



Published in final edited form as:

Arthritis Rheumatol. 2023 September ; 75(9): 1532–1541. doi:10.1002/art.42544.

Phenotype Risk Score but not Genetic Risk Score aids in identifying individuals with Systemic Lupus Erythematosus in the Electronic Health Record

April Barnado, MD, MSCI^{1,2}, Lee Wheless, MD, PhD³, Alex Camai, BA¹, Sarah Green, BA¹, Bryan Han¹, Anish Katta, BA¹, Joshua C. Denny, MD, MS⁴, Amr H. Sawalha, MD⁵

¹Division of Rheumatology & Immunology, Department of Medicine, Vanderbilt University Medical Center, Nashville, TN

²Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN

³Department of Dermatology, Division of Epidemiology, Vanderbilt University Medical Center, Nashville, TN

⁴All of Us Research Program, National Institutes of Health, Bethesda, MD

⁵Departments of Pediatrics, Medicine, and Immunology & Lupus Center of Excellence, University of Pittsburgh School of Medicine, Pittsburgh, PA

Abstract

Objective: Systemic lupus erythematosus (SLE) poses diagnostic challenges. This study aimed to evaluate the utility of a phenotype risk score (PheRS) and a genetic risk score (GRS) to identify SLE individuals in a real-world setting.

Methods: Using a de-identified electronic health record (EHR) with an associated DNA biobank, we identified 789 SLE cases and 2,261 controls with available MEGA^{EX} genotyping. A PheRS for SLE was developed using billing codes that captured ACR SLE criteria. We developed a GRS with 58 SLE risk SNPs.

Results: SLE cases had a significantly higher PheRS (7.7 ± 8.0 vs. 0.8 ± 2.0 , $p < 0.001$) and GRS (12.6 ± 2.3 vs. 11.0 ± 2.0 , $p < 0.001$) compared to controls. Black SLE individuals had a higher PheRS vs. White individuals (10.0 ± 10.1 vs. 7.1 ± 7.2 , $p = 0.002$) but a lower GRS (9.0 ± 1.4 , 12.3 ± 1.7 , $p < 0.001$). Models predicting SLE including PheRS had the highest AUC of 0.89. Adding GRS to PheRS did not result in a higher AUC. On chart review, controls with the highest PheRS and GRS had undiagnosed SLE.

Conclusion: We developed a SLE PheRS to identify established and undiagnosed SLE individuals. A SLE GRS using known risk SNPs did not add value beyond the PheRS and was of limited utility in Black SLE individuals. More work is needed to understand the genetic risks of SLE in diverse populations.

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

Introduction

Systemic lupus erythematosus (SLE) poses a diagnostic challenge to clinicians. Patients have diverse presentations (1), and SLE symptoms can mimic other diseases (2). SLE patients can have delays in diagnosis based on atypical or incremental disease presentation (3). One study demonstrated that it took on average 7 years from symptom onset for SLE diagnosis (4). Delays in diagnosis lead to delays in treatment that result in increased SLE disease damage and associated increased morbidity and mortality (5, 6).

We sought to use a SLE phenotype risk score (PheRS) and a genetic risk score (GRS) to identify SLE individuals in the electronic health record (EHR) and potentially find individuals with undiagnosed SLE. A PheRS measures the degree to which an individual's symptoms, as assessed by billing codes, overlap with defined disease criteria. Using billing codes in the EHR, PheRS have identified individuals with unrecognized Mendelian genetic disorders successfully and may help identify individuals earlier with rare diseases (7–9). To the best of our knowledge, PheRS with EHR billing codes have not been used in autoimmune diseases or SLE. We sought to build a PheRS with billing codes that capture SLE disease criteria.

Several studies examined genetic risk scores (GRS) in SLE with varying success to assess an individual's risk of developing SLE (10–17). In some but not all studies, male (18), pediatric-onset (10–12, 17, 19), and SLE nephritis (11, 15–17) individuals all had higher GRS. These studies did not fully incorporate clinical data with GRS to determine if genetic data contributes beyond clinical data. We evaluated if the PheRS and GRS could identify individuals with SLE in the EHR as well as identify individuals with undiagnosed SLE. We also examined the association between the GRS and PheRS to determine if genetic data adds value beyond clinical data.

Methods

Synthetic Derivative and BioVU

After receiving IRB approval from VUMC, we used a large de-identified mirror image of the EHR called the Synthetic Derivative. The Synthetic Derivative contains over 3.2 million subjects with longitudinal clinical data since the 1990s (20). The Synthetic Derivative contains all clinical data including inpatient and outpatient notes for both primary care and subspecialty care. Billing codes (ICD-9 and ICD-10-CM), labs, medications, radiology, and pathology data are also available. Outside records are not available. Clinical data from the Synthetic Derivative is linked to a large genetic biobank called BioVU (20). BioVU allows for systematic collection across all individuals with representations of race and ethnicity that closely align with the demographics of individuals who seek care at VUMC in middle TN. BioVU accrues DNA samples from subjects using remaining blood obtained from routine clinical testing that would otherwise be discarded. This sample collection launched in 2007. As of March 2023, there are over 310,000 genotyped samples. Subjects who are already genotyped have dense EHR data with a mean follow-up of 5.7 years with an individual subject on average having 596 labs, 601 medications, and 132 clinic notes.

Genotyping

Genotyping was performed at the Vanderbilt Technologies for Advanced Genomics (VANTAGE) Core using the MEGA^{EX} chip. The MEGA^{EX} chip contains over 2 million SNPs and covers 65.7% of GWAS catalog SNPs. SLE risk alleles were selected based on replicated SLE GWAS findings. We focused on risk alleles in European ancestry cohorts knowing our population is predominantly White, but also because SLE GWAS in non-White populations are limited (21). SNPs were required to have independent associations through literature review and were not in linkage disequilibrium. Of the 67 candidate SNPs, we assembled 58 SNPs with a 95% sampling and 95% call rate (Supplemental Table 1).

Identifying SLE cases and controls

Within the Synthetic Derivative, we identified potential SLE cases using a previously validated algorithm that requires 4 counts of the SLE International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9) code of 710.0 and antinuclear antibody (ANA) positive titer of 1:160, while excluding dermatomyositis (DM) and systemic sclerosis (SSc) ICD-9 codes (710.3, and 710.1, respectively) (22). This algorithm has a positive predictive value of 90% and a sensitivity of 86% (22). We performed chart review on all potential SLE cases to ensure they were diagnosed with SLE by a Vanderbilt or external specialist (rheumatologist, nephrologist, or dermatologist). We have previously described this SLE EHR cohort (22–27). We then selected which SLE cases had available MEGA^{EX} chip data (Figure 1).

Controls were identified in the Synthetic Derivative and did not have ICD-9 codes under the 710.* heading “Diffuse diseases of connective tissue,” the 714.* heading “Rheumatoid arthritis and other inflammatory polyarthropathies,” or ICD-10 codes M05.* (“Rheumatoid arthritis with rheumatoid factor”), M06.* (“Other rheumatoid arthritis”), M32 (“SLE”), M33.* (Dermatopolymyositis), M34.* (“Systemic sclerosis”), M35.* (“Other systemic involvement of connective tissue”), and M36.* (“Systemic disorders of connective tissue in diseases classified elsewhere”). Controls were frequency-matched to SLE cases in a 5:1 ratio by age (± 5 years), race, and sex to maximize power while allowing close matching. Control subjects were “medical home patients” who received longitudinal care at VUMC (28) with at least 3 outpatient visits within 5 years to ensure density of records was similar to that of cases. These controls have previously been described (23, 24, 28). We then selected controls with available MEGA^{EX} chip data.

