Arthritis Care & Research
Vol. 69, No. 5, May 2017, pp 687–693
DOI 10.1002/acr.22989

ORIGINAL ARTICLE

Developing Electronic Health Record Algorithms That Accurately Identify Patients With Systemic Lupus Erythematosus

APRIL BARNADO, CAROLYN CASEY, ROBERT J. CARROLL, LEE WHELESS, JOSHUA C. DENNY, AND LESLIE J. CROFFORD

Objective. **To study systemic lupus erythematosus (SLE) in the electronic health record (EHR), we must accurately identify patients with SLE. Our objective was to develop and validate novel EHR algorithms that use International Classification of Diseases, Ninth Revision (ICD-9), Clinical Modification codes, laboratory testing, and medications to identify SLE patients.**

Methods. **We used Vanderbilt's Synthetic Derivative, a de-identified version of the EHR, with 2.5 million subjects. We selected all individuals with at least 1 SLE ICD-9 code (710.0), yielding 5,959 individuals. To create a training set, 200 subjects were randomly selected for chart review. A subject was defined as a case if diagnosed with SLE by a rheumatologist, nephrologist, or dermatologist. Positive predictive values (PPVs) and sensitivity were calculated for combinations of code counts of the SLE ICD-9 code, a positive antinuclear antibody (ANA), ever use of medications, and a keyword of "lupus" in the problem list. The algorithms with the highest PPV were each internally validated using a random set of 100 individuals from the remaining 5,759 subjects.**

Results. **The algorithm with the highest PPV at 95% in the training set and 91% in the validation set was 3 or more counts of the SLE ICD-9 code, ANA positive (≥1:40), and ever use of both disease-modifying antirheumatic drugs and steroids, while excluding individuals with systemic sclerosis and dermatomyositis ICD-9 codes.**

Conclusion. **We developed and validated the first EHR algorithm that incorporates laboratory values and medications with the SLE ICD-9 code to identify patients with SLE accurately.**

INTRODUCTION

Electronic health records (EHRs) are an increasingly important tool in clinical research and are near ubiquitous in the US due to meaningful use standards (1). EHRs provide longitudinal information on a patient's disease course that can be linked to genetic data for discovery research (2). For less common diseases such as systemic lupus erythematosus (SLE), using EHRs can be an efficient and cost-effective tool to study many patients from diverse settings (3). The first step of any EHR-based study is to identify a cohort with the target condition accurately. Identifying patients with SLE is challenging given the heterogeneity of the disease phenotype and the frequency of false positive diagnoses, in part because of the high prevalence of false positive antinuclear antibody (ANA) tests.

Many epidemiologic studies have used the International Classification of Diseases, Ninth Revision (ICD-9), Clinical Modification billing code data, specifically 2 or 3 counts of the SLE ICD-9 code 710.0, to identify patients with SLE within administrative databases (4–9). A recent systematic review highlights that this method has not been rigorously validated and performs poorly, with positive predictive values (PPVs) of 50–60% in general populations (10). Liao et al (11) developed an algorithm for rheumatoid arthritis (RA) that used not only ICD-9 codes but also laboratory values, medication data, and natural language processing, with a PPV of 94% and a sensitivity of 63%. This algorithm was internally and externally validated by our group (11,12). We also developed similar algorithms for atrial fibrillation, Crohn's disease, multiple sclerosis, and type 2 diabetes mellitus (3,13) and have used the EHR for genome- and phenome-wide studies (14–16). In this study, we developed

Dr. Barnado's work was supported by the NIH/National Institute of Arthritis and Musculoskeletal and Skin Diseases (5T32-AR-059039-05), the National Institute of Child Health and Human Development (5K12-HD-043483-12), the National Center for Research Resources (UL1-RR-024975), and the National Center for Advancing Translational Science (UL-TR-000445).

April Barnado, MD, Carolyn Casey, MD, Robert J. Carroll, PhD, Lee Wheless, MD, PhD, Joshua C. Denny, MD, MS, Leslie J. Crofford, MD: Vanderbilt University Medical Center, Nashville, Tennessee.

Address correspondence to April Barnado, MD, 1161 21st Avenue South, T3113 MCN, Nashville, TN 37232. E-mail: april.barnado@vanderbilt.edu.

Submitted for publication February 10, 2016; accepted in revised form July 5, 2016.

Significance & Innovations

- We developed and validated the first electronic health record (EHR) algorithms that incorporate laboratory and medication values with International Classification of Diseases, Ninth Revision codes to identify patients with systemic lupus erythematosus (SLE) accurately within the EHR.
- We present 3 EHR algorithms with positive predictive values greater than 90% that are widely applicable to EHRs.
- EHR algorithms would allow translational and clinical researchers to identify and study large populations of patients with SLE for discovery research.

and validated novel algorithms to identify patients with SLE accurately in the EHR that leverages laboratory data, medications, keywords, and ICD-9 codes.

