict	73	92	85	83	84
		Inclusive	65	93	88	76	80
	≥2 ICD codes, ≥30 days apart	Strict	55	95	88	76	79
		Inclusive	50	96	92	69	75

Performance of LN diagnosis codes for identifying prevalent cases of LN.
Inclusive, definite LN, potential LN or diagnostic uncertainty; NPV, negative predictive value; PPV, positive predictive value; Strict, definite LN only; UCSF, University of California San Francisco; ZSFG, Zuckerberg San Francisco General.

test set and 92%, 85% and 83% in the validation set. The diagnostic criterion of ≥1 ICD9/10 LN code used with the strict definition of LN yielded the highest accuracy in both health systems (table 2). In summary, across both health systems, LN codes were specific when present, but sensitivity was inadequate. This finding led to the focus of algorithm development, discussed below, to prioritise sensitivity.

Scoring systems

Scoring system 1, including LN diagnosis codes: 'LN-Code'

The top 20 important predictors were identified by GBM (online supplemental figure 1). In addition to more specific diagnosis codes for LN, presence of diagnosis codes for acute or chronic kidney disease or proteinuria, younger age at first SLE diagnosis code, and use of mycophenolate mofetil or mycophenolic acid were identified as key predictors and included in the final logistic regression model. UPCr >0.5 g/g, abnormal C3 levels, any use of hydroxychloroquine, azathioprine or rituximab, and glucocorticoid dose were also identified as important predictors but were omitted from the final logistic regression model as their inclusion did not further improve sensitivity. The final logistic regression model had an AUC, sensitivity and PPV of 0.93, 0.88 and 0.84, respectively, for identifying LN using the inclusive definition, in the test set. The model performed similarly (although a lower PPV of 0.70) with a strict definition of LN and had good external validity when tested in the second health system (online supplemental table 1 and online supplemental figures 2 and 3). Predicted and observed probabilities of LN had good calibration (online supplemental table 1). Model sensitivity ranged from 0.83 among Hispanics and Asians to 1.00 among white patients, with differences not reaching statistical significance (p>0.434; online supplemental table 2). The first scoring system was derived from this model (table 3).

Table 3 Scoring system 1: LN-Code

Patient characteristic All patients are required to have at least one diagnosis code (ICD9 or ICD10) for SLE	Points (*add unless stated*)	Total score	Predicted probability
One diagnosis code (ICD10) for LN	35	<80	<50%
Or Two or more diagnosis codes (ICD9 or ICD10) for LN	115	80–110	50–60%
One diagnosis code for acute or chronic kidney disease or proteinuria	70	110–140	60–70%
Or Two or more diagnosis codes for acute or chronic kidney disease or proteinuria	95	140–175	70–80%
Any use of mycophenolate mofetil or mycophenolic acid	45	175–230	80–90%
Age (in years) at first SLE diagnosis code	Subtract age	>230	>90%

Diagnosis codes for LN included ICD10 codes: M32.14 or M32.15, or ICD9 code 710.0 in combination with ICD9 codes: 583.81, 581.81 or 583.89. Diagnosis codes for acute or chronic kidney disease or proteinuria included ICD10 codes: N00–N08, N17–N19 and R80, or ICD9 codes: 580–586 and 791.0.
ICD, International Classification of Diseases; LN, lupus nephritis.

Table 4 Scoring system 2: LN-No Code

Patient characteristic All patients are required to have at least one diagnosis code (ICD9 or ICD10) for SLE	Points (*add unless stated*)	Total score	Predicted probability
Highest urine protein–creatinine (UPCr) ratio test greater than or equal to 1 but less than 3 g/g *Or* Highest UPCr ratio test greater than or equal to 3 g/g	25	<35	<50%
	30	35–55	50–60%
One diagnosis code for acute or chronic kidney disease or proteinuria *Or* Two or more diagnosis codes for acute or chronic kidney disease or proteinuria	50	55–70	60–70%
	80	70–90	70–80%
One or more urinalysis tests positive for protein	35	90–125	80–90%
Age (in years) at first SLE diagnosis code	Subtract age	>125	>90%

Diagnosis codes for LN included ICD10 codes: M32.14 or M32.15, or ICD9 code: 710.0 in combination with ICD9 codes: 583.81, 581.81 or 583.89. Diagnosis codes for acute or chronic kidney disease or proteinuria included ICD10 codes: N00–N08, N17–N19 and R80, or ICD9 codes: 580–586 and 791.0.
ICD, International Classification of Diseases; LN, lupus nephritis.

