Comparison of an Administrative Algorithm for SLE Disease Severity to Clinical SLE Disease Activity Index Scores

Cameron B. Speyer[1], **Daniel Li**[1], **Hongshu Guan**[1], **Kazuki Yoshida**[1], **Emma Stevens**[1], **April M. Jorge**[2], **Brendan M. Everett**[1], **Candace H. Feldman**[1], **Karen H. Costenbader**[3]

[1]Section of Clinical Sciences, Division of Rheumatology, Inflammation and Immunity, Brigham and Women's Hospital, Hale Building, 6th floor, 60 Fenwood Road, Boston, MA, 02115, USA.

[2]Arthritis Unit, Massachusetts General Hospital, Fruit Street, Boston, MA, 02114, USA.

[3]Section of Clinical Sciences, Division of Rheumatology, Inflammation and Immunity, Brigham and Women's Hospital, Hale Building, 6th floor, 60 Fenwood Road, Boston, MA, 02115, USA.

Abstract

Objective: Systemic lupus erythematosus (SLE) severity, reflecting both disease intensity and duration, is heterogeneous making it challenging to study in administrative databases where severity may confound or mediate associations with outcomes. Garris *et al* developed an administrative claims-based algorithm employing claims over a 1-year period to classify SLE severity as mild, moderate or severe. We sought to compare this administrative algorithm to a measure of SLE activity, the SLE Disease Activity Index-2000 (SLEDAI-2K) score at clinical visits.

Methods: We identified 100 SLE patients followed in the Brigham and Women's Hospital (BWH) Lupus Center (in 2008–2010) with SLEDAI-2K scores at each visit over a 1-year period per person. We obtained data for the Garris algorithm for the same year per subject. We compared

Garris SLE severity to the highest SLEDAI-2K in that year, with SLEDAI-2K categories of mild <3, moderate 3–6, and severe >6. We compared classification using weighted kappa statistics, and positive and negative predictive values (PPV, NPV). We also assessed the binary comparison of mild vs. moderate/severe. We calculated sensitivity, specificity, and McNemar's test.

Results: We analyzed 377 SLEDAI-2K assessments (mean 3.8 [SD 2.6] per subject/year). For classifying moderate/severe vs. mild SLE severity, the sensitivity was 85.7%, specificity 67.6%, PPV 81.8% and NPV 73.5%.

Conclusion: The Garris algorithm for classifying SLE severity in administrative datasets had moderate agreement for classification of mild vs. moderate/severe SLE activity assessed by SLEDAI-2K assessments in an academic lupus center. It may be a useful tool for classifying SLE severity in administrative database studies.

Keywords

Introduction

Systemic lupus erythematosus (SLE) is a multi-system inflammatory autoimmune disease with diverse clinical manifestations. Affecting multiple organ systems, it has heterogeneous severity; some patients experience only rashes and arthralgias, others experience severe multi-organ involvement with nephritis or vasculitis. The concept of SLE severity reflects both the intensity and duration of ongoing disease activity. The heterogeneity of SLE severity presents a challenge for the study of SLE in administrative databases as severity may confound, mediate or be an effect modifier of associations with outcomes and analyses ideally should stratify by or account for this severity. In the clinical setting, indices for SLE activity incorporate symptoms and laboratory results to estimate SLE activity [1]. However, as medical notes and laboratory results are not available in claims data, assessing variation in SLE severity is difficult.

Garris *et al* developed an algorithm to classify lupus patients according to their SLE severity using administrative claims data [2]. Developed to predict healthcare utilization over a one-year period using claims data, it employs International Classification of Diseases (ICD-9), Current Procedural Terminology (CPT) and National Drug Code (NDC) codes to assess disease severity over a one-year period. It was derived from the items in the Systemic Lupus Erythematosus Disease Activity Index (SLEDAI [3]), Systemic Lupus Activity Measure (SLAM[4]) and British Isles Lupus Assessment Group index (BILAG [5]) assessments and classifies patients as having mild, moderate or severe disease, based on their most severe lupus manifestation over the course of the year [2, 3]. Since the concept of severity encompasses disease activity over time, we hypothesized that the Garris administrative claims algorithm for severity should exhibit a high degree of association with an existing disease activity measure aggregated over the same time interval. Thus, we sought to compare classification by this administrative algorithm to the highest SLEDAI-2K score [3] from clinical visits over the same one-year period in a longitudinal cohort at an academic lupus center.