PATIENTS AND METHODS

Patient selection. An overview of our approach is illustrated in Figure 1. We used data from a de-identified version of Vanderbilt's EHR called the Synthetic Derivative (17), following approval from the Institutional Review Board of Vanderbilt University Medical Center. Vanderbilt is a regional tertiary care center. The Synthetic Derivative contains over 2.5 million subjects with de-identified clinical data from the EHR, collected longitudinally over several decades with approximately equal males and females who are predominantly white. The Synthetic Derivative includes all information available in the EHR, incorporating diagnostic and procedure codes (ICD-9 and Current Procedural Terminology), demographics, text from inpatient and outpatient notes (including both subspecialty and primary care), laboratory values, radiology reports, and medication orders. Outside records scanned into the EHR, however, are not available in the Synthetic Derivative. Medical orders derive from electronic prescribing systems and natural language processing from phone call logs and notes. Users can perform text-based searches of the entire clinical record within seconds to increase the efficiency and accuracy of data extraction. Records from the Synthetic Derivative are linked to a DNA biorepository called BioVU (17).

Within the Synthetic Derivative, we identified potential SLE cases with at least 1 count of the SLE ICD-9 code (710.0). Of these potential cases, we randomly selected 200 for chart review to identify their true disease status and to serve as a training set. Chart review on the 200 potential SLE cases was conducted by a rheumatologist (AB), with a random 50 of the 200 potential SLE cases reviewed by another rheumatologist (CC) to assess agreement on the final case determination. The second rheumatologist (CC) was blinded to the case status given by the first rheumatologist (AB). A subject was defined as a case if diagnosed with SLE by a Vanderbilt or external rheumatologist, dermatologist, or nephrologist, who was mentioned specifically in the note. Subjects with cutaneous lupus or drug-induced lupus were not considered cases. Potential subjects were classified as cases, not cases with alternative diagnoses noted, unconfirmed if hesitancy in the SLE diagnosis, or missing if there was unavailable clinical documentation.

Algorithm development. A priori, the authors decided to use as potential algorithm components the number of counts of the SLE ICD-9 code (710.0), keyword of "lupus" in the problem list, positive ANA, and ever use of medications that are frequently used in SLE, such as antimalarials, systemic corticosteroids, and disease-modifying antirheumatic drugs (DMARDs) (see list below). These components were selected based on SLE disease criteria and management combined with data accessible in the EHR. Subjects who were not classified as true SLE cases frequently had other autoimmune diseases with ICD-9 codes for these diseases. Multiple true SLE cases had an overlap syndrome (e.g., secondary Sjögren's syndrome or RA). We examined the exclusion of individuals with ICD-9 codes for other autoimmune diseases to assess whether potential SLE subjects who were not true SLE cases were appropriately excluded and also to ensure true SLE cases with overlap

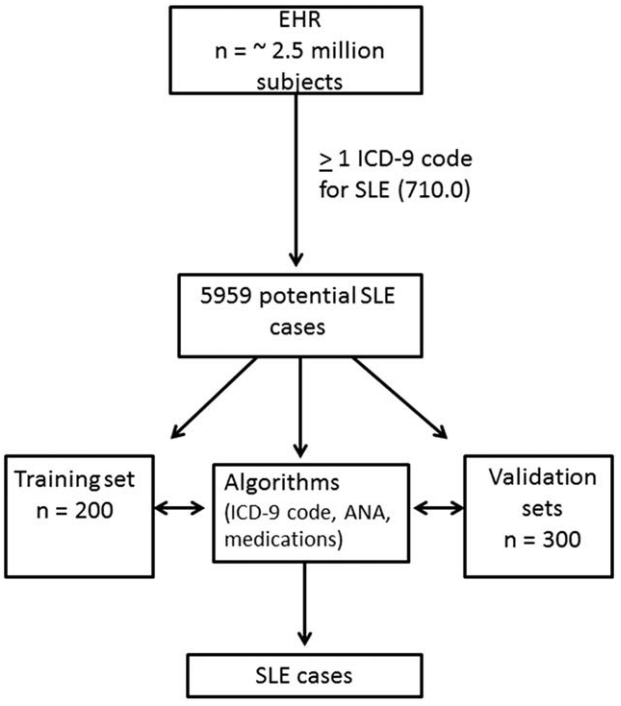

Figure 1. Development of the electronic health record (EHR) to identify patients with systemic lupus erythematosus (SLE). At least a 1-time count of the SLE International Classification of Diseases, Ninth Revision (ICD-9) code (710.0) was applied to the 2.5 million subjects in Vanderbilt's Synthetic Derivative, which resulted in 5,959 potential SLE cases. Of these 5,959, 200 were randomly selected as a training set to develop and test algorithms with various combinations of the SLE ICD-9 code, keywords, positive antinuclear antibody (ANA), and ever medication use. Three of the high-performing algorithms were each internally validated in randomly selected 100 subjects from the remaining 5,759 potential SLE cases. The internally validated algorithms were then used to identify SLE cases.

syndromes were not excluded. These exclusion criteria were selected during chart review of the training set.