Scoring system 2, excluding LN diagnosis codes: 'LN-No Code'

A second scoring system was developed after omitting LN ICD codes as predictors for use in settings where LN diagnosis codes are unavailable or largely underused. GBM identified the top 20 important predictors (online supplemental figure 4). Diagnosis codes for acute or chronic kidney disease or proteinuria, UPCr ≥1 g/g, urinalysis positive for protein and younger age at first SLE diagnosis code were identified as key predictors for identifying patients with prevalent LN and included in the final logistic regression model. UPCr was categorised to <1 g/g, ≥1 and <3 g/g, and ≥3 g/g, to improve model fit. The final logistic regression model had an AUC, sensitivity and PPV of 0.91, 0.95 and 0.72, respectively, for identifying LN using the inclusive definition in the test set. Consistent with the first scoring system, PPV was lower, at 0.61, with a strict definition of LN. The model had good external validity when tested in ZSFG and performed comparably with the first scoring system (online supplemental table 3 and online supplemental figures 2 and 3). Predicted and observed probabilities of LN had good calibration (online supplemental table 3). Model sensitivity ranged from 0.86 among Asians to 1.00 among black and white patients, with differences not reaching statistical significance (p>0.194; online supplemental table 2). The second scoring system was derived from this model (table 4).

DISCUSSION

This study addresses the challenge of accurately identifying prevalent cases of LN using structured EHR data. We first characterised performance of LN ICD codes documented in EHRs of two large health systems and found high specificity (92–97%) but low sensitivity (43–73% across health systems). We then went on to develop two scoring systems using a broader range of EHR data, prioritising sensitivity. Given the morbidity and mortality associated with LN, accurate identification of this manifestation of SLE is vital for patient management, clinical research and public health initiatives, but has been hampered by underutilisation of specific LN diagnosis codes. Prediction of prevalent LN using structured data elements available in EHRs was feasible, had good accuracy and external validity. The scoring systems proposed have the potential to identify prevalent LN accurately across different health systems.

We found that the specificity of LN diagnosis codes was high, but their sensitivity was low. We were able to quantify the impact of underutilisation of LN ICD codes on the ability to accurately identify patients with prevalent LN. LN diagnosis codes are associated with a high false-negative rate as the code does not discern a diagnosis of 'no LN' from the lack of a documentation of LN diagnosis. The sensitivity of diagnosis codes for LN was suboptimal in both health systems examined, especially when using an inclusive definition of LN. This finding suggests that relying solely on these codes may result in missing a significant proportion of prevalent LN cases (up to 52% in this study) and underscores the importance of combining multiple data elements and refined algorithms for accurate LN identification.

We developed two scoring systems, each designed to enhance LN identification in different settings. LN-Code, which includes LN diagnosis codes, demonstrated strong predictive performance with an AUC of 0.93 for the inclusive definition of LN and good sensitivity (0.88). This system is valuable when specific LN diagnosis codes are routinely used and could be employed to improve the accuracy of prevalent LN identification in EHRs. In this scenario, patients with two or more LN codes and two or more codes for kidney disease or proteinuria have a high predicted probability of having LN; use of

mycophenolate acid and younger age further increase this probability. LN-No Code, excluding LN diagnosis codes, was designed to address situations where LN codes are underused or unavailable. This model achieved a remarkable sensitivity of 0.95 for the inclusive definition of LN (0.97 for the strict definition) while maintaining a high AUC of 0.91. While other clinical indicators, such as diagnoses of chronic kidney disease, proteinuria and higher UPCr, lack the specificity of LN ICD codes, their inclusion in this scoring system offers a practical solution for identifying LN in cases where specific LN codes are lacking.

We chose to develop a scoring system rather than a binary system (diagnosis present vs absent) to permit a more granular assessment of LN diagnosis. This approach allows investigators to apply definitions according to the use case and the degree of accuracy required for the intended task. Such an approach permits adjustment of the threshold for classification based on clinical or operational needs. The scoring systems outline which data elements are required for high accuracy and highlight where more limited data might compromise accuracy. Moreover, the scoring systems provide transparency, allowing users to directly observe how different variables are weighted, making the model more interpretable and clinically relevant.

