HHS Public Access

Author manuscript

Arthritis Care Res (Hoboken). Author manuscript; available in PMC 2023 May 01.

Published in final edited form as:

Arthritis Care Res (Hoboken). 2022 May ; 74(5): 849–857. doi:10.1002/acr.24522.

Developing and Validating Methods to Assemble Systemic Lupus Erythematosus Births in the Electronic Health Record

April Barnado, MD, MSCI[1], **Amanda M. Eudy, PhD**[2], **Ashley Blaske, MD**[1], **Lee Wheless**[3], **Katie Kirchoff, MSHI**[4], **Jim C. Oates, MD**[5], **Megan E.B. Clowse, MD, MPH**[2]

[1]Department of Medicine, Vanderbilt University Medical Center, Nashville, TN, USA

[2]Department of Medicine, Duke University Medical Center, Durham, NC, USA

[3]Department of Dermatology, Data Science Institute, Vanderbilt University Medical Center, Nashville, TN, USA

[4]Biomedical Informatics Center, Medical University of South Carolina, Charleston, SC, USA

[5]Department of Medicine, Medical University of South Carolina, Charleston, SC, USA

Abstract

Objective: Electronic health records (EHRs) represent powerful tools to study rare diseases. We developed and validated EHR algorithms to identify SLE births across centers.

Methods: We developed algorithms in a training set using an EHR with over 3 million subjects and validated algorithms at two other centers. Subjects at all 3 centers were selected using ≥ 1 SLE ICD-9 or SLE ICD-10-CM codes and ≥ 1 ICD-9 or ICD-10-CM delivery code. A subject was a case if diagnosed with SLE by a rheumatologist and had a birth documented. We tested algorithms using SLE ICD-9 or ICD-10-CM codes, antimalarial use, a positive antinuclear antibody ≥ 1:160, and ever checked dsDNA or complements using both rule-based and machine learning methods. Positive predictive values (PPVs) and sensitivities were calculated. We assessed the impact of case definition, coding provider, and subject race on algorithm performance.

Results: Algorithms performed similarly across all three centers. Increasing the number of SLE codes, adding clinical data, and having a rheumatologist use the SLE code all increased the likelihood of identifying true SLE patients. All the algorithms had higher PPVs in African American vs. Caucasian SLE births. Using machine learning methods, total number of SLE codes and a SLE code from a rheumatologist were the most important variables in the model for SLE case status.

Conclusion: We developed and validated algorithms that use multiple types of data to identify SLE births in the EHR. Algorithms performed better in African American mothers than Caucasian mothers.

Corresponding Author: April Barnado, MD, MSCI, Address: 1161 21st Avenue South, T3113 MCN, Nashville, TN 37232, Telephone: (615) 322 – 4746, Fax: (615) 322 – 6248, april.barnado@vumc.org.

Conflict of Interest: none

Keywords

systemic lupus erythematosus; electronic health records; electronic phenotyping; delivery; birth; pregnancy

Introduction

Systemic lupus erythematosus (SLE) is an autoimmune disease that primarily affects women of childbearing age. Studying pregnancy outcomes in SLE is difficult given its relative rarity. SLE pregnancy studies are typically limited to a single-center cohort (1, 2) that may not reflect real-world pregnancy outcomes. Population-based studies have investigated SLE pregnancies but mainly in European populations (3–5).

Electronic health records (EHRs) contain rich, longitudinal data and serve as a powerful research tool (6). To harness the power of EHR data, validated methods are needed to identify subjects accurately. Using billing codes alone does not accurately identify SLE patients in the EHR (7, 8). Adding clinical data to billing codes (8) and using non-coded data (9) improves the accuracy of identifying SLE patients in the EHR. These methods did not focus on identifying SLE births. There is a paucity of EHR studies in SLE pregnancy outcomes with only one study using delivery data from multiple EHRs (10). This study used a one-time SLE ICD-9 billing code at delivery or discharge to identify SLE births and did not conduct chart review to confirm SLE case status. Building upon our prior SLE algorithms that incorporate clinical data with billing codes, we incorporated ICD-9 or ICD-10-CM codes to identify SLE deliveries in the EHR. With the transition of ICD-9 to ICD-10-CM codes in the US, we focused on developing algorithms that used either SLE ICD-9 or SLE ICD-10-CM codes. We also validated algorithms at multiple centers and investigated the impact of patient and provider factors on algorithm performance and portability. We then applied these algorithms to assemble a large, multi-center EHR cohort of SLE deliveries at three tertiary care centers in the Southeastern US.

Methods

Patient Selection

This study was approved by the institutional review board for each center. Due to center differences in how EHR data are stored and accessed, the methods for identifying SLE deliveries were slightly different as described below. An overview of our approach is illustrated in Figure 1. Vanderbilt University Medical Center (VUMC) served as a training set. Duke University Medical Center (DUMC) and Medical University of South Carolina (MUSC) served as external validation sets. Chart review rules were consistent across all three centers. A subject was defined as a case if diagnosed with SLE by a rheumatologist and had a delivery documented at the institution and after SLE diagnosis. Remaining subjects were classified as not a case, probable case, or missing. Subjects with cutaneous or drug-induced lupus or other autoimmune diseases were counted as not cases. Subjects who were given a SLE diagnosis by a non-rheumatology provider were counted as not cases. Probable cases were subjects who had a “probable SLE” diagnosis by a rheumatologist or

who were labeled as undifferentiated connective tissue disease or mixed connective disease by a rheumatologist. Probable subjects were counted as not cases in the primary analysis. Missing subjects who had no clinical documentation to determine case status were excluded. Delivery status was assessed on chart review at all three centers.