PPVs and sensitivity were calculated for every combination of ≥1, 2, 3, and 4 counts of the SLE ICD-9 code; a positive ANA (titer ≥1:40 and titer ≥1:160); ever use of antimalarials, systemic corticosteroids, and DMARDs; and a keyword of "lupus" in the problem list using "and" or "or" between the criteria. PPVs were also calculated, excluding ICD-9 codes for systemic sclerosis (SSc) (710.1) and dermatomyositis (DM) (710.3). All algorithms included individuals with at least 1 count of the SLE ICD-9 code. The PPV was calculated as the number of subjects who fit the algorithm and were confirmed cases on chart review divided by the total number of subjects who fit the algorithm. Sensitivity was calculated as the number of subjects who fit the algorithm and were confirmed cases on chart review divided by total number of confirmed cases. To fit the algorithm, the subject had to have available data for that particular algorithm's criteria. If an ANA was not checked at Vanderbilt, it was considered missing. The F-score, which is the harmonic mean of the PPV and sensitivity ([2 × PPV × sensitivity]/[PPV + sensitivity]), was calculated for all algorithms. The F-score is commonly used in informatics because it provides a single number to compare algorithms accounting for both PPV and sensitivity.

Antimalarials included in the medication search were hydroxychloroquine, plaquenil, chloroquine, quinacrine, and aralen. Oral and intravenous corticosteroids included were cortisone acetate, hydrocortisone, Cortef, prednisone, dexamethasone, dexamethasone Intensol, decadron, prednisolone sodium phosphate, Pediapred, prednisone Intensol, methylprednisolone, Medrol, Medrol Dosepak, prednisolone, Orapred, and Prelone. DMARD search terms included were azathioprine, Imuran, methotrexate sodium, methotrexate, Trexall, mycophenolate mofetil, CellCept, mycophenolic acid, Myfortic, cyclophosphamide, Cytoxan, rituximab, Rituxan, etanercept, Enbrel, Enbrel Sureclick, adalimumab, Humira, Humira Pen, infliximab, Remicade, abatacept, and Orencia.

Alternative search strategies. By requiring at least 1 count of the SLE ICD-9 code, SLE patients could possibly have been missed. To test this hypothesis, we searched for potential SLE subjects without an SLE ICD-9 code but who had a keyword of "systemic lupus erythematosus," "systemic lupus," or "lupus" in the subjects' problem lists, to approximate a negative predictive value (NPV) for algorithms using an SLE ICD-9 code. We then randomly selected 50 of these potential SLE subjects for chart review to determine case status. In our training set, we investigated the frequency of the 695.4 "lupus erythematosus" ICD-9 code under "diseases of the skin and subcutaneous tissue" and calculated the PPV and sensitivity of adding 695.4 to 710.0.

Algorithm validation. The 3 algorithms with the highest PPVs and highest combined PPV and sensitivity were validated in 3 distinct sets of 100 randomly selected potential SLE cases not reviewed previously. Chart review was conducted by a rheumatologist (AB) with the same case definition defined above.

Statistical analysis. We assessed differences between subjects who met the SLE case definition versus those who did not using the Mann-Whitney U test for continuous variables, as there were non-normal distributions in the data, and chi-square or Fisher's exact test for categorical variables. Our null hypothesis was that there is no difference in the PPVs of EHR algorithms that incorporate medication and laboratory values with an SLE ICD-9 code compared with algorithms that only use an SLE ICD-9 code. Our preliminary data showed that the algorithm using only 2 or more code counts of the SLE ICD-9 code had a PPV of 65%. Using 90 SLE cases and 95 subjects who were not SLE cases in the training set, we calculated there would be 98% power to detect a PPV of 90% for an algorithm incorporating medication and laboratory values with the SLE ICD-9 code, with an alpha of 0.05 using Fisher's exact test (18). Two-sided *P* values of less than 0.05 were considered to indicate statistical significance. Analyses were conducted using SPSS software, version 23.0. Random numbers to select the training and validation sets were generated in the R statistical package (19) with a set seed of 1. The PS program, version 3.1.2, was used to compute the sample size (18).

