LUPUS SCIENCE & MEDICINE™

# Adverse pregnancy outcomes in women with systemic lupus erythematosus: can we improve predictions with machine learning?

Melissa J Fazzari ,[1] Marta M Guerra,[2] Jane Salmon,[2,3] Mimi Y Kim[1]

**To cite:** Fazzari MJ, Guerra MM, Salmon J, *et al*. Adverse pregnancy outcomes in women with systemic lupus erythematosus: can we improve predictions with machine learning?. *Lupus Science & Medicine* 2022;**9**:e000769. doi:10.1136/lupus-2022-000769

► Additional supplemental material is published online only. To view, please visit the journal online (http://dx.doi.org/10.1136/lupus-2022-000769).

JS and MYK contributed equally.

Received 30 June 2022
Accepted 1 September 2022



[1]Epidemiology and Population Health, Albert Einstein College of Medicine, Bronx, New York, USA
[2]Rheumatology, Hospital for Special Surgery, New York, New York, USA
[3]Department of Medicine, Weill Cornell Medicine, New York, New York, USA

**Correspondence to**
Dr Melissa J Fazzari; melissa.fazzari@einsteined.edu

## ABSTRACT

**Objectives** Nearly 20% of pregnancies in patients with SLE result in an adverse pregnancy outcome (APO). We previously developed an APO prediction model using logistic regression and data from Predictors of pRegnancy Outcome: bioMarkers In Antiphospholipid Antibody Syndrome and Systemic Lupus Erythematosus (PROMISSE), a large multicentre study of pregnant women with mild/moderate SLE and/or antiphospholipid antibodies. Our goal was to determine whether machine learning (ML) approaches improve APO prediction and identify other risk factors.

**Methods** The PROMISSE data included 41 predictors from 385 subjects; 18.4% had APO (preterm delivery due to placental insufficiency/pre-eclampsia, fetal/neonatal death, fetal growth restriction). Logistic regression with stepwise selection (LR-S), least absolute shrinkage and selection operator (LASSO), random forest (RF), neural network (NN), support vector machines (SVM-RBF), gradient boosting (GB) and SuperLearner (SL) were compared by cross-validated area under the ROC curve (AUC) and calibration.

**Results** Previously identified APO risk factors, antihypertensive medication use, low platelets, SLE disease activity and lupus anticoagulant (LAC), were confirmed as important with each algorithm. LASSO additionally revealed potential interactions between LAC and anticardiolipin IgG, among others. SL performed the best (AUC=0.78), but was statistically indistinguishable from LASSO, SVM-RBF and RF (AUC=0.77 for all). LR-S, NN and GB had worse AUC (0.71–0.74) and calibration scores.

**Conclusions** We predicted APO with reasonable accuracy using variables routinely assessed prior to the 12th week of pregnancy. LASSO and some ML methods performed better than a standard logistic regression approach. Substantial improvement in APO prediction will likely be realised, not with increasingly complex algorithms but by the discovery of new biomarkers and APO risk factors.

### WHAT IS ALREADY KNOWN ON THIS TOPIC

⇒ All SLE pregnancies are intensely monitored given their increased risk for adverse outcomes (APO).
⇒ APO risk was previously modelled using logistic regression and data from a large study of SLE pregnancies (Predictors of pRegnancy Outcome: bioMarkers In Antiphospholipid Antibody Syndrome and Systemic Lupus Erythematosus (PROMISSE)).
⇒ Complex machine learning (ML) approaches are increasingly used for prediction of patient outcomes, but little is known about their utility in comparison to standard statistical models, particularly in predicting APO in SLE pregnancies.

### WHAT THIS STUDY ADDS

⇒ In the PROMISSE dataset, ML methods did not offer clear improvement over penalised regression-based approaches in identifying pregnancies with APOs.
⇒ Novel interaction effects, including between lupus anticoagulant and anticardiolipin IgG, were identified via penalised regression to be validated in future studies.

### HOW THIS STUDY MIGHT AFFECT RESEARCH, PRACTICE OR POLICY

⇒ A highly interpretable penalised regression approach provides a simple APO predictive model for SLE pregnancies using routinely collected patient information to be further validated in new patient datasets.
⇒ While larger SLE pregnancy databases would be beneficial, further improvements to this APO predictive model will likely require the discovery of new biomarkers and risk factors.

SLE predominantly affects women and presents during their childbearing years. Pregnancy in patients with SLE, particularly those with antiphospholipid antibodies (aPL), is associated with increased risk for maternal and fetal morbidity and mortality. The clinical consequences of abnormal placental development, inflammatory injury and placental hypoperfusion that can occur in lupus pregnancies affect the maternal/fetal dyad. These include pre-eclampsia in the mother, intrauterine fetal death, neonatal death, fetal growth restriction and preterm delivery due to placental insufficiency. Nearly 20% of pregnancies in patients with SLE result in an adverse pregnancy outcome (APO), despite quiescent lupus disease activity. Yet, there are no established instruments to predict outcomes in individual

patients. Thus, all SLE pregnancies are intensely monitored at an emotional cost to patients and financial cost to society. Furthermore, without validated risk stratification models, trials to prevent APOs are challenging to design and execute. The ability to identify, early in pregnancy, patients at high risk of APO would enhance our capacity to manage these patients and conduct trials of new treatments to prevent pre-eclampsia and placental insufficiency. At the same time, the costly and intensive monitoring during pregnancy, as well as patient stress, could be reduced for those identified with high confidence to be at very low risk of APO.