Vanderbilt University Medical Center (VUMC)

We used a de-identified version of VUMC's EHR called the Synthetic Derivative (6), which contains over 3.2 million subjects. We searched for potential SLE deliveries restricting to female subjects between the ages of 12 to 65 using ≥ 1 count of the SLE ICD-9 (710.0) or SLE ICD-10-CM codes (M32.1*, M32.8, M32.9) and ≥ 1 ICD-9 or ICD-10-CM code for delivery-related diagnoses. The ICD-9 delivery codes have been validated with positive predictive values > 90% (11) and used to assess pregnancy outcomes in other chronic diseases (12, 13) (Supplemental Table 1). Of these potential SLE cases, we randomly selected 100 for chart review to identify case status and to serve as a training set for algorithm development (Figure 1A).

Duke University Medical Center (DUMC)

At DUMC, potential patients with ≥ 1 SLE ICD-9 or ICD-10-CM code and ≥ 1 ICD-9 or ICD-10-CM delivery code restricting to female subjects between the ages of 12 to 65 were selected from the DEDUCE (Duke Enterprise Data Unified Content Explorer) dataset (Figure 1B). Of those potential patients, exclusions (i.e. no delivery at Duke or unknown pregnancy outcome) were applied to facilitate chart review. A full list of exclusions is in Supplemental Table 2.

Medical University of South Carolina (MUSC)

At MUSC, female patients between the ages of 12 and 65 with ≥ 1 SLE ICD-9 or ICD-10-CM code were selected from the Enterprise Data Warehouse from 2007 to 2017. As delivery data is stored in a different data warehouse (Research Data Warehouse), a second step was performed where subjects were selected who had ≥1 ICD-9 or ICD-10-CM delivery code and delivery data available (Figure 1C). Of these potential subjects, chart review was conducted.

Algorithm Development and Validation

A priori, we selected clinically important criteria that would be available in the EHR. We selected SLE ICD-9 and ICD-10-CM code counts, ever documented antimalarials, a positive antinuclear antibody (ANA) ≥ 1:160, and ever checked dsDNA or complements (C3 or C4). Occurrences of billing codes represent distinct days. Antimalarials included were hydroxychloroquine, plaquenil, chloroquine, quinacrine, and aralen. We tested algorithms using ≥ 1, ≥ 2, ≥3, and ≥ 4 code counts of the SLE ICD-9 code, SLE ICD-10-CM codes, or SLE ICD-9 or SLE ICD-10-CM codes. With the transition of ICD-9 to ICD-10-CM codes in the US on October 1, 2015, our EHR data spans this date. To ensure both historical and more recent SLE patients with deliveries are both captured, we focused on developing algorithms that used either SLE ICD-9 or SLE ICD-10-CM codes. We then combined these codes with the above clinical data using "and" or "or" for possible algorithms. The positive

predictive value (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 the total number of confirmed cases. To fit the algorithm, the subject had to have available data for that particular algorithm's criteria. If labs were not checked at the center, they were 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.

Cohort assembly

The algorithm with the highest F-score (≥ 4 counts of the SLE ICD-9 or ICD-10-CM codes) was applied across all centers to identify potential deliveries. All the subjects that fit the algorithm were chart reviewed to determine SLE case status, defined as SLE diagnosis by a rheumatologist. Only pregnancies that delivered at the center with available outcomes that occurred after SLE diagnosis were included. Data were available for VUMC from 1993 – 2017, DUMC 2007 – 2018, and MUSC 2007–2017.

Sensitivity Analyses

We focused on the performance of algorithms that used SLE ICD-9 or ICD-10-CM codes with performance of algorithms using only SLE ICD-9 or only SLE ICD-10-CM codes available in the supplement. The primary analysis defined cases as diagnosed with SLE by a rheumatologist on chart review and allowed ICD-9 or ICD-10-CM codes to be billed by any provider. One sensitivity analysis changed the case definition to also include "probable" SLE cases. A second sensitivity analysis included only SLE ICD-9 or ICD-10-CM codes billed by a rheumatology provider. Additionally, we determined differences in algorithm performance by maternal race.

Machine Learning Methods

In addition to rule-based algorithms, we used machine learning methods, random forest (RF) and extreme gradient boosting (XGB) for algorithm development. RF builds multiple classification trees (a "forest") using a random sample of input variables for each tree (14, 15). The final classification is an average of the forest. XGB is an ensemble method that is the summation of multiple models where each successive model attempts to correct errors in the previous model to improve overall performance. Data across 3 centers were randomly divided in training (80%) and testing (20%) sets. For race-stratified analyses, to increase sample size, training and testing sets were 70% and 30%, respectively. Models were constructed using the training set with 5-fold cross validation, and were tuned using the caret package (16, 17). Final model performance was assessed using the test set. The ranger package was used for RF models (18), and the xgboost package for XGB models using method = "xgbTree" in the caret framework (19). We reported algorithms with the highest PPVs in the test set and identified the most important variables in the models. Model input variables including the following: total number of SLE ICD codes, SLE code from a rheumatologist, ever antimalarial use, ANA positive, ever checked dsDNA, ever checked C3, ever checked C4, age, race, SLE duration defined as first SLE code to delivery date, EHR

duration defined as first code for any condition to delivery date, and center. All analyses were conducted in R version 3.5.1.