Methods

Patient Population and Data Collection

We randomly selected a sub-sample of 100 SLE patients followed in the Lupus Center at Brigham and Women's Hospital (BWH), a large urban academic hospital, during 2008–2010. The BWH Lupus Center sees more than 800 individual SLE patients per year. The patients identified had an average of 3.8 visits per year. We collected SLEDAI-2K assessments from previous research studies for all visits to our clinic over the course of one year per person, including assessments from hospitalizations. The SLEDAI-2K assessment assigns disease severity on a continuous scale and has also been validated as a categorical scale; a SLEDAI-2K score <3 indicates mild disease, 3–6 indicates moderate disease, and any SLEDAI-2K score >6 indicates severe disease[3, 6].

We then obtained the diagnosis, medication and procedural codes included in the Garris algorithm from the Partners HealthCare System (PHS) Research Patient Data Repository (RPDR)[7], a centralized longitudinal electronic medical records (EMR) database containing billing claims and health record data from all institutions within PHS including BWH, Massachusetts General Hospital, and affiliate community hospitals. The Garris algorithm employs ICD-9, CPT and NDC codes for conditions and medications associated with SLE manifestations, such as psychosis and end stage renal disease for severe disease, hemolytic anemia and pericarditis for moderate disease, and medications including immunosuppressants and corticosteroids. Garris algorithm codes were collected for the same year per person as the SLEDAI-2K assessments.

Statistical Analyses

We compared the Garris severity classification for each person to their highest SLEDAI-2K score from the same year. We compared the Garris algorithm categories of mild, moderate and severe to the SLEDAI-2K categories of mild, moderate and severe, and assessed the agreement of the indices in a 3×3 table and using weighted kappa statistics. A weighted kappa accounts for the difference between 1-level vs. 2-level disagreement, and gives less weight to 1-level disagreements[8]. For example, a patient classified as mild by SLEDAI-2K and moderate by Garris algorithm would constitute a 1-level disagreement, and if that patient was classified as severe by the algorithm, it would be a 2-level disagreement. If the algorithm classified this patient as severe this would be a 2-level disagreement. The original SLEDAI-2K was considered the gold standard.

We then assessed a dichotomous comparison of mild vs. moderate/severe SLE. It is clinically relevant and useful to pull out patients with very mild disease, as treatment and outcomes differ significantly from patients with more severe disease. The definition of moderate disease activity is more variable, but together moderate and severe disease represent clinically relevant disease activity. We assessed the performance of the Garris algorithm, calculating its sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV) comparing overall classification to SLEDAI-2K. In this context, sensitivity was defined as the number of "positives" identified by the Garris algorithm (i.e. moderate/severe disease) divided by the total number of true positives as defined by the

SLEDAI-2K. Specificity was defined as the number of "negatives" (i.e. mild disease) identified by the Garris algorithm divided by the total number of true negatives by SLEDAI-2K. A McNemar's Chi-square test to detect disagreement in the binary categorization of moderate/severe vs. mild was also performed. Institutional review board approval was obtained from Partners' Healthcare for all aspects of this study.

Results

Of the 100 patients with prevalent SLE, the mean (SD) patient age was 46 (13) years, 95% were female, and 77% were white. Mean (SD) duration of SLE disease was 15.6 (11.4) years and mean (SD) number of clinical visits per year was 3.8 (2.6) (Table 1). Of the 100 subjects studied, 15 met classification for highest SLE severity, another 51 were classified by the Garris algorithm as moderate severity, and the rest were by default classified as having mild severity. Of the 15 classified as having highest severity disease, 10 (66%) met this classification based on their conditions, the most common being end-stage renal disease (3 cases, 7%). Of those meeting classification as moderate severity, 21 (41%) met this classification based on conditions (the rest based on medications received). The most common moderate severity conditions were nephritis (7 cases, 14%), ischemic necrosis of bone (5 cases, 10%), and pleurisy/pleural effusion (3 cases, 6%).