RESULTS

An overview of our approach and the counts of individuals is illustrated in Figure 1. In the Synthetic Derivative, we identified 5,959 potential SLE cases with at least 1 SLE ICD-9 code (710.0). Of the randomly selected 200 of the 5,959 potential SLE cases, 90 subjects (45%) were defined as SLE cases by chart review. Of the remaining 110 subjects, 76 were classified as not being SLE cases, 19 as having an unconfirmed diagnosis of SLE, and 15 had missing clinic notes. Of the 76 subjects not classified as SLE, many had alternative autoimmune diagnoses, including RA (n = 14), SSc (n = 7), cutaneous lupus (n = 4), Sjögren's syndrome (n = 4), DM (n = 4), inflammatory arthritis (n = 3), undifferentiated connective tissue disease (n = 2), and drug-induced lupus (n = 2). Many of the 19 unconfirmed subjects had a positive ANA or a self-reported history of SLE with no diagnosis made by a rheumatologist, nephrologist, or dermatologist. Further, rheumatologists in the medical record explicitly questioned the diagnosis of SLE. These 19 unconfirmed subjects were analyzed with the 76 subjects not classified as SLE, resulting in a total of 95 subjects who were not SLE cases. The 15 subjects with missing notes were excluded from the analysis, as case status could not be determined, and the various algorithms that required laboratory and medication data could not be applied.

Of the 90 SLE cases, 7 had a secondary or overlap autoimmune disease in addition to SLE; 4 subjects had both SLE and Sjögren's syndrome, 1 SLE and DM, 1 SLE and RA, and 1 SLE and Behçet's syndrome. All 7 of these subjects who were classified as SLE had ICD-9 codes for both SLE and the other autoimmune disease.

A second rheumatologist (CC) reviewed a randomly selected 50 of the 200 charts in the training set. Of the 50 charts, the second rheumatologist's determination of case status was the same as the original rheumatologist (AB) with 96% agreement. For the 2 charts with initial disagreement, a

Table 1. Characteristics of SLE cases versus non-SLE cases in the training set*

Characteristics	SLE cases (n = 90)	Non-SLE cases (n = 95)†	*P*‡
Age, years	53 ± 15	61 ± 15	< 0.001
Female, %	91	87	0.41
White, %	68	82	0.04
No. of counts of SLE ICD-9 code (710.0)	17 ± 19	4 ± 7	< 0.001
Years of followup	8 ± 6	8 ± 5	0.72
Specialty visits, %§	91	73	0.001
ICD-9 code used by a specialist, %§	91	37	< 0.001

* Values are the mean ± SD unless indicated otherwise. SLE = systemic lupus erythematosus; ICD-9 = International Classification of Diseases, Ninth Revision.
† Non-SLE cases include subjects who were not classified as having SLE (n = 76) and not having a confirmed diagnosis of SLE (n = 19).
‡ Mann-Whitney U test for continuous variables and chi-square test for categorical variables. Of the 200 subjects in the training set, 15 subjects were missing sufficient clinical information.
§ Specialty visits included rheumatology, dermatology, and nephrology.

final consensus was reached, with both subjects determined to have mixed connective tissue disease.

The 90 SLE cases and the 95 subjects who were not SLE cases are compared in Table 1. Both SLE cases and subjects who were not cases were predominantly women (91% versus 87%; $P = 0.41$). Compared to subjects who were not cases, the SLE cases were significantly younger (age 53 ± 15 versus 61 ± 15 years; $P = 0.001$) and less likely to be white (68% versus 82%; $P = 0.04$). SLE cases and subjects who were not SLE cases had similar years of EHR followup (8 ± 6 versus 8 ± 5 years; $P = 0.72$), but SLE cases had significantly more occurrences of the SLE ICD-9 code than subjects who were not cases (17 ± 19 versus 4 ± 7; $P < 0.001$). Compared to the subjects who were not SLE cases, a higher proportion of SLE cases had visits (91% versus 73%; $P = 0.001$) and an SLE code (91% versus 37%; $P < 0.001$) used by a rheumatologist, dermatologist, or nephrologist. Excluding the 19 unconfirmed subjects from the non-SLE group did not significantly change the above results.

Of the 90 SLE cases, only 8 did not see a Vanderbilt rheumatologist, dermatologist, or nephrologist. Of these, 5 followed with an external rheumatologist, who was mentioned in the note, and 3 had renal pathology consistent with SLE nephritis in the EHR. Only 66 had documentation of the American College of Rheumatology (ACR) SLE criteria, while 26% who saw a rheumatologist did not have documented ACR SLE criteria (20).