Development of phenotype risk score

We developed the phenotype risk score (PheRS) by identifying billing codes (both ICD-9 and ICD-10-CM) and corresponding PheWAS codes that capture the 1997 ACR SLE criteria (24, 29), as most cases were diagnosed with SLE prior to 2019. A list of both ICD-9 and ICD-10-CM codes that are mapped to PheWAS codes is available at <http://phewascatalog.org>. PheWAS and their associated billing codes that correspond to SLE disease criteria (24) were used in the phenotype risk score (Supplemental Table 2). The SLE PheRS was the sum of these codes in a given individual weighted by the log inverse prevalence of that code in the entire EHR. As is typical for a PheRS, billing codes with a direction mention of SLE (i.e. 710.0 and M32*) were excluded from the PheRS, as our goal

was to identify individuals not diagnosed with SLE. Scores were calculated for both SLE cases and controls.

Development of genetic risk score

We developed the genetic risk score (GRS) by reviewing SLE GWAS studies and focusing on SNPs with replication in predominantly European ancestry cohorts, as our population is predominantly White. We then ensured SNPs of interest had independent effects and were not in linkage disequilibrium (Supplemental Table 1). To calculate the GRS, we weighted each SNP by the inverse log of the effect size (i.e. odds ratio) reported in the literature and then summed all the SNPs for a total score (18, 30). Scores were calculated for both SLE cases and controls.

Chart Review

We conducted chart review on the 50 controls with the highest PheRS and the 50 controls with the highest GRS. We assessed both inpatient and outpatient notes and labs for 1997 updated ACR and SLICC SLE criteria (29, 31). Race and ethnicity were based on both self-report and administrative data. Prior studies have validated that these EHR race assignments reflect genetic ancestry (32). Age was defined as current age at time of analysis. We performed chart review to assess for presence of SLE nephritis, defined by a renal biopsy or diagnosis by a rheumatologist or nephrologist. To estimate age at SLE diagnosis, we performed chart review to examine for date of SLE diagnosis documented in a rheumatologist's note. If this date was not documented, we used date of first SLE billing code.

Statistical Analysis

We compared categorical variables using Chi-square or Fisher's exact test and compared continuous variables using the Mann-Whitney U test, as there were non-normal distributions in the data. We performed logistic regression to estimate the association of SLE case status with both PheRS and GRS adjusting for age, sex, and race. We also performed Pearson correlation and linear regression to estimate the association of PheRS with GRS adjusting for age, sex, and race as well as to estimate the association of age of SLE diagnosis with PheRS and GRS. Two-sided p values less than 0.05 were considered significant. Analyses were conducted using R version 4.0.2.

Results

PheRS

Demographics for SLE cases (n = 789) and controls (n = 2,261) are shown in Table 1. There were no significant differences in age, sex, or ethnicity. SLE cases were significantly more likely to be White compared to controls (p < 0.001). Initially, controls were race matched to SLE cases but then only SLE cases and controls were selected who had available genetic data. In BioVU, 79% of subjects are White and 10% Black. SLE cases had significantly higher PheRS compared to controls (7.7 ± 8.0 vs. 0.8 ± 2.0 , p < 0.001) (Figure 2A). Female and male individuals with SLE had similar PheRS (7.7 ± 8.0 vs. 7.3 ± 7.1 , p = 0.61). Compared to White individuals with SLE, Black individuals had higher PheRS (10.0 ± 10.1

vs. 7.1 ± 7.2 , $p = 0.002$). In a logistic model, PheRS was significantly associated with SLE case status after adjusting for age, sex, and race ($p < 0.001$).

Chart review of controls with highest PheRS

We chart reviewed the 50 controls with the highest PheRS (Figure 3). The control with the highest PheRS was a 38-year-old, Black female individual who had a positive ANA and dsDNA and a renal biopsy consistent with class V SLE nephritis with the pathologist noting need for clinical correlation. The individual followed with nephrology who felt the individual did not fit enough 1997 ACR SLE criteria (29), as she didn't have extrarenal SLE manifestations. She later developed end stage renal disease (ESRD). The individual fulfilled both SLICC (31) and 2019 EULAR/ACR SLE criteria (33).

The control with the second highest PheRS was a 55-year-old, Black female individual who presented in the inpatient setting with transverse myelitis and found to have a positive ANA and dsDNA. She also had a history of ESRD and pancytopenia, both with causes unidentified. The individual was managed by neurology and never saw rheumatology. She was treated with corticosteroids, intravenous immunoglobulin, plasma exchange, and cyclophosphamide but was never formally diagnosed as SLE. She fulfilled both SLICC (31) and 2019 EULAR/ACR SLE criteria (33). In addition to the 2 "controls" discussed above, there were an additional 4 individuals who had incomplete lupus or fulfilled 3 ACR SLE (29) or SLICC ACR criteria (31). These individuals had a positive ANA, malar rash, and inflammatory arthritis. There were also an additional 8 individuals that had other autoimmune disease including discoid lupus, Crohn's disease, and seronegative spondyloarthropathy.

Of the 50 controls with the highest PheRS, the lowest PheRS among the top 50 controls was 10.39. This corresponds to an 81% percentile among SLE cases (Supplemental Figure 1). For these same 50 controls with the highest PheRS, there was a mean GRS of 11.07. This GRS corresponds to a 30% percentile among SLE cases.

Genotype risk score (GRS)

We compared SLE genotype risk scores (GRS) in SLE cases ($n = 789$) vs. controls ($n = 2261$) (Figure 2B). Compared to controls, SLE cases had a significantly higher GRS (12.16 ± 2.25 vs. 11.02 ± 2.04 , $p < 0.001$). Compared to female SLE cases, male SLE cases had a higher GRS (11.58 ± 2.21 vs. 11.30 ± 2.15 , $p = 0.03$). White SLE cases had a significantly higher GRS compared to Black SLE cases (12.31 ± 1.67 vs. 9.01 ± 1.36 , $p < 0.001$). We also calculated an unweighted GRS by counting the cumulative number of risk alleles for each SNP in each individual. SLE cases had a significantly higher allele risk count compared to controls (32.95 ± 5.12 vs. 30.46 ± 4.85 , $p < 0.001$).

Chart review of controls with highest GRS

Similar to the PheRS analyses, we conducted chart review on the 50 controls with the highest GRS to determine if SLE or other autoimmune diseases may have been undiagnosed or misdiagnosed (Figure 3). Of the 50 controls, 5 had incomplete SLE with features including positive ANA, joint pain, and serositis with 2 of these individuals having seen

rheumatology. Another 10 controls had other autoimmune diseases including rheumatoid arthritis, psoriatic arthritis, type 1 diabetes, Crohn's disease, autoimmune hepatitis, primary biliary cirrhosis, and antiphospholipid antibody syndrome. While individuals with codes for rheumatic autoimmune diseases were removed from the controls, individuals with codes for all autoimmune diseases (i.e. type 1 diabetes, Crohn's) were not removed. The three control individuals with rheumatic autoimmune diseases (psoriatic arthritis, rheumatoid arthritis) had atypical presentations and also loss of follow-up after diagnosis and thus did not have associated billing codes for these conditions.