Results

Description of the training set

An overview of our approach is illustrated in Figure 1. A training set was created at VUMC by applying at least one SLE and one delivery ICD-9 or ICD-10 CM codes to the Synthetic Derivative resulting in 433 potential SLE deliveries. Of the 433, 100 were randomly selected for chart review. Of these, 40 subjects were SLE cases with 39 subjects having a delivery documented after SLE diagnosis. There were 37 subjects who were not SLE cases, 16 with a "probable" SLE diagnosis, and 7 with missing clinic notes. Of the 37 subjects not classified as SLE, 21 had alternative autoimmune diagnoses with the most common being a subject with a positive autoantibody (Supplemental Table 3).

Description of the validation sets

A validation set was created at DUMC by applying at least one SLE and one delivery ICD-9 or ICD-10 CM codes to DEDUCE resulting in 560 potential SLE deliveries. Of these, 192 had deliveries that occurred after a SLE diagnosis. On chart review of these 192, 95 were a SLE case and 36 "probable" SLE. Of the remaining subjects, 61 did not have SLE of which 31 had alternative autoimmune diagnoses with the most common being cutaneous lupus (Supplemental Table 4).

A second validation set was created at MUSC by applying at least one SLE ICD-9 or ICD-10-CM codes and selecting for female subjects in the EHR. Of these 3,715 potential SLE subjects, subjects with at least one delivery ICD-9 or ICD-10-CM code and a delivery documented at MUSC after SLE diagnosis resulted in 75 potential SLE deliveries. Of these, 38 were a SLE case and 11 "probable" SLE. Of the remaining subjects, 26 did not have SLE of which 15 had alternative autoimmune diagnoses with the most common being cutaneous lupus (Supplemental Table 5).

Algorithms using ICD-9 or ICD-10-CM codes

Algorithm performances using counts of either SLE ICD-9 or ICD-10-CM codes in the training (VUMC) and validation (DUMC, MUSC) sets are shown as a summary in Table 1 with full data in Supplemental Tables 6–8. Algorithm performances using only SLE ICD-9 codes or only SLE ICD-10-CM codes are available in Supplemental Tables 6–8. As data duration for ICD-9 codes differed in training vs. validation centers, we limited the training set data duration to 2007 – 2017 to match the 2 validation centers. Within the training set, the ICD-9 code algorithm performances for the restricted 2007–2017 duration were similar to the algorithm performances for the full data duration 1993 – 2017 (Supplemental Table 9). Requiring more code counts of SLE ICD-9 or ICD-10-CM codes increased PPVs but decreased sensitivities. Across three centers, increasing the number of SLE ICD-9 or ICD-10-CM code counts increased PPVs from 50–56% for ≥ 1 code to 64–81% for ≥ 4 codes.

Algorithms incorporating clinical data

We investigated adding ever antimalarial documented, ever checked labs (dsDNA, C3, or C4), and a positive ANA (≥ 1:160) to ICD-9 and ICD-10-CM codes. Algorithms that incorporated clinical data had higher PPVs compared to algorithms using only codes (Table 1, Supplemental Tables 6–8). This addition, however, lowered sensitivities, as some SLE patients didn't have data documented in the center's EHR. Across all three centers, adding antimalarials to the codes improved PPVs most robustly. Adding ever checked labs to the codes increased PPVs slightly. Adding a positive ANA didn't significantly increase PPVs but decreased sensitivities.

Case definition

For the above analyses, probable SLE subjects were counted as "not cases." We examined algorithm performance with counting probable SLE subjects as cases (Table 1, Supplemental Tables 10–12). With this alternative case definition, PPVs increased substantially while sensitivities decreased for all algorithms.

Billing Code Provider

For the above analyses, we allowed for any provider to use the SLE codes. We investigated if a rheumatology provider using the SLE codes impacted algorithm performance (Table 1, Supplemental Tables 13–15). With requiring a rheumatology provider to use either a SLE ICD-9 or ICD-10-CM code, PPVs significantly increased at all centers with a decrease in sensitivities. Adding clinical data didn't significantly improve PPVs, as PPVs were already relatively high.

Subject Race

We evaluated the impact of subject race on algorithm performance. Prevalence for African American subjects was 31% in the training set and 51% and 55% in the validation sets. While sensitivities were similar, PPVs were significantly higher in African Americans compared to Caucasians (Table 1). Pooling data from the three centers, the algorithm with ≥ 4 counts of the SLE ICD-9 or ICD-10-CM codes had a PPV of 78% in African Americans compared to 54% in Caucasians (Supplemental Table 16). Adding clinical data, such as ever labs checked, to the codes increased PPVs in Caucasians but not in African Americans (Supplemental Table 16). Requiring rheumatology to use the codes increased PPVs significantly in both Caucasians and African Americans (Supplemental Table 17).

Machine Learning Methods

We performed random forest (RF) and extreme gradient boosting (XGB) models for algorithm development. For RF, the algorithm with the highest PPV included 500 trees and sampled two random variables per tree with a PPV of 79%, sensitivity of 80%, an F-score of 80%, negative predictive value (NPV) of 81%, and an AUC of 87% in the training set (Supplemental Table 18). The most important variables in the model were total number of SLE ICD codes and rheumatology using the SLE codes. Model performance varied by race with an F-score of 0.87 in African Americans vs. 0.67 in Caucasians. For XGB, the highest-performing model had a PPV of 79%, sensitivity of 82%, an F-score of 80%,

NPV of 82%, and an AUC of 84% in the training set (Supplemental Table 20). The most important variables in the model were total number of SLE ICD codes and rheumatology using the SLE codes. Model performance varied by race with an F-score of 0.89 in African Americans vs. 0.71 in Caucasians.