Three Category Comparison of Mild vs. Moderate vs. Severe SLE

When comparing the Garris algorithm categories of mild, moderate and severe to the SLEDAI-2K categories of mild, moderate and severe, 60/100 patients were correctly cross-classified. Using the SLEDAI-2K as the gold standard, the algorithm correctly classified 67.6% of mild patients (25/37), 70.6% of moderate patients (24/34), and 37.9% of severe patients (11/29) (Table 2). In total, the algorithm classified 37% of patients in a 1-level disagreement from the SLEDAI-2K and 3% of patients in a 2-level disagreement. The weighted kappa was 0.47 (95% confidence interval 0.34–0.61), suggesting moderate agreement between the two indices[9]. The mean (SD) SLEDAI-2K among patients classified as mild by the algorithm was 2.0 (2.2), moderate was 6.0 (3.9), and severe was 9.4 (5.2).

Dichotomized comparison: Mild vs. Moderate/Severe SLE

We also assessed the performance of the algorithm in a binary comparison: mild vs. moderate/severe. When assessing the mild vs. moderate/severe comparison, the algorithm correctly classified 79% of total patients (Table 3). Twenty-five out of 37 total mild patients were correctly classified, giving a sensitivity of 85.7% to the algorithm in this setting. Additionally, 54/63 total moderate/severe patients were correctly classified by the algorithm, providing high specificity (67.6%) (Table 3). The PPV was 81.8% for moderate/severe vs. mild SLE and the NPV was 73.5% for this binary comparison. The McNemar's test for difference was not significant (Chi-square 0.43, p= 0.52).

Discussion

The heterogeneous nature of SLE severity complicates the ability to study and appropriately adjust for its influence when conducting research using administrative databases. SLE disease activity may influence observed SLE outcomes, however it is challenging to measure using claims data as clinical assessments and laboratories are not available. Although the constructs of SLE severity and SLE activity are distinct, they are related as SLE severity reflects both duration and activity of disease. We thus sought to compare a previously-developed algorithm intended to stratify SLE patients by their disease severity, based on claims for both organ-system involvement and specific medications in administrative data, to classification by clinical SLE disease activity assessments.

The algorithm performed moderately well in the binary comparison; the 85.7% sensitivity, 67.6% specificity and PPV of 81.8% in this sample of patients from the BWH Lupus Center indicate it is a relatively well-discriminating model for classifying patients based on their SLE severity over one year[10]. The ability to discriminate between mild and moderate/severe disease is clinically relevant; patients with moderate/severe disease have more organ involvement, comorbidities, and mortality compared to those with mild disease, warranting more intensive treatment and medications[6]. Our assessment of the Garris algorithm for SLE severity in administrative data included 100 well-characterized SLE patients, closely followed at BWH with an average 3.8 annual visits and a range of SLE severity. Each subject had SLEDAI-2K severity assessments completed for every clinical visit over a one-year period. Although the algorithm cannot distinguish between sustained high disease activity and shorter-term flares, it would be expected to capture severe organ system involvement.

There are several limitations to this study. It is possible that claims from outside of our healthcare network were not captured, however, as all patients had a BWH rheumatologist and several visits per year, it is likely we captured most all SLE claims and medications. Additionally, SLEDAI-2K assessments are intermittent and capture SLE disease activity over a shorter duration of 10 days, whereas the Garris algorithm captures all codes and claims for a 1-year period. Several of the misclassifications we found resulted from ICD-9 codes propagated from prior years. For example, patients receive ESRD billing codes post-kidney transplant, when SLE may be quiescent. We also found ICD-9 codes can be nonspecific; for example, codes for "unspecified arteritis" or "aortitis" capture peripheral vasculitis. Additionally several of this algorithm's items, developed by another group, may not be supported by current clinical practice or are not included or differently weighted by the SLEDAI-2K [2]. For example, seizure is a moderate SLE severity condition in the algorithm, but heavily weighted by the SLEDAI-2K when attributable to SLE activity. Myocarditis and hepatitis are not included in the SLEDAI-2K. We found that several conditions were not represented in this population and much of the administrative algorithm classification was based on medication use (59% of those classified as moderate severity). The relationship between the two instruments obviously would be much stronger if medications were considered in addition to the SLEDAI-2K. Furthermore, SLEDAI-2K items such as leukopenia and positive anti-dsDNA are rarely independent diagnoses and thus were not included in the algorithm.

Administrative billing codes also cannot differentiate between conditions due to SLE or other causes. Despite these limitations, we found the algorithm to have moderately acceptable performance characteristics for stratification of mild vs. more severe SLE patients in administrative datasets.