Of the 50 controls with the highest GRS, the lowest GRS among the top 50 controls was 15.02. This corresponds to a 91% percentile among SLE cases (Supplemental Figure 2). For these same 50 controls, there was a mean PheRS of 2.58, corresponding to a 25% percentile among SLE cases.

GRS and association with clinical variables

We compared the GRS in cases with SLE nephritis ($n = 147$) to cases without SLE nephritis ($n = 640$). Cases without SLE nephritis had a higher GRS compared to cases with nephritis (12.3 ± 2.2 vs. 11.5 ± 2.6 , $p < 0.001$). This result remained the same when we restricted the analyses to SLE nephritis individuals with diagnoses confirmed on renal biopsies ($n = 119$). Stratifying results by race demonstrated that White cases with nephritis ($n = 76$) had significantly higher GRS compared to Black cases with nephritis ($n = 55$) (13.3 ± 1.8 vs. 9.1 ± 1.6 , $p < 0.01$). White cases with SLE nephritis had a higher GRS compared to White cases without SLE nephritis (13.3 ± 1.8 vs. 12.8 ± 1.8 , $p = 0.05$). There was no difference in GRS in Black cases with vs. without nephritis (9.1 ± 1.6 vs. 9.1 ± 1.6 , $p = 0.92$).

For age of SLE diagnosis, Black cases were diagnosed at a significantly younger age than White cases (33 ± 16 vs. 42 ± 15 , $p < 0.001$). We observed an adjusted R^2 of 0.18 with age of SLE diagnosis and GRS in all SLE cases ($p < 0.001$), an adjusted R^2 of 0.001 in White cases ($p = 0.58$), and an adjusted R^2 of 0 in Black cases ($p = 0.77$) (Figure 4). In a linear regression model for GRS, age of SLE diagnosis was not significant after adjusting for sex and race ($p = 0.99$).

PheRS and association with clinical variables

We performed similar analyses for age of diagnosis and PheRS. We observed an adjusted R^2 of 0.03 in all SLE cases ($p < 0.001$), an adjusted R^2 of 0.01 in White cases ($p = 0.02$), and an adjusted R^2 of 0.06 in Black cases ($p = 0.004$) (Figure 4). Age at SLE diagnosis ($p < 0.001$), race ($p < 0.001$), and the interaction between race and age at SLE diagnosis ($p = 0.01$) were all significantly associated with PheRS after adjusting for sex ($p = 0.30$).

Models with PheRS and GRS

We then performed analyses that incorporated both the SLE GRS and PheRS. We examined for the association of individual SNPs with the PheRS. There was only one SNP (rs2476601, PTPN22) with a significant association with the PheRS ($p = 0.03$). In SLE cases, there was no correlation between the PheRS and GRS ($R^2 = 0.03$). Using linear regression, we examined the association of PheRS with GRS adjusting for age, sex, and race. GRS was

not significantly associated with the PheRS ($p = 0.52$) after adjusting for age and sex. Both Black race ($p < 0.01$) and younger age ($p < 0.01$) were associated with higher PheRS after adjusting for GRS and sex.

In a logistic regression model for SLE case status, both the PheRS (OR = 1.05, 95% CI 1.04–1.05, $p < 0.01$) and GRS (OR = 1.04, 95% CI 1.03–1.04, $p < 0.01$) were significantly associated with SLE case status after adjusting for age and sex. As we observed different distributions of PheRS and GRS by race, we conducted stratified analyses by race. In a model with White subjects, both the PheRS (OR = 1.05, 95% CI 1.04–1.05, $p < 0.01$) and GRS (OR = 1.04, 95% CI 1.03–1.05, $p < 0.01$) were significantly associated with SLE case status after adjusting for sex and age. In a model with Black subjects, the PheRS (OR = 1.04, 95% CI 1.04–1.05, $p < 0.01$) but not the GRS ($p = 0.57$) was associated with SLE case status after adjusting for sex and age.

We compared the performance of models for SLE case status in all individuals (overall) and in stratified analyses with White and Black individuals (Figure 5). AUCs for overall and stratified models with PheRS, GRS, and both PheRS and GRS are shown in Supplemental Table 3. Models with PheRS + GRS had the highest AUCs (overall AUC: 0.89, 95% CI 0.87–0.90) while models with only the GRS had the lowest AUC (overall AUC: 0.65, 95% CI 0.63–0.67).

Discussion

We developed a PheRS that identified individuals with SLE in the EHR and even identified individuals with SLE clinical criteria who were not formally diagnosed with SLE. The SLE PheRS is important because it could be deployed within the EHR to assemble individuals with established diagnoses of SLE and to identify individuals who have concerning features for SLE that are misdiagnosed or undiagnosed. To the best of our knowledge, PheRS have been developed for Mendelian genetic disorders (7–9) but not for SLE or other systemic autoimmune diseases. We also developed a SLE GRS using available and validated GWAS findings that also identified individuals with SLE in the EHR. The SLE GRS did not add value in our models beyond clinical data from the SLE PheRS.

Our SLE PheRS, which only uses billing codes, performed well in distinguishing SLE cases from controls with a robust AUC of 0.87 in the overall model. The controls with the highest scores were undiagnosed SLE cases, while other controls with high PheRS had other systemic autoimmune diseases. The SLE PheRS was higher in Black vs. White SLE cases with a higher AUC for the PheRS model in Black vs. White SLE cases. Multiple studies demonstrate a more severe SLE disease course in Black SLE individuals with higher rates of SLE nephritis and SLE mortality (35, 36). Within our own SLE EHR cohort, we have demonstrated this racial health disparity (24).

Identifying individuals with autoimmune diseases, including SLE, can be challenging due to SLE disease heterogeneity. Individuals with SLE and other autoimmune diseases face significant diagnostic delays (5, 6, 37). While SLE risk models have been developed to identify individuals with SLE, these models have not been validated, deployed in real-time

in the EHR, or assessed to determine if they improve outcomes (4, 38). Our SLE PheRS only requires billing codes and could be easily deployed within the EHR or other administrative databases. The SLE PheRS could potentially be used to assemble individuals with SLE across a healthcare system, not just known SLE individuals followed in a rheumatology clinic. Based on finding 2 SLE individuals in our control sample with high SLE PheRS (0.0088%), applying this estimate to our entire de-identified EHR, while excluding known SLE individuals, could result in identifying approximately 3,200 individuals with possible SLE and 1,700 individuals if we restrict to age < 45 years. The SLE PheRS could work as part of a learning health system (39–41) where billing codes used in real time identify individuals with a high SLE PheRS. These individuals could then be prioritized for rheumatology evaluation.

