MEDINFO 2023 — The Future Is Accessible
J. Bichel-Findlay et al. (Eds.)

doi:10.3233/SHTI230947

Phenotype Systemic Lupus Erythematosus Patients from EPIC Cosmos

Jay PATEL[a,1], Lixia YAO[a], Ernest VINA[b], David FLEECE[c], Jayatilleke ARUNDATHI[b], Roberto CARICCHIO[d] and Huanmei WU[a]
[a] *Department of Health Services Administration and Policy, College of Public Health Temple University, Philadelphia, PA, USA*
[b] *Division of Rheumatology, Temple University School of Medicine, PA, USA*
[c] *Temple University School of Medicine and Temple University Hospital, PA, USA*
[d] *Division of Rheumatology, UMass Memorial Health, Worcester, MA, USA*

Abstract. Systemic Lupus Erythematosus (SLE) is a widespread autoimmune disease for which early diagnosis is paramount in improving clinical outcomes. In this project, we used the de-identified patients from Epic Cosmos to retrieve the ICD code for SLE, checked data quality based on the EULAR/ACR classification systems, created an approach to determine the SLE patients, and performed statistical analyses on lab tests and clinical characteristics. Our preliminary results showed that clinical notes must be reviewed to improve the completeness, as structured EHR data fields provide limited information in determining if a patient meets the established classification criteria.

Keywords. Systemic lupus erythematosus, phenotype, EPIC COSMOS

1. Introduction

Systemic Lupus Erythematosus (SLE) is a heterogeneous autoimmune disease that causes widespread inflammation and tissue damage in the affected organs, such as joints, skin, brain, lungs, kidneys, and blood vessels, which affects 20 to 150 people per 100,000 in the US, especially in women (3-11 times higher than in men) [1]. Unfortunately, SLE is not curable, with treatment options aimed at therapeutic care and minimizing its complications. Moreover, SLE patients still have significant mortality and carry a risk of progressive organ damage responsible for reduced quality of life and increased healthcare costs [2,3]. One factor exacerbating the high mortality, morbidity, and low treatment rate is the delay between SLE symptom onset and diagnosis, about two years on average [3].

Recently, the European League Against Rheumatism (EULAR) and the American College of Rheumatology (ACR) have introduced a classification system for SLE for diagnosis with excellent sensitivity and specificity [4,5]. Based on this classification system, SLE is diagnosed using objective (e.g., presence of antiphospholipid antibodies, SLE-specific antibodies) and clinical findings. Unfortunately, there is currently no model that can predict the risk of SLE to prevent and delay its progression. To our

[1] Corresponding Author: Jay Patel, email: patel.jay@temple.edu.

knowledge, no study has attempted to create a prediction model for early diagnosis of SLE using longitudinal EHR data.

Our long-term goal is to create an early SLE prediction model using longitudinal EHR data. We hypothesize that the EHR data would be able to provide us with patients' objective and clinical information with at least 50% completeness on phenotype SLE patients. The objective of this paper is twofold: 1) To determine the quality of the EHR data to assess the completeness of objective and clinical findings required for SLE diagnose using the EULAR/ACR classification. 2) To determine how many patients with a positive diagnosis of SLE (ICD code M32.*) meet the EULAR/ACR classification.

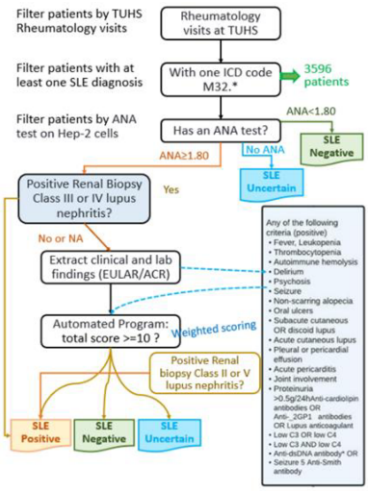

Figure 1. The overall workflow.

2. Methods

2.1. Data source and acquisition

We accessed patient information originating from the EHR system used by Temple University Health System (Epic Systems, Inc., Verona, WI) using Epic's Cosmos platform, which combines billions of clinical data points to form high-quality, representative, and integrated data sets. The information includes patients' medical diagnoses, procedures, lab reports, vitals, and socio-demographic information. Our overall workflow is presented in Figure 1. We included those patients who visited rheumatology at TUHS and had at least one physician-assigned ICD code for lupus (*M32.**) between 01/2017-03/2022. Next, we grouped patients based on their available lab values and subjective and clinical findings. We then determined the completeness of the patient-level information and checked the data quality. Finally, we reported the completeness of the objective and clinical findings documented in the TUHS Cosmos system.

2.2. Manual review guidelines

After obtaining the patient cohort, we created a manual review guideline using bottom-up and top-down approaches. In the top-down approach, we conducted an in-depth literature review especially peer-reviewed research articles, focused on the development of EULAR/ACR classification, testing performance of the EULAR/ACR classification, and utilization of EHR data to validate the EULAR/ACR classification system. In the bottom-up approach, we identified the availability of information with the EHR data.