Investigators may choose to apply these scoring systems separately or together based on their needs. For example, whether for research or clinical work, if high sensitivity is desired and at least some LN codes are available, investigators may choose to use LN-Code first to identify patients with LN; this may then be followed by the application of LN-No Code to identify any additional patients who do not have codes for LN to increase sensitivity. In another use case, an investigator may be interested in applying a system with very high specificity; in this case, using LN-Code alone may be preferred.

As evident by receiver operating characteristic curve plots, differences in performance between the two scoring systems were more subtle in the ZSFG health system (our external validation set) than at UCSF (the test set). We observed greater improvements in sensitivity by both scoring systems (compared with LN diagnosis codes alone) in UCSF than ZSFG. This may be explained by a higher sensitivity of LN diagnosis codes in ZSFG, which is an integrated public health system, leaving fewer unidentified true positives. Another reason may be differences in workflow within EHRs as well as documentation differences of laboratory results, diagnosis or procedure codes across the two health systems. In addition, patient demographics differed significantly across the two health systems.

We see many clinical applications of these scoring systems to identify individuals with LN. LN is a condition associated with significant morbidity, causing 2% of all end-stage renal disease in the USA. While there has been some progress in treating LN, with two new drug approvals in 2020–2021, there are few comparative effectiveness studies using real-world data. Tools for disease surveillance of LN are also lacking. Application of the scoring systems presented here could facilitate such studies. The scoring systems could also help identify eligible patients for clinical trials requiring identification of prevalent disease. Moreover, in clinical settings, individuals with LN may be lost to follow-up or not under appropriate specialty care. Using algorithms such as those presented here for population health management and quality improvement may be useful for ensuring all patients receive timely and appropriate care.

This study employed a rigorous methodology involving chart reviews conducted by rheumatologists, machine learning techniques to optimise predictive performance and evaluation of a comprehensive range of metrics to validate the scoring systems. We also acknowledge the limitations of our study. First, our research was conducted using data from two academic health systems in the San Francisco Bay Area, potentially limiting generalisability. Future studies should validate the scoring systems in community healthcare settings. Second, our scoring systems rely on structured EHR data and therefore may not fully capture the complexity of LN that is included in clinical notes. Incorporating unstructured clinical notes and imaging data could further improve accuracy, although many large datasets lack unstructured data, and even when available, health systems and clinics may lack the infrastructure or resources to analyse clinical notes.

In conclusion, our study corroborates previous studies demonstrating that LN diagnosis codes are underused in EHRs and presents two novel scoring systems to enhance LN identification. These systems, tailored to different data availability scenarios, offer practical solutions for healthcare providers and researchers. Improving the accuracy of LN identification has potential to facilitate patient care, inform research and facilitate public health initiatives in the field of LN. Further research and validation are warranted to ensure the robustness and applicability of our scoring systems across diverse healthcare settings.

Acknowledgements We acknowledge financial support from Aurinia.

Contributors ZI and MG directly accessed and verified the underlying data reported in the manuscript. ZI designed the study and data analysis, performed the literature search and the statistical analyses, developed the manuscript, figures and tables, and takes responsibility for the data and analyses. ZI and MG contributed to data quality control. All authors contributed to the interpretation of data and writing of the manuscript. JY supervised the work. All authors approved the final version to be submitted. ZI acts as the guarantor.

Funding This work was additionally supported by NIH/NIAMS K24 AR074534, a mentoring grant, and NIH/NIAMS P30 AR070155, for data extraction.

Disclaimer Aurinia was not involved in the design and conduct of the study; collection, management, analysis and interpretation of the data; preparation, review or approval of the manuscript; and decision to submit the manuscript for publication.

Competing interests ZI is currently employed by BMS. MG reports grants from the National Institutes of Health and NIAMS, outside the submitted work. She is currently employed by Pfizer. CA has received funding from Rheumatology Research Foundation Scientist Development Award. GS reports no conflicts of interest. JY's work is supported by grants from the National Institutes of Health (K24 AR074534 and P30 AR070155). Outside of this work, she has received

research grants or performed consulting for Gilead, BMS Foundation, Pfizer, Aurinia and AstraZeneca.

Patient and public involvement Patients and/or the public were not involved in the design, or conduct, or reporting, or dissemination plans of this research.

Patient consent for publication Not applicable.