Highest performing algorithms

We assessed algorithm performance with PPV, sensitivity, and F-score, a measure that accounts for both PPV and sensitivity. Algorithms' performances varied across the three centers leading to different high performing algorithms at each center. Algorithms with the highest PPVs included higher SLE code counts, incorporated clinical data, expanded the case definition to include probable and definite SLE cases, and required rheumatology to use the SLE codes (Figure 2). Algorithms that incorporated ≥ 4 SLE codes coded by rheumatology along with ever antimalarial documented had PPVs from 90 to 100% across the three centers. Algorithms with the highest sensitivities were algorithms that used fewer SLE code counts and incorporated either SLE ICD-9 or ICD-10-CM codes (Figure 2). The algorithm with the highest F-score across the three centers used ≥ 4 counts of the SLE ICD-9 or ICD-10-CM codes and was 87% at VUMC, 79% at DUMC, and 73% at MUSC (Table 1).

Cohort assembly

Deploying the algorithm with the highest F-score (≥ 4 counts of the SLE ICD-9 or ICD-10-CM codes) resulted in 438 possible SLE deliveries across the three centers (Table 2). In this cohort, mean age at first delivery was 29.5 ± 1.2 years with Caucasian deliveries at 51% and 42% African American, 3% Asian, 4% other. Only 5% of deliveries were of Hispanic ethnicity.

Discussion

We have harnessed the power of the EHR to develop, validate, and deploy algorithms that assemble a rare event across multiple EHRs. To the best of our knowledge, this is one of the first successful applications of assembling SLE and SLE deliveries from several centers using EHR data in the United States. This is important work because it establishes valuable methods for researchers to not only identify SLE and SLE deliveries but also other outcomes across EHRs. In summary, increasing number of SLE codes, adding ever antimalarial documented to codes, expanding case definition to probable and definite SLE cases, and requiring rheumatology to use SLE codes all improved algorithms' PPVs. Subject race had a significant impact on algorithm performance with significantly higher PPVs in African Americans compared to Caucasians.

While there are validated algorithms to identify SLE in the EHR (8, 9), there are no studies on accurately identifying SLE deliveries in the EHR. Literature in sickle cell anemia (12) has evaluated and validated delivery codes in studying pregnancy outcomes (11,13, 20, 21). Our study used these validated delivery codes and built upon our work in SLE EHR algorithms (8, 9). Our EHR data spanned the ICD-9 to ICD-10-CM code transition time. Some historical patients in our dataset only have ICD-9 codes while

more recently diagnosed patients only have ICD-10-CM codes. Some patients with more longitudinal data have both ICD-9 and ICD-10-CM codes. Therefore, we focused on algorithms that incorporated ICD-9 or ICD-10-CM SLE codes to capture both historical and newly diagnosed SLE pregnancies. This approach is more generalizable to EHRs that likely contain both ICD-9 and ICD-10-CM codes and not just solely ICD-9 or ICD-10-CM codes. Including either ICD-9 or ICD-10-CM codes also limits cohort effects on the algorithms' performances. We developed multiple algorithms that incorporate different types of data to meet researchers' diverse goals. We also performed validation and found good portability of the algorithms. Lastly, we identified key factors such as clinical data, case definition, subject race, and coding provider that all significantly impact algorithm performance.

As expected, requiring higher counts of SLE ICD-9 or ICD-10-CM codes resulted in algorithms with higher PPVs but lower sensitivities. The more visits a potential SLE patient has, the more times a SLE code is used with the clinician feeling confident with the diagnosis. In general, adding clinical data to the codes improved algorithms' PPVs but decreased sensitivities. For example, some SLE patients did not have clinical data such as a positive ANA within the center's EHR, as they were followed by an outside rheumatologist. Using SLE ICD-9 or ICD-10-CM codes from only rheumatologists resulted in algorithms with high PPVs without requiring clinical data. These algorithms would be useful if clinical data is not available but limits the sample to SLE women managed by center rheumatologists.

As expected, broadening the SLE case definition to include "probable" patients increased PPVs while decreasing sensitivities. Compared to definite SLE patients, probable SLE patients were more likely to have fewer SLE codes. Algorithms requiring higher counts of the SLE codes would then exclude more of these probable SLE patients, resulting in lower sensitivities. We used a specialist diagnosis for SLE versus using ACR SLE criteria (22), as ACR SLE criteria are not documented systematically in notes. We previously demonstrated that requiring documentation of ACR SLE criteria excludes approximately 26% of true SLE patients (8). Researchers can, however, select a case definition based on their study's goals and available data.

Across all three centers, algorithms had higher PPVs but similar sensitivities in African American vs. Caucasian patients. While increasing code counts and adding clinical data improved algorithms' PPVs for Caucasians somewhat, requiring rheumatology coding dramatically improved PPVs. Our results suggest a high rate of SLE over-labeling in Caucasians, particularly by physicians other than rheumatologists. Thus, different algorithms may be needed for different races. Specifically, algorithms to identify Caucasians accurately may require rheumatologists to use SLE billing codes. The impact of race on EHR phenotyping has not been explored in SLE or other chronic conditions. We hypothesize that the higher prevalence of SLE in African Americans compared to Caucasians (23–26) may contribute to this observation. PPVs are a function of disease prevalence while sensitivities are a function of the algorithm.