We assessed if genetic data would add value beyond clinical data to identify individuals with SLE. While genetics contributes to SLE risk, it does not explain the risk completely with a monozygotic twin concordance rate of 24% and a dizygotic twin concordance rate of 2% (42). We developed a SLE GRS as a method to summarize important SLE risk SNPs. As in other SLE GRS studies, we found a similar effect size when comparing GRS in SLE cases to controls. Similar to Hughes et al. (18), males with SLE had a higher GRS compared to females. We did not find a significant association of GRS with age of SLE diagnosis. Our findings contrast with other studies that conducted similar analyses with correlation and linear regression (10, 19). Our study likely did not find an association of GRS with age of SLE diagnosis due to the limited utility of the GRS in Black SLE cases, with Black SLE cases significantly younger at SLE diagnosis compared to White SLE cases. Age at SLE diagnosis not correlating with GRS further demonstrates that the GRS using SNPs from a predominantly European cohort did not capture the genetic risks for Black SLE cases. Age of SLE diagnosis was significantly associated with PheRS, particularly in Black cases. These findings align with studies demonstrating more severe disease with younger age of onset (10, 43–45) as well as younger age of disease onset in Black individuals with SLE (36, 46).

In contrast to other SLE GRS studies (11, 15, 17), in our overall cohort, individuals with SLE nephritis did not have a higher GRS compared to individuals without nephritis. In stratified analyses, White individuals with SLE nephritis had higher GRS compared to SLE individuals without nephritis. This finding was similar to a study where an HLA-based GRS was associated with SLE nephritis only in European ancestry individuals (16). In our study, there was no significant difference in GRS in Black individuals with vs. without SLE nephritis. As Black SLE individuals were overrepresented in individuals with nephritis, we did not observe a significant association with nephritis and GRS in our overall cohort.

Most of the prior SLE GRS studies have not incorporated clinical data with genetic risk scores to determine if genetic data adds value beyond clinical data and have focused on SLE individuals of European and Asian ancestry. The SLE GRS did not add value to the logistic regression model for SLE case status in either the overall or stratified cohorts. Notably, the GRS was not significantly different in Black individuals with SLE compared to Black controls. The GRS is likely to be less helpful in Black individuals, as currently published SLE GWAS do not include SLE individuals of African ancestry (21). Some

studies have shown that different race and ethnic ancestry groups share risk loci for SLE (47, 48), particularly in White and Asian populations. Other studies, however, have shown unique SLE risk loci in Asian (49) and Black individuals (50). Until a GWAS of Black SLE individuals is published, the clinical utility of current SLE GRS may be limited in Black individuals with SLE (21).

While we developed and deployed a SLE PheRS and a GRS in a relatively large sample of SLE individuals, our study has limitations. Our data comes from a single center in the Southeastern US, so our results may not be generalizable to other SLE populations. As our study used EHR data, we do not have access to SLE disease activity or damage measures. These measures are not collected routinely in clinical practice and thus are not available currently in the EHR. Date of SLE diagnosis is also not systematically collected in the EHR, but we had a date of diagnosis documented for 90% of SLE cases and only used first SLE billing code to estimate diagnosis in the remaining 10%. While the demographics of individuals in our de-identified EHR match the demographics of individuals in our genetic biobank, we observed slight differences in race in our SLE and control individuals with EHR data compared to SLE and control individuals with genetic data, due to projects that have generated extant genotype data available to this project. We also had a low proportion of Hispanic individuals. Further, while our controls were initially age, sex, and race matched to our SLE cases, we then selected SLE cases and controls that had available genetic data, which resulted in some imbalances in race between these groups. To account for this difference, we adjusted for race in our models and performed stratified analyses. While we included 3 SNPs in the MHC region which tag the genetic association between SLE and the HLA region, we did not, however, directly examine the effects of HLA classical alleles, as these data were not available.

We developed a SLE PheRS that can identify SLE individuals accurately in the EHR and even found individuals who were undiagnosed. We propose that the SLE PheRS could serve as a clinical tool to identify SLE individuals who may be lost on a diagnostic odyssey in the healthcare system, simply by assessing for patterns of specific billing codes. Further, we demonstrate that genetic data may not add value beyond the clinical data in identifying SLE individuals. The limited utility of genetic data was most evident in Black SLE individuals but may improve as Black SLE GWAS data become available and identify more relevant SLE risk SNPs for this population.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements:

The authors would like to thank Leslie J Crofford, MD for review of the manuscript.

Financial Support:

This work was supported by the National Institutes of Health/National Institute of Arthritis and Musculoskeletal and Skin Diseases (1K08 AR072757-01, R01AR080629, Barnado); the Rheumatology Research Foundation (K Supplement Award, Barnado); US Department of Veterans Affairs Clinical Science Research and Development Service (IK2 CX002452, Wheless); National Institutes of Health/National Institute of Allergy and Infectious

Diseases (R01 AI097134, Sawalha); National Institutes of Health/National Center for Research Resources (UL1 RR024975, VUMC); National Institutes of Health/National Center for Advancing Translational Sciences (ULTR000445, VUMC). Dr. Denny's involvement in this project was primarily as faculty at Vanderbilt University Medical Center prior to joining the NIH.

References

1. Durcan L, Petri M. Why targeted therapies are necessary for systemic lupus erythematosus. *Lupus* 2016;25:1070–9. [PubMed: 27497251]
2. Pramanik B Diagnosis of systemic lupus erythematosus in an unusual presentation: what a primary care physician should know. *Curr Rheumatol Rev* 2014;10:81–6. [PubMed: 25599686]
3. Nightingale AL, Davidson JE, Molta CT, Kan HJ, McHugh NJ. Presentation of SLE in UK primary care using the Clinical Practice Research Datalink. *Lupus Sci Med* 2017;4:e000172.
4. Rees F, Doherty M, Lanyon P, Davenport G, Riley RD, Zhang W, et al. Early Clinical Features in Systemic Lupus Erythematosus: Can They Be Used to Achieve Earlier Diagnosis? A Risk Prediction Model. *Arthritis Care Res (Hoboken)* 2017;69:833–41. [PubMed: 27588834]
5. Olsen NJ, Karp DR. Finding lupus in the ANA haystack. *Lupus Sci Med* 2020;7:e000384.
6. Sebastiani GD, Prevete I, Iuliano A, Piga M, Iannone F, Coladonato L, et al. Early Lupus Project: one-year follow-up of an Italian cohort of patients with systemic lupus erythematosus of recent onset. *Lupus* 2018;27:1479–88. [PubMed: 29779436]
7. Bastarache L, Hughey JJ, Hebring S, Marlo J, Zhao W, Ho WT, et al. Phenotype risk scores identify patients with unrecognized Mendelian disease patterns. *Science* 2018;359:1233–9. [PubMed: 29590070]
8. Zhong X, Yin Z, Jia G, Zhou D, Wei Q, Faucon A, et al. Electronic health record phenotypes associated with genetically regulated expression of CFTR and application to cystic fibrosis. *Genet Med* 2020;22:1191–200. [PubMed: 32296164]
9. Bastarache L Using Phecodes for Research with the Electronic Health Record: From PheWAS to PheRS. *Annu Rev Biomed Data Sci* 2021;4:1–19. [PubMed: 34465180]
10. Webb R, Kelly JA, Somers EC, Hughes T, Kaufman KM, Sanchez E, et al. Early disease onset is predicted by a higher genetic risk for lupus and is associated with a more severe phenotype in lupus patients. *Ann Rheum Dis* 2011;70:151–6. [PubMed: 20881011]
11. Taylor KE, Chung SA, Graham RR, Ortmann WA, Lee AT, Langefeld CD, et al. Risk alleles for systemic lupus erythematosus in a large case-control collection and associations with clinical subphenotypes. *PLoS Genet* 2011;7:e1001311.
12. Joo YB, Lim J, Tsao BP, Nath SK, Kim K, Bae SC. Genetic variants in systemic lupus erythematosus susceptibility loci, XKR6 and GLT1D1 are associated with childhood-onset SLE in a Korean cohort. *Sci Rep* 2018;8:9962. [PubMed: 29967481]
13. Langefeld CD, Ainsworth HC, Cunninghame Graham DS, Kelly JA, Comeau ME, Marion MC, et al. Transancestral mapping and genetic load in systemic lupus erythematosus. *Nat Commun* 2017;8:16021. [PubMed: 28714469]
14. Gianfrancesco MA, Balzer L, Taylor KE, Trupin L, Nititham J, Seldin MF, et al. Genetic risk and longitudinal disease activity in systemic lupus erythematosus using targeted maximum likelihood estimation. *Genes Immun* 2016;17:358–62. [PubMed: 27467283]
15. Chen L, Wang YF, Liu L, Bielowka A, Ahmed R, Zhang H, et al. Genome-wide assessment of genetic risk for systemic lupus erythematosus and disease severity. *Hum Mol Genet* 2020;29:1745–56. [PubMed: 32077931]
16. Webber D, Cao J, Dominguez D, Gladman DD, Levy DM, Ng L, et al. Association of systemic lupus erythematosus (SLE) genetic susceptibility loci with lupus nephritis in childhood-onset and adult-onset SLE. *Rheumatology (Oxford)* 2020;59:90–8. [PubMed: 31236574]
17. Reid S, Alexsson A, Frodlund M, Morris D, Sandling JK, Bolin K, et al. High genetic risk score is associated with early disease onset, damage accrual and decreased survival in systemic lupus erythematosus. *Ann Rheum Dis* 2020;79:363–9. [PubMed: 31826855]