Using 185 subjects with sufficient data in the training set, PPVs and sensitivity were calculated for the counts of the SLE ICD-9 code (Table 2). As the frequency of the code counts increased, the PPV increased. Excluding ICD-9 codes for SSc and DM further increased the PPVs for all the algorithms by 2–5%. Algorithms that used a keyword of "lupus" in the subject's problem list had similar PPVs but lower sensitivities compared with algorithms that used the ICD-9 codes. Adding a positive ANA, defined as either ≥1:40 or ≥1:160, improved the PPV of algorithms compared with algorithms that only used the ICD-9 code. Adding ever systemic corticosteroid use, ever DMARD use, or ever

Table 2. Positive predictive values (PPVs) of algorithms with SLE ICD-9 code counts, keyword, laboratory values, and medications*

No. of counts of SLE ICD-9 code 710.0	ICD-9 code alone	"Lupus" keyword in the problem list†	ICD-9 code plus ANA positive		ICD-9 code plus ever drug use		
			≥1:40	≥1:160	Antimalarial	DMARD	Corticosteroid
≥1							
PPV	49	57	53	51	63	52	50
PPV excluded‡	54	58	56	55	65	59	54
Sensitivity	NA	72	89	100	76	41	77
≥2							
PPV	65	69	71	71	77	73	70
PPV excluded‡	68	72	74	75	79	79	73
Sensitivity	86	68	83	97	72	37	70
≥3							
PPV	75	78	80	80	88	84	78
PPV excluded‡	78	82	84	84	91	89	81
Sensitivity	77	64	77	89	66	34	66
≥4							
PPV	79	80	83	84	89	83	80
PPV excluded‡	82	82	87	89	92	88	83
Sensitivity	71	61	72	86	61	33	61

* Values are percentages. SLE = systemic lupus erythematosus; ICD-9 = International Classification of Diseases, Ninth Revision; ANA = antinuclear antibody; DMARD = disease-modifying antirheumatic drug; NA = not applicable.
† Algorithm also included ≥1 count of the SLE ICD-9 code.
‡ Excluding dermatomyositis ICD-9 code (710.3) and systemic sclerosis ICD-9 code (710.1).

Table 3. Electronic health record algorithms with the highest positive predictive values*

Algorithm, SLE ICD-9 code counts†	PPV	F-score	PPV excluding DM, SSc‡	F-score	Sensitivity
≥3 plus ANA+ (≥1:40) and ever DMARD use and ever steroid use	91§	0.56§	95§	0.56§	40§
≥4 plus ANA+ and ever DMARD use and ever steroid use	90	0.53	95	0.54	38
≥4 plus ANA+ and ever antimalarial use	89	0.78	92	0.80	70
≥4 plus ever antimalarial use	89	0.72	92	0.73	61
≥3 plus ever antimalarial use	88§	0.75§	91§	0.77§	66§
≥3 plus ever steroid use and ever DMARD use	86	0.49	91	0.50	34
≥4 plus ever steroid use and ever DMARD use	86	0.48	91	0.48	33
≥4 plus ANA ≥1:160	86§	0.86§	89§	0.87§	86§

* Values are percentages, unless indicated otherwise. SLE = systemic lupus erythematosus; ICD-9 = International Classification of Diseases, Ninth Revision; PPV = positive predictive value; DM = dermatomyositis; SSc = systemic sclerosis; ANA = antinuclear antibody; DMARD = disease-modifying antirheumatic drug.
† SLE ICD-9 code 710.0.
‡ DM ICD-9 code (710.3), SSc ICD-9 code (710.1).
§ Internally validated.

antimalarial use to the ICD-9 code improved the PPVs. More combinations of the above criteria are provided in Supplementary Table 1 (available on the *Arthritis Care & Research* web site at http://onlinelibrary.wiley.com/doi/10.1002/acr.22989/abstract).

We performed an alternative search strategy of looking for potential SLE subjects without an SLE ICD-9 code but who had a keyword of "systemic lupus erythematosus," "systemic lupus," or "lupus" in the subjects' problem lists. Only 1 of the 50 randomly chosen subjects fit the SLE case definition on chart review, resulting in an estimated NPV of 98% for algorithms using the SLE ICD-9 code. The keyword search strategy found subjects with cutaneous or discoid lupus, other autoimmune diseases, a positive lupus anticoagulant without concomitant SLE, a family history of SLE, or self-reported diagnoses not confirmed by a rheumatologist.

Using the 695.4 "lupus erythematosus" ICD-9 code in the training set, we found 25 subjects with at least 1 count of 695.4 and 710.0. Of these 25, 13 were classified as SLE cases and 12 as having cutaneous but not systemic lupus. Combining 1 count of 695.4 to 710.0 resulted in a PPV of 14% and a sensitivity of 52%.

The algorithms with the highest PPVs are shown in Table 3. The algorithm with the highest PPV at 95% was 3 or more counts of the SLE ICD-9 code and ANA positive (≥1:40) and ever DMARD use and ever corticosteroid use, while excluding SSc and DM codes. The other algorithm with the highest PPV at 95% was 4 or more counts of the SLE ICD-9 code and ANA positive (≥1:40) and ever DMARD use and ever corticosteroid use, while excluding SSc and DM codes. The algorithm with the highest F-score of 87% was 4 or more counts of the SLE ICD-9 code and ANA positive (≥1:160), while excluding SSc and DM codes, with a PPV of 89% and sensitivity of 86%.

