Supplementary Methods

Clinical setting, discovery and validation cohorts

This is a retrospective observational study exploiting data from the Department of Rheumatology, University Hospital of Heraklion (community-based centre at Heraklion, Crete) and the Rheumatology Clinic, "Attikon" University Hospital (tertiary centre in Athens, Attica) in Greece. Both centres have established SLE registries and use homogenized, structured forms for collecting demographics, clinical characteristics (including classification criteria), use of treatments and disease outcomes.[1-4] We included consecutively registered cases diagnosed during the period 01/2005-06/2019 with SLE or other diseases by consultant rheumatologists with at least 5 years clinical practice. The aforementioned time interval was selected to ensure data completeness and reduce possible information/classification bias. Additional inclusion criteria were: sufficient patient identification and clinical data, age of diagnosis ≥16 years, known status of ANA and serum complements C3/C4, and follow-up at least 6 months to confirm the diagnosis (SLE or other). Disease controls were selected from the electronic patient databases of the two participating centres, and included miscellaneous lupus-mimicking rheumatologic diseases (**Supplementary Table S1**).

We used a randomly-selected sample of 401 SLE patients and 401 disease controls (*discovery cohort*) to construct, train and compare the ML models. Part of this dataset has been used in previous work to evaluate the SLE classification criteria.[1] The balanced (1:1) ratio of SLE cases and controls was chosen in order to minimise any predictive modelling biases. An independent sample of 512 SLE patients and 143 disease controls (*validation cohort*) was used to provide an unbiased estimate of the diagnostic accuracy of the best model selected from the discovery cohort. The study was approved by the local Ethics Committees.

Clinical variables and dataset preparation

For each patient, demographics, rheumatological disease and date of diagnosis, presence and date of earliest reported occurrence of each of the items from all three classification criteria sets (ACR 1997,[5] SLICC 2012,[6] EULAR/ACR 2019[7, 8]) and date of last follow-up visit/assessment were extracted from medical charts. Attribution of the criteria items to SLE or not was arbitrated by rheumatologists (DB, GB, AF) with special interest and experience in the disease, and we followed the EULAR/ACR attribution process[7, 8] We used both criteria items in their original version and after deconvolution into sub-items (e.g. "*maculopapular lupus rash*" sub-item from the EULAR/ACR 2019 "*acute cutaneous lupus*" criterion). In addition, we monitored for the presence of a predefined list of clinical and serological features not included in the criteria (e.g. Raynaud's phenomenon, anti-RNP antibodies) (**Supplementary Table S2**). Missing data were eliminated through vigorous charts review and quality control, and a few patients with no documentation of immunological tests were considered negative for these items.

Computational methods for feature selection, model construction and evaluation

We followed two approaches for developing a predictive model for SLE using the discovery cohort. *First*, we combined each one of the three classification criteria (ACR 1997, SLICC 2012, EULAR/ACR 2019) with additional, non-redundant features from the other two criteria sets and with non-criteria features. Classification criteria were treated both as binary variable (classified/not-classified) and as a continuous score (e.g., sum of ACR-1997 features). *Second*, we developed a *de novo* model based on clinical variables selected from the three classification criteria (in their original form or deconvoluted into sub-items) and non-criteria features.

Univariable LR (**Supplementary Table S3**) was carried out in the discovery cohort to determine the individual association of each feature with SLE diagnosis. Correlation analysis (**Supplementary Table S4**) was performed to identify collinearity between features/predictors. Both analyses were not used to filter the variables, rather for completeness purposes and to assist clinicians in the construction of feature panels. Clinicians (GB, CA) created 20 different panels of features with the aim to introduce alternative feature versions in different panels (based on criteria deconvolution into sub-items). These 20 panels were submitted into two ML algorithms for feature selection, thus yielding a total 40 multivariable models (**Figure 1**). The first algorithm, *Random Forests* (RF), is a complex, model-free learning method with fewer assumptions and an embedded feature selection process to further address redundant information, capable of generating highly accurate – yet less explainable due to its non-linear nature – predictions. The second, *Logistic Regression* (LR), is a standard, less complex method of analysing biomedical datasets with high interpretability, which however, lacks a feature selection process. Therefore, for each panel, a separate feature selection process was carried out prior to LR model construction, using *Least Absolute Shrinkage and Selection Operator* (LASSO) Logistic Regression method. A notable difference between these two methods is that if several highly correlated variables are predictive, LASSO may select one or a few while RF may use all of them (selected in different trees as part of the ensemble algorithm's majority voting process).