18. Hughes T, Adler A, Merrill JT, Kelly JA, Kaufman KM, Williams A, et al. Analysis of autosomal genes reveals gene-sex interactions and higher total genetic risk in men with systemic lupus erythematosus. *Ann Rheum Dis* 2012;71:694–9. [PubMed: 22110124]
19. Dominguez D, Kamphuis S, Beyene J, Wither J, Harley JB, Blanco I, et al. Relationship Between Genetic Risk and Age of Diagnosis in Systemic Lupus Erythematosus. *J Rheumatol* 2021;48:852–8. [PubMed: 33060314]
20. Roden DM, Pulley JM, Basford MA, Bernard GR, Clayton EW, Balsler 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]
21. Harley ITW, Sawalha AH. Systemic lupus erythematosus as a genetic disease. *Clin Immunol* 2022;236:108953.
22. 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;69:687–93. [PubMed: 27390187]
23. Barnado A, Carroll RJ, Casey C, Wheless L, Denny JC, Crofford LJ. Phenome-Wide Association Studies Uncover a Novel Association of Increased Atrial Fibrillation in Male Patients With Systemic Lupus Erythematosus. *Arthritis Care Res (Hoboken)* 2018;70:1630–6. [PubMed: 29481723]
24. Barnado A, Carroll RJ, Casey C, Wheless L, Denny JC, Crofford LJ. Phenome-wide association study identifies marked increased in burden of comorbidities in African Americans with systemic lupus erythematosus. *Arthritis Res Ther* 2018;20:69. [PubMed: 29636090]
25. Barnado A, Carroll RJ, Casey C, Wheless L, Denny JC, Crofford LJ. Phenome-wide association study identifies dsDNA as a driver of major organ involvement in systemic lupus erythematosus. *Lupus* 2019;28:66–76. [PubMed: 30477398]
26. Xiong WW, Boone JB, Wheless L, Chung CP, Crofford LJ, Barnado A. Real-world electronic health record identifies antimalarial underprescribing in patients with lupus nephritis. *Lupus* 2019;28:977–85. [PubMed: 31189414]
27. Boone JB, Wheless L, Camai A, Tanner SB, Barnado A. Low prevalence of bone mineral density testing in patients with systemic lupus erythematosus and glucocorticoid exposure. *Lupus* 2021;30:403–11. [PubMed: 33307984]
28. Gandelman JS, Khan OA, Shuey MM, Neal JE, McNeer E, Dickson A, et al. Increased Incidence of Resistant Hypertension in Patients With Systemic Lupus Erythematosus: A Retrospective Cohort Study. *Arthritis Care Res (Hoboken)* 2020;72:534–43. [PubMed: 30875459]
29. Tan EM, Cohen AS, Fries JF, Masi AT, McShane DJ, Rothfield NF, et al. The 1982 revised criteria for the classification of systemic lupus erythematosus. *Arthritis Rheum* 82;25:1271–7.
30. Morris DL, Sheng Y, Zhang Y, Wang YF, Zhu Z, Tomblinson P, et al. Genome-wide association meta-analysis in Chinese and European individuals identifies ten new loci associated with systemic lupus erythematosus. *Nat Genet* 2016;48:940–6. [PubMed: 27399966]
31. Petri M, Orbai AM, Alarcon GS, Gordon C, Merrill JT, Fortin PR, et al. Derivation and validation of the Systemic Lupus International Collaborating Clinics classification criteria for systemic lupus erythematosus. *Arthritis Rheum* 2012;64:2677–86. [PubMed: 22553077]
32. Dumitrescu L, Ritchie MD, Brown-Gentry K, Pulley JM, Basford M, Denny JC, et al. Assessing the accuracy of observer-reported ancestry in a biorepository linked to electronic medical records. *Genet Med* 2010;12:648–50. [PubMed: 20733501]
33. Aringer M EULAR/ACR classification criteria for SLE. *Semin Arthritis Rheum* 2019;49:S14–S7. [PubMed: 31779843]
34. Aref L, Bastarache L, Hughey JJ. The phers R package: using phenotype risk scores based on electronic health records to study Mendelian disease and rare genetic variants. *Bioinformatics* 2022.
35. Costenbader KH, Desai A, Alarcon GS, Hiraki LT, Shaykevich T, Brookhart MA, et al. Trends in the incidence, demographics, and outcomes of end-stage renal disease due to lupus nephritis in the US from 1995 to 2006. *Arthritis Rheum* 2011;63:1681–8. [PubMed: 21445962]