Three high-performing algorithms were selected, and each internally validated, on 100 randomly selected subjects who were not part of the training set (Table 4). We internally validated the algorithm of 3 or more counts of the SLE ICD-9 code and ANA positive (≥1:40) and ever DMARD use and ever corticosteroid use and excluding DM and SSc codes, with a PPV of 91%. We internally validated the algorithm with the highest F-score of 4 or more counts of the ICD-9 code and ANA positive (≥1:160) and excluding DM and SSc codes, with a PPV of 94%. All 31 cases that fulfilled this algorithm in the training set had an SLE ICD-9 code used by a Vanderbilt rheumatologist, dermatologist, or nephrologist. For the third algorithm, we selected an algorithm with the highest PPV among algorithms that did not incorporate an ANA value. We internally validated this algorithm of 3 or more counts of the ICD-9 code and ever antimalarial use while excluding DM and SSc codes and found a PPV of 88%. We conducted sensitivity analyses to determine the impact of missing clinical notes on the internally validated PPVs. Using best and worst case scenarios of counting the missing subjects as either SLE cases or not cases, missing notes impacted the PPV by ≤4% in either direction.

The algorithm with the highest F-score (4 or more counts of the SLE ICD-9 code and ANA ≥1:160 and excluding DM and SSc codes) was applied to the entire Synthetic Derivative, resulting in 1,098 cases. The SLE cases were of current

Table 4. Internal validation of the high-performing electronic health record algorithms*

Algorithm, SLE ICD-9 code counts†	PPV excluding DM, SSc in training set‡	PPV excluding DM, SSc in validation set‡
≥3 plus ANA+ (≥1:40) and ever DMARD use and ever steroid use	95	91
≥4 plus ANA+ (≥1:160)	89	94
≥3 plus ever antimalarial use	91	88

* Values are percentages. SLE = systemic lupus erythematosus; ICD-9 = International Classification of Diseases, Ninth Revision; PPV = positive predictive value; DM = dermatomyositis; SSc = systemic sclerosis; ANA = antinuclear antibody; DMARD = disease-modifying antirheumatic drug.
† SLE ICD-9 code 710.0.
‡ DM ICD-9 code (710.3), SSc ICD-9 code (710.1).

Table 5. Characteristics of 1,098 SLE cases with ≥4 SLE ICD-9 code counts and ANA positive*

Characteristics	Values
Age, years	51 ± 17
Age at first use of SLE ICD-9 code, years†	40 ± 17
Female, no. (%)	986 (90)
Race, no. (%)	
White	715 (65)
African American	270 (25)
Hispanic	30 (3)
Asian	25 (2)
Alaskan/Indian	2 (0.2)
Missing/unknown	56 (5)
Number of counts of SLE ICD-9 code†	20 ± 22
Years of followup	9 ± 5

* Values are the mean ± SD unless indicated otherwise. ANA positive ≥1:160. SLE = systemic lupus erythematosus; ICD-9 = International Classification of Diseases, Ninth Revision; ANA = antinuclear antibody.
† SLE ICD-9 code 710.0.

mean ± SD age of 50 ± 17 years and mean ± SD age at first use of the SLE ICD-9 code of 40 ± 17 years (Table 5). The SLE cases were predominantly women and white, with a minimum count of the SLE ICD-9 code at 4 and a maximum at 182, with a mean ± SD of 20 ± 22. The mean ± SD years of followup in the Synthetic Derivative was 9 ± 5 years, with a range of 1 to 24 years. A random 100 subjects were selected from these 1,098, and ACR SLE criteria documented on 85. The mean ± SD ACR SLE criteria reported in the clinical notes was 4.0 ± 1.6. As the ACR SLE criteria are not systematically documented, the number of criteria is likely underestimated. The ACR SLE criteria obtained from clinical notes were 72% with immune, 39% arthritis, 35% malar rash, 33% hematologic, 29% serositis, 28% renal, 25% oral or nasal ulcers, 18% photosensitivity, 13% neurologic, and 9% discoid. For autoantibodies, 57% had a positive double-stranded DNA, 25% positive RNP, 11% positive Smith, 37% positive SSA, and 13% positive SSB.

DISCUSSION

We developed and validated 3 novel algorithms to identify patients with SLE using multiple classes of data available in the EHR. This development is important, because it is the first instance of validated algorithms to incorporate laboratory and medication values with the SLE ICD-9 code. These algorithms exhibited PPVs of 95%, 89%, and 91% in a training set, and PPVs of 91%, 94%, and 88% in a validation set. Since one algorithm incorporates the ANA and medications with the SLE ICD-9 code, another uses only the ANA and the SLE ICD-9 code, and the third uses medications and the SLE ICD-9 code, investigators can select which algorithm is best suited to their EHR or administrative database.