In addition to rule-based algorithms, we performed machine learning models. These models had a similar F-score to the high-performing rule-based algorithm of ≥ 4 counts of

SLE ICD-9 or ICD-10-CM codes. The machine learning methods confirmed results from the rule-based algorithms including variable model performance based on subject race. Machine learning methods, particularly XGB, are robust (27) and have advantages including automatic model tuning and the ability to model complex interactions.

We developed and validated algorithms to identify SLE deliveries in the EHR. Similar methodology can identify and assemble other rare diseases or outcomes in the EHR. Researchers can choose methods based on available data and research goals (Figure 2). If the goal is to identify subjects with high certainty, one would select an algorithm with the highest PPV. In contrast, if one wants to select as many subjects as possible to increase sample size, one would select an algorithm with a high sensitivity and F-score. We chose to use an algorithm with a high F-score to amass the largest number of true SLE deliveries.

While we validated multiple EHR-based algorithms, our study has limitations. We started our search for possible SLE deliveries using at least one SLE ICD-9 or ICD-10-CM codes. This search strategy in finding SLE in the EHR has a NPV of 98% (8), so we anticipate very few potential SLE deliveries were missed. Search strategies for identifying training and validation sets varied slightly due to differences in how data is stored and accessed at the three centers. In machine learning methods, model performance did not vary by center. Therefore, center heterogeneity had minimal impact on algorithm performance. Our algorithms were developed and validated at three tertiary care referral centers in the Southeastern US, which may limit generalizability to other centers.

While the EHR does not substitute for prospective cohort studies, EHRs contain longitudinal, real-word data that can dramatically increase the efficiency and sample size of a study. Using one of our validated, high-performing algorithms, we assembled over 400 potential SLE deliveries across three centers. With this large SLE delivery cohort, we will have the power in future studies to examine the impact of disease and provider factors on important outcomes such as preterm delivery and preeclampsia in SLE.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Financial Support: Supported by grants NIH/NIAMS 1K08 AR072757-01 (Barnado), NIH/NCRR UL1 RR024975 (VUMC), NIH/NCATS ULTR000445 (VUMC), NIH/NCATS UL1TR002553 (Duke), NIH/NCATS 1KL2TR002554 (Eudy), NIH/NIAMS P60 AR062755 (MUSC), NIH/NIAMS P30 AR072582 (MUSC), and NIH/NCATS UL1 TR001450 (MUSC).

References

1. Andrade RM, McGwin G Jr, Alarcón GS, Sanchez ML, Bertoli AM, Fernández M, et al. Predictors of post-partum damage accrual in systemic lupus erythematosus: data from LUMINA, a multiethnic US cohort (XXXVIII). Rheumatology (Oxford) 2006; 45: 1380–4. [PubMed: 16880189]
2. Buyon JP, Kim MY, Guerra MM, Laskin CA, Petri M, Lockshin MD, et al. Predictors of Pregnancy Outcomes in Patients With Lupus: A Cohort Study. Ann Intern Med 2015; 163: 153–63. [PubMed: 26098843]