36. 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]
37. Sloan M, Harwood R, Sutton S, D’Cruz D, Howard P, Wincup C, et al. Medically explained symptoms: a mixed methods study of diagnostic, symptom and support experiences of patients with lupus and related systemic autoimmune diseases. *Rheumatol Adv Pract* 2020;4:rkaa006.
38. Adamichou C, Genitsaridi I, Nikolopoulos D, Nikoloudaki M, Repa A, Bortoluzzi A, et al. Lupus or not? SLE Risk Probability Index (SLERPI): a simple, clinician-friendly machine learning-based model to assist the diagnosis of systemic lupus erythematosus. *Ann Rheum Dis* 2021;80:758–77. [PubMed: 33568388]
39. Walker SC, Creech CB, Domenico HJ, French B, Byrne DW, Wheeler AP. A Real-time Risk-Prediction Model for Pediatric Venous Thromboembolic Events. *Pediatrics* 2021;147:e2020042325.
40. Freundlich RE, Li G, Domenico HJ, Moore RP, Pandharipande PP, Byrne DW. A Predictive Model of Reintubation After Cardiac Surgery Using the Electronic Health Record. *Ann Thorac Surg* 2022;113:2027–35. [PubMed: 34329600]
41. Wang L, McGregor TL, Jones DP, Bridges BC, Fleming GM, Shirey-Rice J, et al. Electronic health record-based predictive models for acute kidney injury screening in pediatric inpatients. *Pediatr Res* 2017;82:465–73. [PubMed: 28486440]
42. Deapen D, Escalante A, Weinrib L, Horwitz D, Bachman B, Roy-Burman P, et al. A revised estimate of twin concordance in systemic lupus erythematosus. *Arthritis Rheum* 1992;35:311–8. [PubMed: 1536669]
43. Brunner HI, Gladman DD, Ibanez D, Urowitz MD, Silverman ED. Difference in disease features between childhood-onset and adult-onset systemic lupus erythematosus. *Arthritis Rheum* 2008;58:556–62. [PubMed: 18240232]
44. Hersh AO, von Scheven E, Yazdany J, Panopalis P, Trupin L, Julian L, et al. Differences in long-term disease activity and treatment of adult patients with childhood- and adult-onset systemic lupus erythematosus. *Arthritis Rheum* 2009;61:13–20. [PubMed: 19116979]
45. Tucker LB, Uribe AG, Fernandez M, Vila LM, McGwin G, Apte M, et al. Adolescent onset of lupus results in more aggressive disease and worse outcomes: results of a nested matched case-control study within LUMINA, a multiethnic US cohort (LUMINA LVII). *Lupus* 2008;17:314–22. [PubMed: 18413413]
46. 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]
47. Wang C, Ahlford A, Jarvinen TM, Nordmark G, Eloranta ML, Gunnarsson I, et al. Genes identified in Asian SLE GWASs are also associated with SLE in Caucasian populations. *Eur J Hum Genet* 2013;21:994–9. [PubMed: 23249952]
48. Cui J, Raychaudhuri S, Karlson EW, Speyer C, Malspeis S, Guan H, et al. Interactions Between Genome-Wide Genetic Factors and Smoking Influencing Risk of Systemic Lupus Erythematosus. *Arthritis Rheumatol* 2020;72:1863–71. [PubMed: 32969204]
49. Han JW, Zheng HF, Cui Y, Sun LD, Ye DQ, Hu Z, et al. Genome-wide association study in a Chinese Han population identifies nine new susceptibility loci for systemic lupus erythematosus. *Nat Genet* 2009;41:1234–7. [PubMed: 19838193]
50. Ruiz-Narvaez EA, Fraser PA, Palmer JR, Cupples LA, Reich D, Wang YA, et al. MHC region and risk of systemic lupus erythematosus in African American women. *Hum Genet* 2011;130:807–15. [PubMed: 21695597]

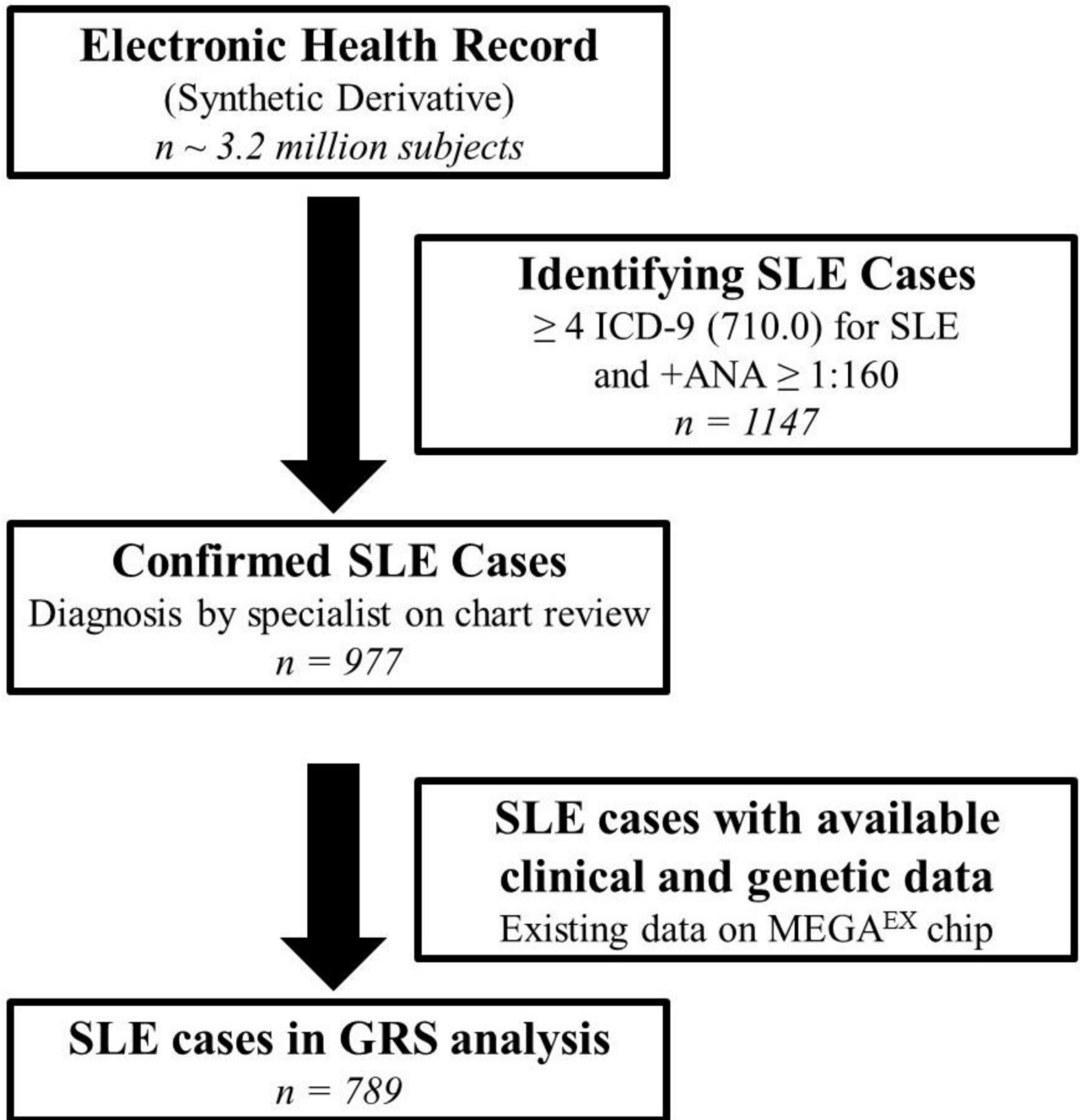


Figure 1. Flow chart of selection of SLE cases.