A recent systematic review of algorithms to identify patients with SLE highlights that many studies do not describe algorithm validation (10). Of the 12 studies that performed some form of validation, PPVs for algorithms using the SLE ICD-9 code (710.0) ranged from 50–60% in the general population to 70–90% in selected populations, such as patients seen in a rheumatology clinic (10). Our PPVs of 49% and 65%, for 1 or 2 code counts of the SLE ICD-9 code, respectively, agree with this review. The authors of the review suggested that adding laboratory and medication values to the SLE ICD-9 code would likely result in a stronger algorithm (10).

We confirmed that incorporating pertinent medications and a positive ANA with the SLE ICD-9 code increased the PPVs of EHR algorithms. As expected, adding a positive ANA, even at a low titer of ≥1:40, improved the PPVs. Adding ever use of commonly prescribed medications in SLE, particularly the DMARD class, also improved the PPVs but at the expense of decreasing the sensitivity. This decrease in sensitivity likely reflects the variable clinical courses of patients with SLE, with some not requiring a DMARD.

While the SLE ICD-9 codes are arguably the most available data, medications and ANA are also available in many EHRs. Relying solely on 1 or even 2 counts of the SLE ICD-9 code is not accurate in identifying patients with SLE, with PPVs of 49% and 65%, respectively. The low PPVs of the ICD-9 codes alone may reflect the fact that physicians use ICD-9 codes differently, with their own biases of diagnosis and treatment and their own institution's practices (21). A physician may use the SLE ICD-9 code to justify further laboratory testing on a patient suspected of having SLE but who may ultimately not have this diagnosis. For example, 51% of subjects not classified as having SLE in the training set had an SLE ICD-9 code but did not have an SLE diagnosis documented. In addition, patients may report a history of SLE that has not been confirmed by a rheumatologist, and other providers may then use the SLE ICD-9 code. We expect that the ICD-10 codes will perform similarly to the ICD-9 code due to the same biases discussed above.

We have limitations in our study. To start our search for patients with SLE within Vanderbilt's Synthetic Derivative of over 2.5 million subjects, we applied a 1-time use of the SLE ICD-9 code to identify potential patients with SLE on which to develop and validate our algorithms. An SLE patient could possibly have been missed who did not have at least 1 encounter with an SLE ICD-9 code. In our alternative search strategy starting with a keyword for SLE instead of the SLE ICD-9 code, we estimated an NPV of 98% for algorithms using an SLE ICD-9 code. Thus, using an SLE ICD-9 code to start our search did not exclude significant numbers of patients with SLE. Given the low prevalence of SLE in the general population, we anticipate that the NPV would be unlikely to be lower than this estimate.

We defined an SLE case based on an SLE diagnosis given by a rheumatologist, nephrologist, or dermatologist, as not all these specialists document the ACR SLE criteria in clinic notes, which could cause exclusion of true SLE cases, based on physicians' documentation. Specifically, 26% of the SLE cases in the training set who saw Vanderbilt rheumatology did not have documented ACR SLE criteria. Tumor necrosis factor (TNF) inhibitors were included in the DMARD criterion to capture potential SLE patients who may have an overlap with RA and have been exposed to TNF inhibitors

soon after they were approved, but this criterion could capture drug-induced SLE cases. In our training set, however, this criterion did not capture any patients with drug-induced SLE.

A prior study of an EHR algorithm for RA used natural language processing for narrative EHR data, and a regression model to develop their algorithm (11). We used more accessible search criteria for our algorithms to increase the generalizability to different EHRs and administrative databases. Our algorithms were developed at a single center (Vanderbilt), so biases inherent to our institution may affect the portability of our algorithms. Prior studies, however, have demonstrated significant portability of EHR algorithms (22). We have future plans to validate our algorithms within other institutions' EHRs. In conclusion, we have developed and validated 3 EHR algorithms with high PPVs to identify patients with SLE accurately in the EHR. These algorithms represent powerful tools for clinical and translational researchers to identify and study patients with SLE efficiently and accurately.

AUTHOR CONTRIBUTIONS

All authors were involved in drafting the article or revising it critically for important intellectual content, and all authors approved the final version to be submitted for publication. Dr. Barnado had full access to all of the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.

Study conception and design. Barnado, Carroll, Denny, Crofford.
Acquisition of data. Barnado, Casey, Carroll.
Analysis and interpretation of data. Barnado, Carroll, Wheless, Denny, Crofford.