LASSO hyper-parameter λ (*lambda*) was optimized in the discovery cohort (from 100 sequential parameter values of equal intervals ranging from λ=0, which includes all features, to the lowest λ value that excludes all features from the model). As optimal λ was selected the value that yielded the minimum deviance plus one standard deviation in a 10-fold CV process and was also higher to the λ value of minimum deviance (this is a slightly stricter feature selection approach than the standard approach of selecting the λ of minimum deviance since it includes higher λ penalty and thus lower number of selected features). Features with non-zero beta coefficients in the LASSO logistic regression of the optimal λ value were subsequently included in a LR model. RF algorithm was implemented with hyper-parameter number of trees value=50. For other parameters of the algorithms, the default values or settings were used.

We performed a 10-fold stratified cross-validation (CV 10-fold) process (division of the dataset into 10 folds of near-equal size without re-substitution) to construct and compare the 40 multivariable models (20 RF, 20 LR) in terms of their predictive capability. Each fold (10%) was used as a *test dataset* to determine the model performance, while the remaining nine folds (90%) were used as the *training dataset* for the model construction. In the test dataset, we evaluated the following diagnostic metrics: sensitivity, specificity, accuracy and Area Under the Receiver-Operating-Characteristic Curve (AUC). These metrics were averaged from the 10 CV test datasets for each model. The model with the highest accuracy was selected as the best model. The best model was further evaluated in an independent *validation cohort* of 512 SLE patients and 143 disease controls by calculating the AUC, sensitivity, specificity, accuracy, positive and negative likelihood ratios, thus obtaining an unbiased estimate of its diagnostic performance.

Feature ranking

The drop-column methodology was used (10-fold CV process in the discovery cohort) to evaluate the importance of each feature in the model predictive functionality, by excluding it from the construction process and estimating the change in the model accuracy.

Statistical analysis

The model accuracy derived from the validation cohort was corrected based on an expected ratio 3:17 of SLE patients:controls (mimicking rheumatological diseases) receiving diagnosis at a general rheumatology outpatient clinic according to the following formula:

$$\frac{(sensitivity \times SLE\ count) + (specificity \times controls\ count)}{total\ count}$$

References

1. Adamichou C, Nikolopoulos D, Genitsaridi I*, et al.* In an early SLE cohort the ACR-1997, SLICC-2012 and EULAR/ACR-2019 criteria classify non-overlapping groups of patients: use of all three criteria ensures optimal capture for clinical studies while their modification earlier classification and treatment. *Ann Rheum Dis* 2020;79:232-41.

2. Gergianaki I, Fanouriakis A, Repa A*, et al.* Epidemiology and burden of systemic lupus erythematosus in a Southern European population: data from the community-based lupus registry of Crete, Greece. *Ann Rheum Dis* 2017;76:1992-2000.

3. Nikolopoulos D, Kostopoulou M, Pieta A*, et al.* Evolving phenotype of systemic lupus erythematosus in Caucasians: low incidence of lupus nephritis, high burden of neuropsychiatric disease and increased rates of late-onset lupus in the 'Attikon' cohort. *Lupus* 2020;29:514-22.

4. Nikolopoulos DS, Kostopoulou M, Pieta A*, et al.* Transition to severe phenotype in systemic lupus erythematosus initially presenting with non-severe disease: implications for the management of early disease. *Lupus Sci Med* 2020;7.

5. Hochberg MC. Updating the American College of Rheumatology revised criteria for the classification of systemic lupus erythematosus. *Arthritis Rheum* 1997;40:1725.

6. Petri M, Orbai AM, Alarcon GS*, et al.* Derivation and validation of the Systemic Lupus International Collaborating Clinics classification criteria for systemic lupus erythematosus. *Arthritis Rheum* 2012;64:2677-86.

7. Aringer M, Costenbader K, Daikh D*, et al.* 2019 European League Against Rheumatism/American College of Rheumatology Classification Criteria for Systemic Lupus Erythematosus. *Arthritis Rheumatol* 2019;71:1400-12.

8. Aringer M, Costenbader K, Daikh D*, et al.* 2019 European League Against Rheumatism/American College of Rheumatology classification criteria for systemic lupus erythematosus. *Ann Rheum Dis* 2019;78:1151-9.