2.3. Three-step filtering approach to determine SLE diagnosis

Next, we developed an automated program to determine the SLE diagnosis with the following three-step filtering approach. First, we determined if the patients' ANA test was recorded or not. If the ANA test was documented, then its values must be examined. If ANA is >=1:80 on Hep-2 cells or an equivalent positive test at least once, then we review further records and lab reports. However, if the ANA value is <1:80 on Hep-2 cells, we consider the SLE diagnosis negative. In the second-round filter, for patients with positive ANA tests, we examined if the patient had a positive Renal Biopsy Class III or IV Lupus Nephritis. If this information is not present, then in step three, we further examined patients' clinical criteria (*e.g.,* fever, Leukopenia, Thrombocytopenia, delirium, psychosis, oral ulcers, etc.) or lab values (*e.g.,* anti-cardiolipin, C3 OR C4 level, Anti-Ds DNA antibody levels, etc.), as described in the EULAR/ACR classification [2]. Weightage was then assigned for each of these clinical findings as well as lab values to categorize patients into one of the three diagnostic criteria: 1) Positive, 2) uncertain, or 3) negative SLE.

3. Results

3.1. Patient cohort and characteristics

There are over 138 million patient records documented in EPIC Cosmos, with 873,521 patients received at least a single care at the TUHS. Among the TUHS patients, 3,596 patients had at least one ICD code related to SLE (M32*). Most of these SLE patients are female (89%) between 50 to 65 years of age and identified as African American.

3.2. Completeness of objective and clinical fundings reported in the EHR

Out of the 3,596 patients with at least one physician-assigned diagnosis of SLE, 475 (13%) had information about their ANA test and only 342 (10%) patients had a positive ANA test result. Out of 342 patients, 100 (29 %) patients reported having fever, 86 (25%) reported had proteinuria, 84 (25%) had leukopenia, and 68 (20%) had thrombocytopenia. Complement C3 lab value was reported in 280 (81%) patients, and Smith extractable nuclear antibody in Serum by Immunoassay was reported in 50 (15%) patients.

3.3. Patients who met EULAR/ACR classification criteria

With our automate program, we classified patients' SLE diagnosis into 1) positive, 2) uncertain, and 3) negative. In this study, out of the 3596 patients with SLE coded, we

were able to confirm a partial positive SLE diagnosis of 342 patients (10%) who had a positive ANA test as well as at least one positive objective or clinical finding. We classified 139 patients (4%) as SLE negative because their ANA lab value was not in the suggested range. We classified the remaining 3,114 patients (86%) as uncertain because we were not able to obtain their complete medical history and lab values.

4. Discussion

This is the first study to examine the quality of the EHR data for SLE research. Our long-term goal is to develop a prediction model to diagnose SLE patients early and improve early treatment to improve patient outcomes. However, as EHR data is not collected for research purposes, the quality of EHR data must be evaluated before its intended use. We found that only 10% of patients met EULAR/ACR classification criteria when only structured categories were utilized in the EHR (EPIC COSMOS). Most patients (86%) were classified as uncertain cases because we lacked access to their complete diagnosis, lab reports, and clinical notes in the structured category of EPIC COSMOS. Patients' rich information, such as patient-reported symptoms, physical examination findings, and medical histories, are typically documented in the clinical notes, such as arthritis, photosensitive rash, ulcers, pericarditis, and nephritis, but we were unable to include information documented in the clinical notes as it would require the development and testing of natural language processing (NLP) methods. Since we found poor completeness of SLE-related information in this study, we will develop NLP pipelines to extract patients' other relevant information from the clinical notes to improve completeness before developing prediction models. We also presented a step-by-step process to phenotype SLE patients using EHR datasets using top-down and bottom-up approaches. These steps satisfy both the expert-driven EULAR/ACR classification criteria and data-driven approach to phenotype using EHR.

One another possible reason for only 10% of patients meeting EULAR/ACR classification could also be that not all physicians follow the EULAR/ACR classification to diagnose SLE patients. It is possible that patients may not demonstrate all clinical symptoms that meet the EULAR/ACR criteria, but clinically Rheumatologists may suspect SLE. Relying solely on these classification systems may lead to a higher frequency of false negative cases, especially for SLE patients who already have a delay in diagnosis. We found that despite only 10% of patients meeting EULAR/ACR classification criteria for SLE, 90% of the patients still had at least one ICD code for SLE present. Therefore, we will add clinical note information using NLP to calculate the number of patients who meet the EULAR/ACR classification criteria. This can be applied to study other immune diseases (Sjogren's Syndrome).

Comparing our study results with other studies [6–8], only one study attempted to identify SLE patients from EHR data. However, our study result is distinct because it utilizes the EPIC COSMOS dataset that includes millions of patients' information for research and uses a comparatively large sample size given the rarity of SLE. Walunas et al. [7] concluded that it might be possible to characterize the spectrum of disease in people with lupus as described through care documented in medical records, which is further confirmed by our study results. Other studies have used machine learning and NLP methods to extract information from the EHR about SLE-related complications

such as renal flares in lupus nephritis [6–8]. However, these studies did not include the early diagnosis component in their prediction modeling.