SLE severity heterogeneity complicates administrative studies. Although the Garris administrative algorithm goes beyond the SLEDAI-2K activity measure by including SLE medication data and irreversible organ damage items, it has acceptable sensitivity and PPV for discriminating mild vs. moderate/severe SLE compared to the SLEDAI-2K, and overall good specificity and PPV. This algorithm may be useful for assessing SLE severity in administrative studies.

Acknowledgments:

References

1. Petri M, Hellmann D, Hochberg M (1992) Validity and reliability of lupus activity measures in the routine clinic setting. J Rheumatol 19:53–9 [PubMed: 1556700]

2. Garris C, Jhingran P, Bass D, Engel-Nitz NM, Riedel A, Dennis G (2013) Healthcare utilization and cost of systemic lupus erythematosus in a US managed care health plan. J Med Econ 16:667–77 [PubMed: 23425294]

3. Gladman DD, Ibanez D, Urowitz MB (2002) Systemic lupus erythematosus disease activity index 2000. J Rheumatol 29:288–91 [PubMed: 11838846]

4. Liang MH, Socher SA, Larson MG, Schur PH (1989) Reliability and validity of six systems for the clinical assessment of disease activity in systemic lupus erythematosus. Arthritis Rheum 32:1107–18 [PubMed: 2775320]

5. Isenberg DA, Rahman A, Allen E, Farewell V, Akil M, Bruce IN, et al. (2005) BILAG 2004. Development and initial validation of an updated version of the British Isles Lupus Assessment Group's disease activity index for patients with systemic lupus erythematosus. Rheumatology (Oxford) 44:902–6 [PubMed: 15814577]

6. Polachek A, Gladman DD, Su J, Urowitz MB (2017) Defining Low Disease Activity in Systemic Lupus Erythematosus. Arthritis Care Res (Hoboken) 69:997–1003 [PubMed: 27696791]

7. Nalichowski R, Keogh D, Chueh HC, Murphy SN (2006) Calculating the benefits of a Research Patient Data Repository. AMIA Annu Symp Proc:1044 [PubMed: 17238663]

8. Cohen J (1968) Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. Psychol Bull 70:213–20 [PubMed: 19673146]

9. McHugh ML (2012) Interrater reliability: the kappa statistic. Biochem Med (Zagreb) 22:276–82 [PubMed: 23092060]

10. Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, et al. (2010) Assessing the performance of prediction models: a framework for traditional and novel measures. Epidemiology 21:128–38 [PubMed: 20010215]

Table 1.

Baseline Characteristics of 100 Subjects in the BWH Lupus Cohort 2008–2010

Age, mean (SD)	46 (13)
Female, n (%)	95 (95)
Race, n (%)	
White	77 (77)
Black	10 (10)
Asian	5 (5)
Hispanic/Latino	2 (2)
Other/Unknown	6 (6)
Duration of SLE disease, mean (SD)	15.6 (11.4)
Number of visits per year, mean (SD)	3.8 (2.6)
Number of ACR criteria, mean (SD)	5.1 (1.6)

Table 2.

Cross-Classification of SLEDAI-2K and Garris Algorithm for SLE Severity over one year among 100 Subjects in BWH Lupus Cohort 2008–2010

Mild vs. Moderate vs. Severe				
Garris category (n)	**Highest SLEDAI-2K category (n)**			
Frequency	**Mild**	**Moderate**	**Severe**	**Total**
Mild	**25**	8	1	34
Moderate	10	**24**	17	51
Severe	2	2	**11**	15
Total	**37**	**34**	**29**	100

Table 3.

Cross-Classification and Performance Characteristics for Garris Algorithm for SLE Severity vs. Highest SLEDAI-2K among 100 Patients in BWH Lupus Cohort 2008–2010

Mild vs. Moderate/Severe			
Garris category (n)	**Highest SLEDAI-2K category (n)**		
Frequency	**Moderate/ Severe**	**Mild**	**Total**
Mild	9	25	34
Moderate/Severe	**54**	**12**	66
Total	**63**	**37**	100
Statistics for moderate/severe vs. mild classification			
Sensitivity	85.70%		
Specificity	67.60%		
PPV	81.80%		
NPV	73.50%		
McNemar's Chi-squared test for difference	0.43 (p= 0.52)		