The Predictors of pRegnancy Outcome: bioMarkers In Antiphospholipid Antibody Syndrome and Systemic Lupus Erythematosus (PROMISSE) study is the largest multicentre, multi-ethnic and multiracial study to date to prospectively assess clinical and laboratory predictors of APO in women with SLE and/or aPL with inactive or mild/moderate disease activity at conception. We previously analysed the PROMISSE data from the first trimester using standard logistic regression and identified lupus anticoagulant (LAC+), current use of antihypertensive medication, SLE disease activity, low platelets and non-white race as significant baseline predictors of APO.[1] Logistic regression has been widely applied in clinical and epidemiological studies to assess predictors of a binary outcome (eg, APO, no APO), and yields easily interpretable results, in the form of ORs, of the association of each predictor in the model with the outcome. While our previously published logistic regression model for APO was shown to perform reasonably well, more complex functions of the predictor variables, such as interactions, were not considered, and logistic regression cannot capture potentially important relationships between predictors and the outcome unless they are explicitly characterised in the model a priori. As a result, the current model may have underfit the data producing biased estimates of APO risk.

'Black box' machine learning (ML) algorithms, capable of fitting complex and flexible models without explicit specification, are increasingly used in medicine for diagnostic and predictive purposes. For example, random forest was successfully used to predict lupus disease activity using gene expression data, and gradient boosting was used to predict lupus nephritis flares.[2 3] In this paper, we aimed to determine whether ML methods improve the ability to predict APO in patients with SLE based on data obtained early in pregnancy. We also compared the most important APO predictors identified with each method and our original model to obtain a more complete understanding of the major risk factors for APOs and to generate hypotheses for future studies that would inform the development of strategies to mitigate APO risk in patients with lupus.

## METHODS

### PROMISSE data

The training data to fit the ML prediction models was from PROMISSE, a multicentre, prospective observational study of pregnancies in women with SLE (≥4 revised American College of Rhematology criteria),[4] women with SLE with and without aPLs, aPLs alone and healthy women at low risk for APOs (≥1 successful pregnancy, no prior fetal death and<2 miscarriages at <10 weeks' gestation). Pregnant women at <12 weeks' gestation were enrolled in the National Institute of Arthritis and Muscloskeletal and Skin Diseases (NIAMS)-funded study between September 2003 and December 2012; additional patients, not included in previous reports, were enrolled until 2018. This paper focuses on data from the 385 patients with SLE (with or without aPL) currently available in PROMISSE.

Inclusion criteria for enrolment were singleton intrauterine pregnancy, age 18–45 years and haematocrit >26%. To identify risk factors for APOs specifically attributable to lupus and/or aPLs, exclusion criteria were prednisone use >20 mg/day, urinary protein/creatinine ratio >1000 mg/g, erythrocyte casts on urinalysis, serum creatinine level >1.2 mg/dL, diabetes mellitus and blood pressure >140/90 mm Hg at screening. aPL (LAC, anticardiolipin (aCL) IgG and IgM and anti-β2GPI IgG and IgM) assays were performed in core labs, and positivity was defined as previously described.[5]

The primary outcome, APO, is a composite end point that includes (1) unexplained fetal death after 12 weeks' gestation; (2) neonatal death before hospital discharge due to complications of prematurity and placental insufficiency; (3) preterm delivery at <36 weeks' gestation due to gestational hypertension, pre-eclampsia or placental insufficiency and (4) small for gestational age neonate (<5th percentile).

### Patient variables

The original PROMISSE dataset consisted of demographic, clinical and laboratory variables of interest. Our goal is to develop an APO clinical prediction tool that can be used by physicians practising in different specialties and by patients. Therefore, we focused on 41 predictor variables routinely assessed in the clinical management of patients with SLE: demographic characteristics (age, ethnicity, race, body mass index), clinical history (SLE criteria, renal status, thrombosis history), baseline clinical and laboratory values (blood pressure, platelets, urine protein/creatinine ratio, lupus serologies, aPL tests), medications and disease activity measures (online supplemental table 1). Other potential risk factors and biomarkers (ie, angiogenic factors, complement activation products) were considered, but not included because they were deemed to be more costly, time consuming, invasive or impossible to assess in routine care. Only the first measurement obtained during the first trimester from each patient was analysed to enable early identification of those at highest risk of APO and to maximise the time window for potentially intervening and improving their outcomes.

No variables in the dataset were missing in >10% of patients; however, 98 (25.4%) patients were missing at least one variable, and 34 (8.8%) were missing at least two.

We used a single iteration of the multivariate imputation by chained equations procedure[6] to obtain a complete dataset of 385 patients and 41 predictors.

To explore potential interactions between predictor variables, an expanded dataset consisting of the primary 41 clinical covariates, plus an additional 820 pairwise interaction terms, was also generated. All variables were centred and scaled prior to analysis.