A major limitation of this study is that we did not examine patients' clinical notes to phenotype SLE patients due to the expectation that we would be able to obtain their complete diagnostic information from the structured format. This is because it requires NLP and manual labeling, and creating gold standard dataset tasks can be time consuming. We also wanted to examine the functionality of new COSMOS EPIC data system about the completeness of the data. In future studies, we will develop NLP pipelines to mine clinical notes of SLE patients. We will then determine early diagnosis patterns to build a prediction model using machine learning models.

5. Conclusions

Only 10% of patients met EULAR/ACR classification criteria when considering only structured data from the EHR. Examination of clinical notes is critical to confirm the diagnosis of the remaining patients. Further studies are warned to also evaluate the effectiveness of the new EULAR/ACR classification in a clinical setting.

References

[1] Mok CC, Lau CS. Pathogenesis of systemic lupus erythematosus. J Clin Pathol. 2003 Jul;56(7):481-90, doi: 10.1136/jcp.56.7.481.

[2] Bernatsky S, Boivin JF, Joseph L, Manzi S, Ginzler E, Gladman DD, Urowitz M, Fortin PR, Petri M, Barr S, Gordon C, Bae SC, Isenberg D, Zoma A, Aranow C, Dooley MA, Nived O, Sturfelt G, Steinsson K, Alarcón G, Senécal JL, Zummer M, Hanly J, Ensworth S, Pope J, Edworthy S, Rahman A, Sibley J, El-Gabalawy H, McCarthy T, St Pierre Y, Clarke A, Ramsey-Goldman R. Mortality in systemic lupus erythematosus. Arthritis Rheum. 2006 Aug;54(8):2550-7, doi: 10.1002/art.21955.

[3] Peña Y, Tse K, Hanrahan LM, de Bruin A, Morand EF, Getz K. Establishing Consensus Understanding of the Barriers to Drug Development in Lupus. Ther Innov Regul Sci. 2020 Sep;54(5):1159-65, doi: 10.1007/s43441-020-00134-2.

[4] Aringer M, Leuchten N, Johnson SR. New Criteria for Lupus. Curr Rheumatol Rep. 2020 May;22(6):18, doi: 10.1007/s11926-020-00896-6.

[5] Aringer M, Costenbader K, Daikh D, Brinks R, Mosca M, Ramsey-Goldman R, Smolen JS, Wofsy D, Boumpas DT, Kamen DL, Jayne D, Cervera R, Costedoat-Chalumeau N, Diamond B, Gladman DD, Hahn B, Hiepe F, Jacobsen S, Khanna D, Lerstrøm K, Massarotti E, McCune J, Ruiz-Irastorza G, Sanchez-Guerrero J, Schneider M, Urowitz M, Bertsias G, Hoyer BF, Leuchten N, Tani C, Tedeschi SK, Touma Z, Schmajuk G, Anic B, Assan F, Chan TM, Clarke AE, Crow MK, Czirják L, Doria A, Graninger W, Halda-Kiss B, Hasni S, Izmirly PM, Jung M, Kumánovics G, Mariette X, Padjen I, Pego-Reigosa JM, Romero-Diaz J, Rúa-Figueroa Fernández Í, Seror R, Stummvoll GH, Tanaka Y, Tektonidou MG, Vasconcelos C, Vital EM, Wallace DJ, Yavuz S, Meroni PL, Fritzler MJ, Naden R, Dörner T, Johnson SR. 2019 European League Against Rheumatism/American College of Rheumatology Classification Criteria for Systemic Lupus Erythematosus. Arthritis Rheumatol. 2019 Sep;71(9):1400-12, doi: 10.1002/art.40930.

[6] Deng Y, Pacheco JA, Chung A, Mao C, Smith JC, Zhao J, Wei WQ, Barnado A, Weng C, Liu C, Cordon A. Natural language processing to identify lupus nephritis phenotype in electronic health records. 2021 Dec;1, doi:10.48550/arxiv.2112.10821.

[7] Walunas TL, Ghosh AS, Pacheco JA, Mitrovic V, Wu A, Jackson KL, Schusler R, Chung A, Erickson D, Mancera-Cuevas K, Luo Y, Kho AN, Ramsey-Goldman R. Evaluation of structured data from electronic health records to identify clinical classification criteria attributes for systemic lupus erythematosus. Lupus Sci Med. 2021 Apr;8(1):e000488, doi: 10.1136/lupus-2021-000488.

[8] Chen Y, Huang S, Chen T, Liang D, Yang J, Zeng C, Li X, Xie G, Liu Z. Machine Learning for Prediction and Risk Stratification of Lupus Nephritis Renal Flare. Am J Nephrol. 2021;52(2):152-60, doi: 10.1159/000513566.