REFERENCES

1. Blumenthal D, Tavenner M. The "meaningful use" regulation for electronic health records. N Engl J Med 2010;363:501–4.
2. Pathak J, Kho AN, Denny JC. Electronic health records-driven phenotyping: challenges, recent advances, and perspectives. J Am Med Inform Assoc 2013;20:e206–11.
3. Ritche MD, Denny JC, Crawford DC, Ramirez AH, Weiner JB, Pulley JM, et al. Robust replication of genotype-phenotype associations across multiple diseases in an electronic medical record. Am J Hum Genet 2010;86:560–72.
4. Li T, Carls GS, Panopalis P, Wang S, Gibson TB, Goetzel RZ. Long-term medical costs and resource utilization in systemic lupus erythematosus and lupus nephritis: a five-year analysis of a large Medicaid population. Arthritis Rheum 2009;61:755–63.
5. Karve S, Candrilli S, Kappelman MD, Tolleson-Rinehart S, Tennis P, Andrews E. Healthcare utilization and comorbidity burden among children and young adults in the United Sates with systemic lupus erythematosus or inflammatory bowel disease. J Pediatr 2012;161:662–70.
6. Feldman CH, Hiraki LT, Liu J, Fischer MA, Solomon DH, Alarcón GS, et al. Epidemiology and sociodemographics of systemic lupus erythematosus and lupus nephritis among US adults with Medicaid coverage, 2000–2004. Arthritis Rheum 2013;65:753–63.
7. Yazdany J, Marafino BJ, Dean ML, Bardach NS, Duseja R, Ward MM, et al. Thirty-day hospital readmissions in systemic lupus erythematosus: predictors and hospital- and state-level variation. Arthritis Rheumatol 2014;66:2828–36.
8. Feldman CH, Hiraki LT, Winkelmayer WC, Marty FM, Franklin JM, Kim SC, et al. Serious infections among adult Medicaid beneficiaries with systemic lupus erythematosus and lupus nephritis. Arthritis Rheumatol 2015;67:1577–85.
9. Murray SG, Schmajuk G, Trupin L, Gensler L, Katz PP, Yelin E. National lupus hospitalization trends reveal rising rates of herpes zoster and declines in pneumocystis pneumonia. PLoS One 2016;11:e0144918.
10. Moores KG, Sathe NA. A systematic review of validated methods for identifying systemic lupus erythematosus (SLE) using administrative or claims data. Vaccine 2013;31:K62–73.
11. Liao KP, Cai T, Gainer V, Goryachev S, Zeng-Treitler Q, Raychaudhuri S, et al. Electronic medical records for discovery research in rheumatoid arthritis. Arthritis Care Res (Hoboken) 2010;62:1120–7.
12. Carroll RJ, Thompson WK, Eyler AE, Mandelin AM, Cai T, Zink RM, et al. Portability of an algorithm to identify rheumatoid arthritis in electronic health records. J Am Med Inform Assoc 2012;19:e162–9.
13. Kho AN, Pacheco JA, Peissig PL, Rasmussen L, Newton KM, Weston N, et al. Electronic medical records for genetic research: results of the eMERGE consortium. Sci Transl Med 2011;3:1–7.
14. Denny JC, Ritchie MD, Crawford DC, Schildcrout JS, Ramirez AH, Pulley JM, et al. Identification of genomic predictors of atrioventricular conduction: using electronic medical records as a tool for genome science. Circulation 2010;122:2016–21.
15. Denny JC, Crawford DC, Ritchie MD, Bielinski SJ, Basford MA, Bradford Y, et al. Variants near FOXE1 are associated with hypothyroidism and other thyroid conditions: using electronic medical records for genome- and phenome-wide studies. Am J Hum Genet 2011;89:529–42.
16. Shameer K, Denny JC, Ding K, Jouni H, Crosslin DR, de Andrade M, et al. A genome- and phenome-wide association study to identify genetic variants influencing platelet count and volume and their pleiotropic effects. Hum Genet 2014;133:95–109.
17. Roden DM, Pulley JM, Basford MA, Bernard GR, Clayton EW, Balser JR, et al. Development of a large-scale de-identified DNA biobank to enable personalized medicine. Clin Pharmacol Ther 2008;84:362–9.
18. Dupont WD, Plummer WD. Power and sample size calculations: a review and computer program. Control Clin Trials 1990;11:116–28.
19. R Core Team (2014). R: a language and environment for statistical computing. URL: http://www.R-project.org/.
20. Hochberg MC. Updating the American College of Rheumatology revised criteria for the classification of systemic lupus erythematosus [letter]. Arthritis Rheum 1997;40:1725.
21. Wei WQ, Teixeira PL, Mo H, Cronin RM, Warner JL, Denny JC. Combining billing codes, clinical notes, and medications from electronic health records provides superior phenotyping performance. J Am Med Inform Assoc 2016;23:e20–7.
22. Kho AN, Hayes MG, Rasmussen-Torvik L, Pacheco JA, Thompson WK, Armstrong LL, et al. Use of diverse electronic medical record systems to identify genetic risk for type 2 diabetes within a genome-wide association study. J Am Med Inform Assoc 2012;19:212–8.