3. Palmsten K, Simard JF, Chambers CD, Arkema E. Medication use among pregnant women with systemic lupus erythematosus and general population comparators. Rheumatology (Oxford) 2017; 56: 561–569. [PubMed: 28013193]
4. Skorpen CG, Lydersen S, Gilboe IM, Skomsvoll JF, Salvesen KA, Palm O, et al. Influence of disease activity and medications on offspring birth weight, pre-eclampsia and preterm birth in systemic lupus erythematosus: a population-based study. Ann Rheum Dis 2018; 77: 264–269. [PubMed: 29092851]
5. Zusman EZ, Sayre EC, Avina-Zubieta JA, De Vera MA. Patterns of medication use before, during and after pregnancy in women with systemic lupus erythematosus: a population-based cohort study. Lupus 2019; 28: 1205–1213. [PubMed: 31311418]
6. 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. [PubMed: 18500243]
7. Moores KG, Sathe. A systematic review of validated methods for identifying systemic lupus erythematosus (SLE) using administrative or claims data. Vaccine 2013; 31 Suppl 10: K62–73. [PubMed: 24331075]
8. Barnado A, Casey C, Carroll RJ, Wheless L, Denny JC, Crofford LJ. Developing Electronic Health Record Algorithms That Accurately Identify Patients With Systemic Lupus Erythematosus. Arthritis Care Res (Hoboken) 2017; 69: 687–693. [PubMed: 27390187]
9. Jorge A, Castro VM, Barnado A, Gainer V, Hong C, Cai T, et al. Identifying lupus patients in electronic health records: Development and validation of machine learning algorithms and application of rule-based algorithms. Semin Arthritis Rheum 2019; 49: 84–90. [PubMed: 30665626]
10. Williams A, Grantz K, Seeni I, Robledo C, Li S, Ouidir M, et al. Obstetric and neonatal complications among women with autoimmune disease. J Autoimmun 2019; 103: 102287. [PubMed: 31147159]
11. Yasmeen S, Romano PS, Schembri M, Keyzer JM, Gilbert WM. Accuracy of obstetric diagnoses and procedures in hospital discharge data. Am J Obstet Gynecol 2006; 194: 992–1001. [PubMed: 16580288]
12. Boulet SL, Okoroh EM, Azonobi I, Grant A, Hooper WC. Sickle cell disease in pregnancy: maternal complications in a Medicaid-enrolled population. Matern Child Health J 2013; 17: 200–7. [PubMed: 23315242]
13. Zhang S, Cardarelli K, Shim R, Ye J, Booker KL, Rust G. Racial disparities in economic and clinical outcomes of pregnancy among Medicaid recipients. Matern Child Health J 2013; 17: 1518–25. [PubMed: 23065298]
14. Breiman L, Friedman J, Olshen RA, Stone CJ. Classification and regression trees. In: The Wadsworth statistics/probability series. 1st ed. Belmont(CA): Wadsworth International Group; 1983.
15. Breiman L Random Forests. Machine Learning 2001; 45: 5–32.
16. Kuhn M Building Predictive Models in R Using the caret Package. J Stat Softw 2008; 28: 1–26. [PubMed: 27774042]
17. R Core Team. R: A Language and Environment for Statistical Computing, 2018. https://www.R-project.org/
18. Wright MN, Ziegler A. ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. J Stat Softw 2017; 77: 1–17.
19. Chen T, He T, Benesty M, Khotilovich V, Tang Y, Cho H, et al. xgboost: Extreme Gradient Boosting, 2019. https://CRAN.R-project.org/package=xgboost
20. Hardy JR, Holford TR, Hall GC, Bracken MB. Strategies for identifying pregnancies in the automated medical records of the General Practice Research Database. Pharmacoepidemiol Drug Saf 2004; 13: 749–59. [PubMed: 15386720]
21. Devine S, West S, Andrews E, Tennis P, Hammad TA, Eaton S, et al. The identification of pregnancies within the general practice research database. Pharmacoepidemiol Drug Saf 2010; 19: 45–50. [PubMed: 19823973]
22. Hochberg MC. Updating the American College of Rheumatology revised criteria for the classification of systemic lupus erythematosus. Arthritis Rheum 1997; 40: 1725.

23. Lim SS, Bayakly AR, Helmick CG, Gordon C, Easley KA, Drenkard C. The incidence and prevalence of systemic lupus erythematosus, 2002–2004: The Georgia Lupus Registry. Arthritis Rheumatol 2014; 66: 357–68. [PubMed: 24504808]

24. Somers EC, Marder W, Cagnoli P, Lewis EE, DeGuire P, Gordon C, et al. Population-based incidence and prevalence of systemic lupus erythematosus: the Michigan Lupus Epidemiology and Surveillance program. Arthritis Rheumatol 2014; 66: 369–78. [PubMed: 24504809]

25. Centers for Disease, C. and Prevention. Trends in deaths from systemic lupus erythematosus--United States, 1979–1998. MMWR Morb Mortal Wkly Rep 2002; 51: 371–4. [PubMed: 12018384]

26. Krishnan E, Hubert HB. Ethnicity and mortality from systemic lupus erythematosus in the US. Ann Rheum Dis 2006; 65: 1500–5. [PubMed: 16627544]

27. Fernández-Delgado M, Sirsat MS, Cernadas E, Alawadi S, Barro S, Febrero-Bande M. An extensive experimental survey of regression methods. Neural Netw 2019; 111: 11–34. [PubMed: 30654138]

Significance and Innovations:

- To the best of our knowledge, we are the first to assemble an EHR-based SLE cohort and SLE birth cohort across multiple centers in the United States.
- We develop, validate, and successfully deploy EHR-based algorithms to identify a subset of patients with a rare disease across multiple centers.
- We demonstrate key factors of clinical data, case definition, coding provider, and subject race that impact EHR algorithm performance and portability.
- We employed both traditional, rule-based algorithm methods as well as machine learning techniques including extreme gradient boosting.
- The performance of the SLE delivery algorithms varied by race with higher positive predictive values (PPVs) in African American mothers compared to Caucasian mothers.

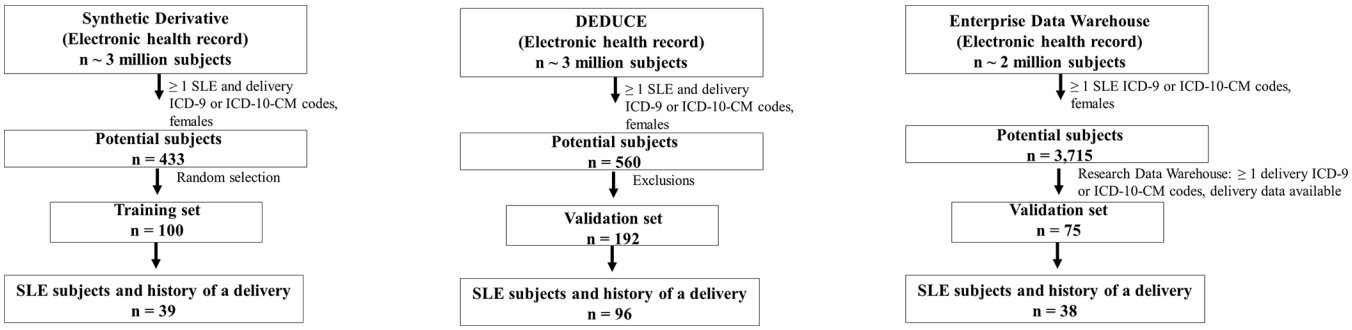

Figure 1. Training and validation sets.