We used a large, de-identified electronic health record called the Synthetic Derivative to select our SLE cases. We used an algorithm with a positive predictive value of 90% requiring 4 or more SLE ICD-9 codes (710.0) and a positive ANA (1:160). We performed chart review to confirm SLE case status and then selected SLE cases that had both available genetic data (existing data on the MEGA^{EX} chip) and clinical data (including billing codes).

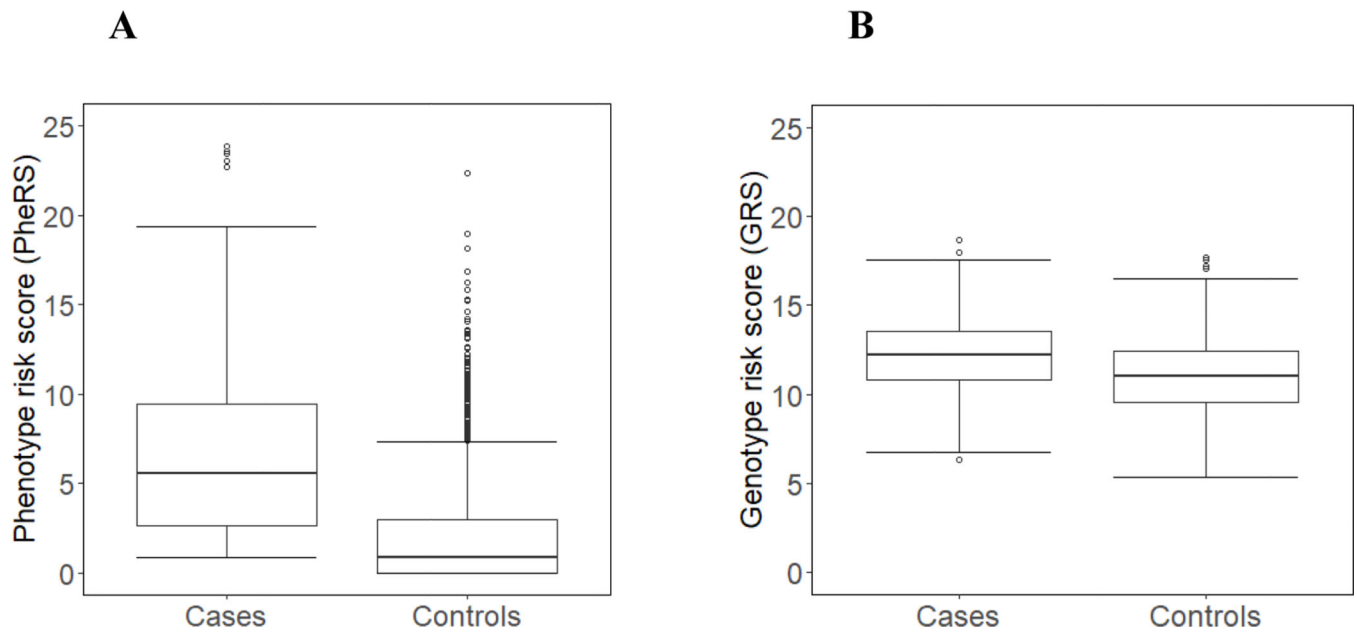


Figure 2. Boxplot of Systemic lupus erythematosus (SLE) phenotype risk scores (PheRS) in SLE cases vs. controls.

(A) We identified 789 SLE cases, all who had a SLE diagnosis confirmed by a rheumatologist. We identified 2,261 controls with no known autoimmune disease diagnoses. The horizontal line indicates the median score. (B) **Boxplot of SLE genetic risk scores (GRS) in SLE cases vs. controls.** Both SLE cases and controls were required to have genetic data available on MEGA^{EX} chip. The genetic risk score consisted of 58 SLE risk SNPs with a 95% sampling and 95% call rate.

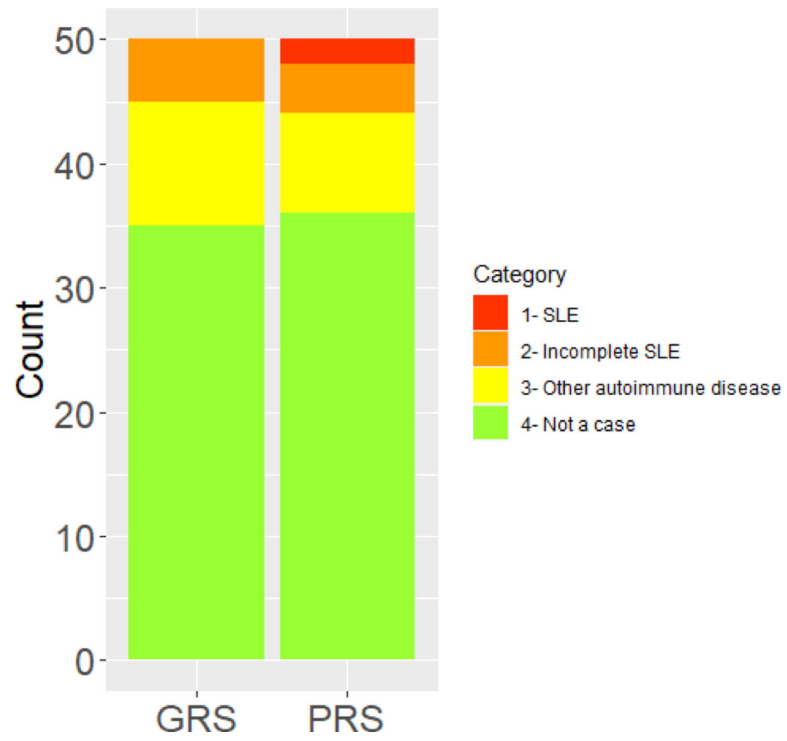


Figure 3. Controls with the highest genetic risk scores (GRS) and phenotype risk scores (PheRS). The bar graphs show the proportion of diagnoses of the controls with the 50 highest GRS and PheRS. Categories included systemic lupus erythematosus (SLE), incomplete SLE, other autoimmune disease, and no autoimmune disease or not a case.