Ethics approval The study was reviewed by and obtained ethical approval from the University of California San Francisco (IRB: 15-17561).

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement No data are available. Data sharing is not available for this study.

ORCID iDs
Zara Izadi http://orcid.org/0000-0002-1867-0905
Jinoos Yazdany http://orcid.org/0000-0002-3508-4094

REFERENCES

1 Alarcón GS. Multiethnic lupus cohorts: what have they taught us *Reumatol Clin* 2011;7:3–6.
2 Bastian HM, Roseman JM, McGwin G Jr, *et al*. Systemic lupus erythematosus in three ethnic groups. XII. risk factors for lupus nephritis after diagnosis. *Lupus* 2002;11:152–60.
3 Cervera R, Khamashta MA, Font J, *et al*. Morbidity and mortality in systemic lupus erythematosus during a 5-year period. A multicenter prospective study of 1,000 patients. *Medicine (Baltimore)* 1999;78:167–75.
4 Hocaoğlu M, Valenzuela-Almada MO, Dabit JY, *et al*. Incidence, prevalence, and mortality of lupus nephritis: a population-based study over four decades using the lupus Midwest network. *Arthritis Rheumatol* 2023;75:567–73.
5 Simard JF, Oates J, Yazdany J, *et al*. SLE and UCTD in the rheumatology Informatics system for effectiveness (RISE) Registry. *Arthritis and Rheumatology* 2016;68:2288–90.
6 Chibnik LB, Massarotti EM, Costenbader KH. Identification and validation of lupus nephritis cases using administrative data. *Lupus* 2010;19:741–3.
7 Yeh WSWSMcCartyKM. Health resource utilization of lupus nephritis patients – a comparison of result across case identification Algorithms. *Value in Health* 2013;16:A21.
8 Wenderfer SE, Chang JC, Goodwin Davies A, *et al*. Using a multi-institutional pediatric learning health system to identify systemic lupus erythematosus and lupus nephritis: development and validation of Computable phenotypes. *Clin J Am Soc Nephrol* 2022;17:65–74.
9 Hahn BH, McMahon MA, Wilkinson A, *et al*. American college of rheumatology guidelines for screening, treatment, and management of lupus nephritis. *Arthritis Care Res (Hoboken)* 2012;64:797–808.
10 Murray SG, Avati A, Schmajuk G, *et al*. Automated and flexible identification of complex disease: building a model for systemic lupus erythematosus using noisy labeling. *J Am Med Inform Assoc* 2019;26:61–5.
11 Moons KGM, Altman DG, Reitsma JB, *et al*. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): explanation and elaboration. *Ann Intern Med* 2015;162:W1–73.
12 Weening JJ, D'Agati VD, Schwartz MM, *et al*. The classification of glomerulonephritis in systemic lupus erythematosus Revisited. *J Am Soc Nephrol* 2004;15:241–50.
13 Bajema IM, Wilhelmus S, Alpers CE, *et al*. Revision of the International society of Nephrology/renal pathology society classification for lupus nephritis: clarification of definitions, and modified National Institutes of health activity and Chronicity indices. *Kidney Int* 2018;93:789–96.
14 Schaefer J, Lehne M, Schepers J, *et al*. The use of machine learning in rare diseases: a Scoping review. *Orphanet J Rare Dis* 2020;15:145.
15 Hong WS, Haimovich AD, Taylor RA. Predicting hospital admission at emergency Department triage using machine learning. *PLoS One* 2018;13:e0201016.
16 Qiao Z, Sun N, Li X, *et al*. Using machine learning approaches for emergency room visit prediction based on electronic health record data. *Stud Health Technol Inform* 2018;247:111–5.
17 Kuhn M. Building predictive models in R using the **Caret** package. *J Stat Soft* 2008;28.
18 Richter AN, Khoshgoftaar TM. A review of statistical and machine learning methods for modeling cancer risk using structured clinical data. *Artificial Intelligence in Medicine* 2018;90:1–14.
19 Philips Z, Ginnelly L, Sculpher M, *et al*. Review of guidelines for good practice in decision-analytic Modelling in health technology assessment. *Health Technol Assess* 2004;8:iii–iv.
20 Shipe ME, Deppen SA, Farjah F, *et al*. Developing prediction models for clinical use using logistic regression: an overview. *J Thorac Dis* 2019;11:S574–84.