Training set (A) was formed at VUMC starting with the Synthetic Derivative, a de-identified electronic health record (EHR), and applying ≥ 1 SLE and delivery ICD-9 or ICD-10-CM codes to female subjects resulting in 433 potential systemic lupus erythematosus (SLE) subjects with deliveries. A random 100 subjects were then selected for chart review to determine SLE case status and to ensure delivery occurred after SLE diagnosis, of which 39 subjects met these criteria. N refers to number of SLE subjects and not pregnancies. **Validation set (B)** was created at DUMC by applying ≥ 1 SLE and ≥ 1 delivery ICD-9 or ICD-10 CM codes to females subjects in a de-identified electronic health record called DEDUCE (Duke Enterprise Data Unified Content Explorer) resulting in 560 potential SLE subjects. Exclusions were then applied to facilitate chart review and ensure SLE subjects with available delivery data that occurred after SLE diagnosis were obtained (see Supplemental Table 2 for full list of exclusions). These 192 subjects were then chart reviewed yielding 96 SLE cases with a history of a delivery. **Validation set (C)** was created at MUSC by applying ≥ 1 SLE ICD-9 or ICD-10-CM codes while restricting to female subjects in the Enterprise Data Warehouse (EDW) resulting in 3,715 subjects. As delivery data is stored in a different data warehouse (Research Data Warehouse), a second step was performed where subjects were then selected who had ≥ 1 ICD-9 or ICD-10-CM delivery code and delivery data available. This step resulted in 75 potential SLE subjects who had pregnancy data available at MUSC and had deliveries that occurred after a SLE diagnosis. Of these 75 subjects, 38 subjects had a confirmed diagnosis by a rheumatologist on chart review.

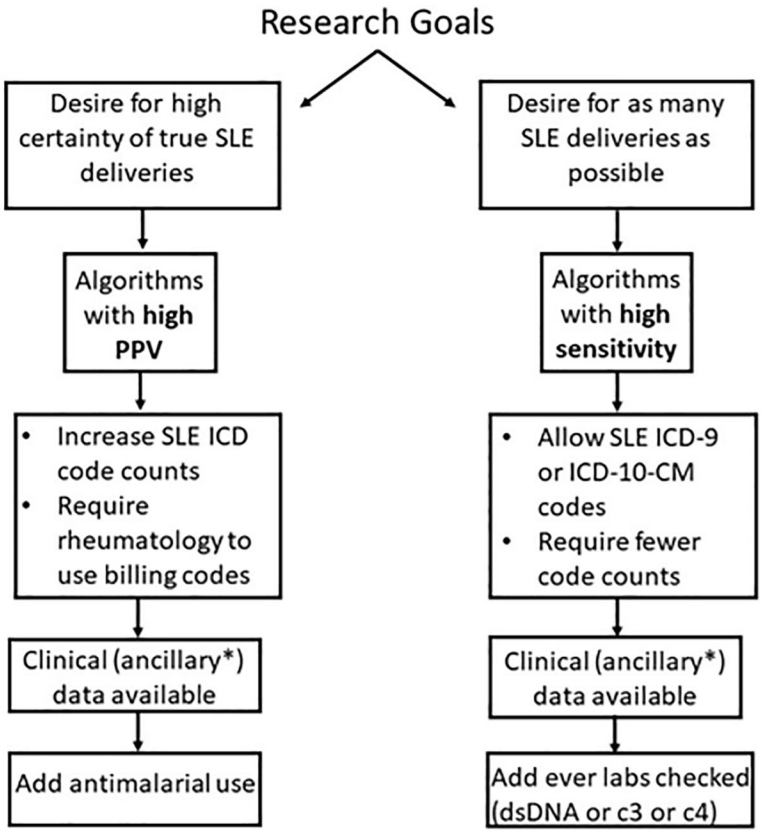

Figure 2. Guide to selecting algorithms to identify SLE deliveries in the Electronic Health Record.

Algorithms can be selected based on the researcher's goals and available clinical or ancillary data. If there is a desire for high certainty for true SLE deliveries, then one would select an algorithm with a high PPV. If chart review is not available or possible to confirm case status, one would also want to select an algorithm with a high PPV. Alternatively, if there is a desire to assemble as many possible SLE deliveries as possible, one would select an algorithm with a high sensitivity. If clinical or ancillary data is available such as structured electronic health record data including laboratory values or medications, this will further influence algorithm selection.

Table 1.

Summary of algorithm performances.