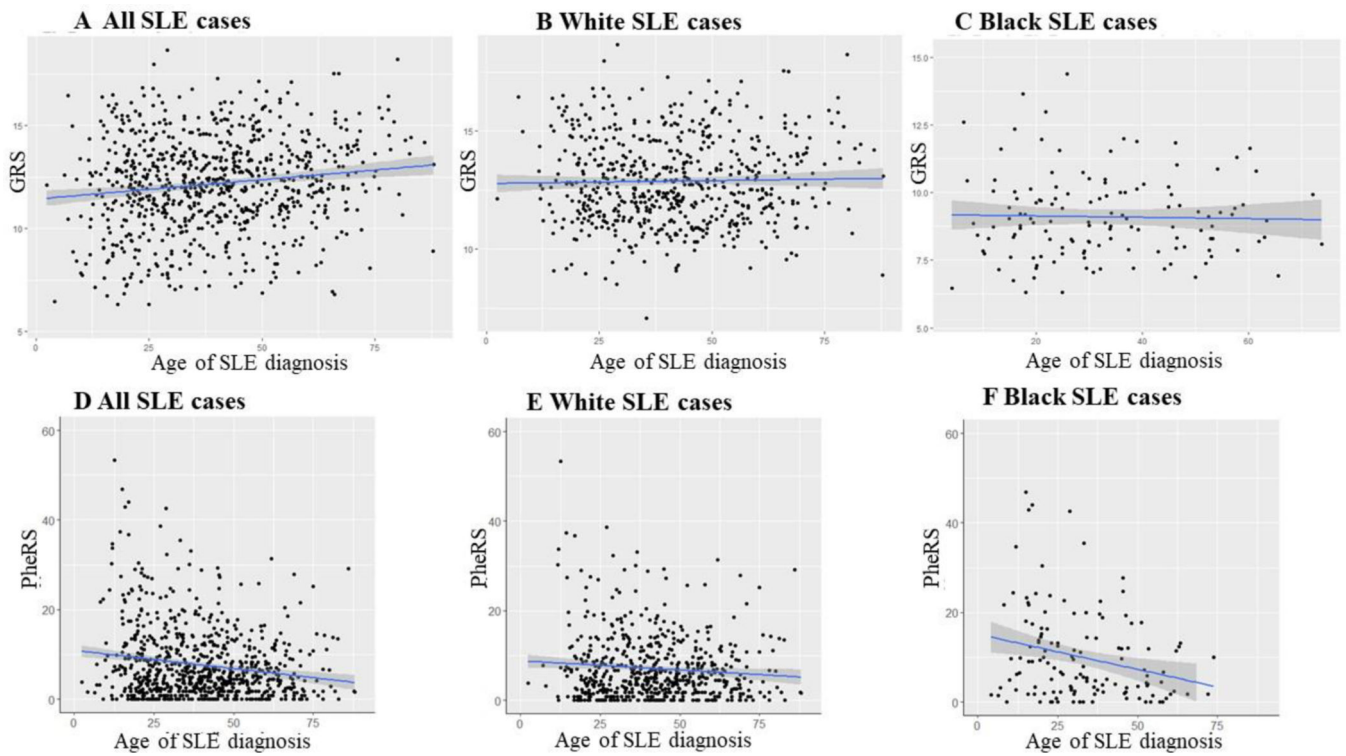
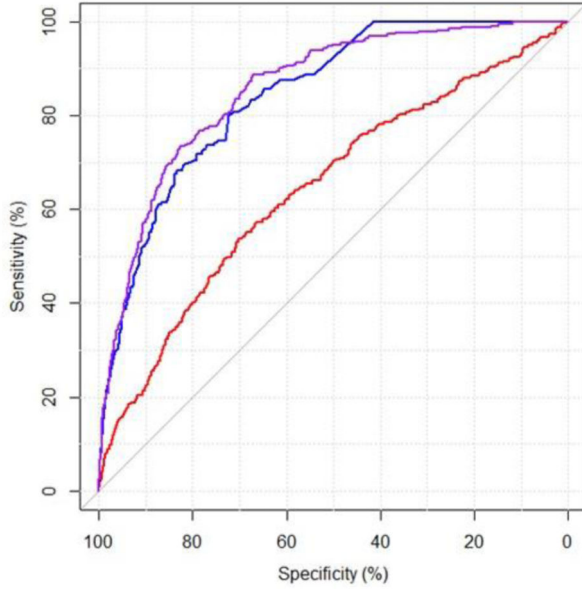


Figure 4. Scatterplot of age of SLE diagnosis and SLE genetic risk score (GRS) and phenotype risk score (PheRS).

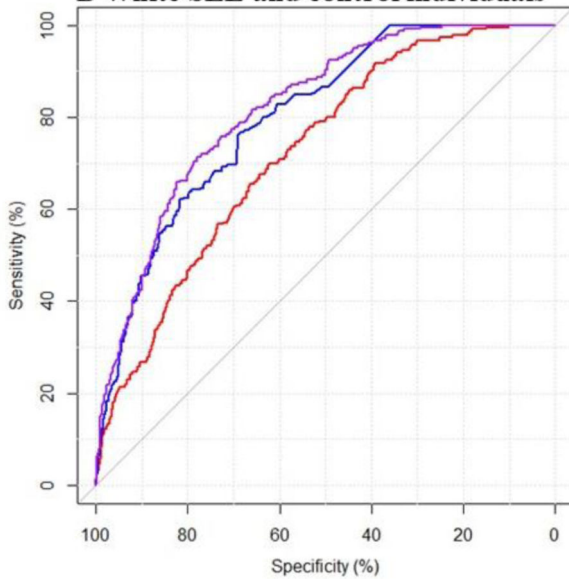
A) Age of SLE diagnosis and GRS in all SLE cases, B) Age of SLE diagnosis and GRS in White SLE cases, C) Age of SLE diagnosis and GRS in Black SLE cases, D) Age of SLE diagnosis and PheRS in all SLE cases, E) Age of SLE diagnosis and PheRS in White SLE cases, and F) Age of SLE diagnosis and PheRS in Black SLE cases. Age of SLE diagnosis was obtained from chart review.

A All SLE and control individuals



Blue = PheRS
Red = GRS
Purple = PheRS + GRS

B White SLE and control individuals



C Black SLE and control individuals

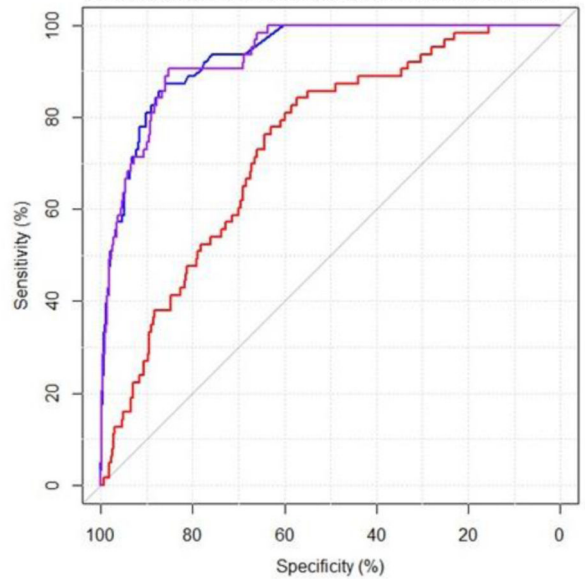


Figure 5. AUC for models for SLE case status using SLE phenotype risk score (PheRS) and genotype risk score (GRS).

A) Models in all SLE and control individuals. B) Models in White SLE and control individuals only. C) Models in Black SLE and control individuals only. The blue line denotes an unadjusted model using only the PheRS. The red line denotes an unadjusted model for SLE case status using only the GRS while the purple line denotes an unadjusted model with both PheRS and GRS.

Table 1.

Comparison of SLE cases and controls.

Characteristics	SLE cases (n = 789)	Controls (n = 2261)	p value
Current mean age \pm standard deviation	56.2 \pm 16.9	55.0 \pm 17.3	0.08
Sex			
Female (%)	89%	90%	0.56
Male (%)	11%	10%	
Race			
White (%)	78%	61%	< 0.001
Black (%)	17%	34%	
Asian (%)	2%	2%	
Other (%)	3%	3%	
Ethnicity			
Hispanic (%)	2%	2%	0.65