	VUMC Training Set			DUMC Validation Set			MUSC Validation Set		
Algorithms	**PPV**	**Sensitivity**	**F-score**	**PPV**	**Sensitivity**	**F-score**	**PPV**	**Sensitivity**	**F-score**
ICD-9 or ICD-10-CM code counts									
≥ 1 count	56%	100%	72%	50%	100%	67%	51%	100%	68%
≥ 4 counts	81%	95%	87%	71%	88%	79%	64%	84%	73%
ICD-9 or ICD-10-CM code counts AND ever antimalarial documented									
≥ 1 count	68%	85%	76%	67%	76%	71%	72%	94%	82%
≥ 4 counts	83%	83%	83%	76%	70%	73%	74%	81%	77%
ICD-9 or ICD-10-CM code counts AND ANA positive [3]									
≥ 1 count	64%	76%	69%	56%	96%	71%	64%	91%	75%
≥ 4 counts	79%	76%	77%	72%	86%	78%	76%	83%	79%
ICD-9 or ICD-10-CM code counts AND ever labs checked [4]									
≥ 1 count	62%	95%	75%	59%	91%	72%	61%	87%	72%
≥ 4 counts	84%	93%	88%	74%	81%	77%	68%	74%	71%
Case Definition: Definite and Probable SLE, ICD-9 or ICD-10-CM code counts									
≥ 1 count	81%	100%	90%	63%	100%	77%	75%	100%	86%
≥ 4 counts	91%	67%	77%	78%	78%	78%	90%	80%	85%
Billing Code Provider: Require Rheumatology to use SLE ICD-9 or ICD-10-CM code counts									
≥ 1 count	79%	78%	78%	82%	81%	81%	77%	54%	63%
≥ 4 counts	85%	73%	79%	96%	56%	71%	100%	24%	39%
Maternal race: African American women, ICD-9 or ICD-10-CM code count									
≥ 1 count	71%	100%	83%	60%	100%	75%	61%	100%	76%
≥ 4 counts	84%	94%	89%	81%	85%	83%	73%	88%	80%
Maternal race: Caucasian women, ICD-9 or ICD-10-CM code count									
≥ 1 count	48%	100%	65%	33%	100%	50%	34%	100%	51%
≥ 4 counts	69%	90%	78%	88%	55%	68%	41%	70%	52%
Maternal race: African American women, ICD-9 or ICD-10-CM code count, require rheumatology to use codes									

	VUMC Training Set			DUMC Validation Set			MUSC Validation Set		
Algorithms	**PPV**	**Sensitivity**	**F-score**	**PPV**	**Sensitivity**	**F-score**	**PPV**	**Sensitivity**	**F-score**
≥ 1 count	93%	76%	84%	88%	76%	82%	80%	67%	73%
≥ 4 counts	92%	71%	80%	97%	51%	67%	100%	33%	50%
Maternal race: Caucasian women, ICD-9 or ICD-10-CM code count, require rheumatology to use codes									
≥ 1 count	68%	75%	71%	71%	83%	77%	60%	30%	40%
≥ 4 counts	82%	70%	76%	93%	54%	68%	100%	10%	18%

[1] SLE ICD-9 code: 710.0

[2] SLE ICD-10-CM codes: M32.1*, M32.8, M32.9

[3] ANA positive (≥ 1:160)

[4] Ever labs checked included dsDNA, C3, or C4

Table 2.

Characteristics of Systemic Lupus Erythematosus deliveries using a high-performing algorithm across three centers.

	≥ 4 SLE ICD-9 or ICD-10-CM and delivery codes[1]				Probable and Definite SLE deliveries[1]				Definite SLE deliveries[1]			
	Overall n[2] = 438	**Center 1 (VUMC) n = 269**	**Center 2 (DUMC) n = 119**	**Center 3 (MUSC) n = 50**	**Overall n = 369**	**Center 1 n = 231**	**Center 2 n = 93**	**Center 3 n = 45**	**Overall n = 286**	**Center 1 n = 170**	**Center 2 n = 84**	**Center 3 n = 32**
Age at Delivery Mean ± standard deviation (Range)	29.5 ± 1.2 (16–46)	28.4 ± 5.6 (16–41)	30.8 ± 5.9 (19–46)	29.4 ± 5.2 (19–41)	29.5 ± 1.4 (16–46)	28.3 ± 5.6 (16–41)	31.0 ± 5.9 (19–46)	29.3 ± 5.2 (19–41)	29.2 ± 1.4 (16–46)	28.1 ± 5.7 (16–41)	30.8 ± 5.9 (19–46)	28.8 ± 5.6 (19–41)
Race, n (%)												
African American	175 (42%)	83 (33%)	62 (62%)	30 (64%)	157 (44%)	74 (34%)	53 (57%)	30 (71%)	133 (48%)	61 (38%)	50 (60%)	22 (76%)
Caucasian	215 (51%)	160 (63%)	38 (32%)	17 (36%)	171 (48%)	134 (61%)	25 (27%)	12 (29%)	121 (44%)	93 (57%)	21 (25%)	7 (24%)
Asian	14 (3%)	7 (3%)	7 (6%)	0 (0%)	12 (4%)	6 (3%)	6 (6%)	0 (0%)	11 (4%)	6 (4%)	5 (6%)	0 (0%)
Other	17 (4%)	5 (2%)	12 (10%)	0 (0%)	14 (4%)	5 (2%)	9 (10%)	0 (0%)	10 (4%)	2 (1%)	8 (10%)	0 (0%)
Unknown	17	14	0	3	15	12	0	3	11	8	0	3
Ethnicity, n												
Hispanic	24 (5%)	17 (6%)	7 (6%)	0 (05)	21 (6%)	17 (7%)	4 (4%)	0 (0%)	14 (5%)	10 (6%)	4 (5%)	0 (0%)

[1]Applying the algorithm of ≥ 4 SLE ICD-9 or ICD-10-CM and delivery codes resulted in 438 possible deliveries across the 3 centers. Of these 438 deliveries, 369 were probable and definite SLE deliveries based on chart review and 286 were definite SLE deliveries.

[2]Note n refers to number of deliveries. Note EHR delivery data available at VUMC 1993–2017, DUMC 2007–2018, and MUSC 2014–2017.