### Predictive algorithms

We evaluated the performance of standard logistic regression with stepwise variable selection (LR-S), penalised logistic regression via least absolute shrinkage and selection operator (LASSO), random forest (RF), support vector machines with radial basis function kernel (SVM-RBF), gradient boosting (GB), neural networks (NN) and SuperLearner (SL). All of these algorithms generate a continuous estimate of APO risk between 0 and 1 for each patient, with higher values indicating higher predicted risk. Using the same logistic regression framework, LR-S iteratively selects predictors that have statistically significant associations with APO after adjusting for the other variables in the model, while LASSO shrinks regression coefficients towards zero, dropping predictor variables with coefficients exactly equal to zero from the model to yield a simpler model.[7 8] We note that our earlier APO prediction model used a variation of LR-S with variable selection based on both statistical significance and clinical factors.[1] NN predicts outcome based on a weighted combination of models (neurons) within hidden layers and can model higher-order interactions and complex relationships depending on the network architecture.[9] We considered a NN with three neurons (NN-1) and another NN with two hidden layers (NN-2), with the first layer containing three neurons and the second containing two neurons. SVM-RBF finds the best way to separate APO from non-APO in an expanded variable space that includes complex functions of the original input variables with the use of a radial basis function kernel.[10] RF builds an ensemble of independent classification trees using perturbed versions of the same dataset and a random selection of variables for tree-building[11]; the final prediction for an individual using RF is based on the proportion of votes across the ensemble members that predict APO for that subject. GB is an ensemble of 'shallow' decision trees, where trees are grown sequentially, with the next tree minimising the loss of the previous tree.[12] Finally, SL is an ensemble method that combines predictions from a diverse set of models using optimised weights.[13] In this study, we included RF, LASSO, GB and SVM-RBF classifiers in the ensemble. Hyperparameters for each method were determined through internal cross-validation (CV). Algorithms and simulations were conducted in R V.10.4.[14]

### Algorithm performance

To evaluate the performance of each modelling approach, we used 5×10-fold CV, with performance summaries representing the average across five independent 10-fold CVs.[15] Within each of the five runs, the 10-fold CV is conducted by randomly splitting the data into 10 distinct groups (folds). For each fold $k$ ($k$=1, 2, … 10), patients in $k$ are treated as the test set, while a model is estimated on the remaining 90% of the data not in $k$. This process is repeated 10 times so that predictions for any specific individual is based on a model developed and estimated without including that individual's data in the training set. Within each of the 10 CV folds, model building and estimation are performed independently. Model discrimination for each algorithm was computed based on the cross-validated area under the receiver operating characteristic curve (AUC) and 95% CIs.[16] Additionally, we computed sensitivity and specificity at an optimal cut-point based on maximising the sum of sensitivity and specificity across all candidate cut-points (Youden index). To assess model calibration, we evaluated the Brier score, which is the average of the squared difference between the model-based prediction and actual outcome (0=non-APO, 1=APO). We additionally computed a reliability score by separating predictions into four equidistant bins of predicted APO risk (≤25%, 26%–50%, 51%–75%, >75%), and computing the average squared difference between actual APO rates versus the binned predictions.[17] The sample size and number of APO events precluded the use of narrower bins. Low Brier and reliability scores are consistent with well-calibrated models. We also computed the Spiegelhalter Z-test statistic for goodness of fit, which tests for extreme values of the Brier score; p values <0.05 indicate poor calibration.[18]

### Variables retained and importance

We examined the number of predictors or features retained by each algorithm using the full PROMISSE dataset. Given similar performance, a more parsimonious model is clearly preferred over a larger and more complex model for greater ease of use and interpretation. Variable importance was examined for the two tree-based methods (RF and GB), penalised regression (LASSO) and SVM via a permutation-based method that computes the reduction in AUC when each variable is permuted and thus offers no prognostic information.[19] To explore two-way interaction effects, we performed LASSO regression using the expanded dataset with all pairwise interactions.

### Patient and public involvement

This study represents a secondary data analysis of the existing PROMISSE dataset, therefore it was not possible to involve patients or the public in the design, or conduct, or reporting, or dissemination plans of our research.

## RESULTS

Among the 385 patients with SLE in the PROMISSE dataset, 71 (18.4%) experienced an APO. A detailed description and statistical summary of the demographic, clinical and laboratory characteristics of the PROMISSE patients with lupus were previously reported.[1]

Table 1 Model discrimination*

| Model | AUC (95% CI) | Sensitivity† | Specificity† | PPV | NPV |
|---|---|---|---|---|---|
| Regression models | | | | | |
| Stepwise selection (LR-S) | 0.74 (0.68 to 0.82) | 0.64 | 0.79 | 0.41 | 0.91 |
| Penalised (LASSO) | 0.77 (0.71 to 0.83) | 0.67 | 0.78 | 0.41 | 0.91 |
| Neural networks (NN) | | | | | |
| One hidden layer (NN-1) | 0.74 (0.67 to 0.80) | 0.65 | 0.75 | 0.39 | 0.90 |
| Two hidden layers (NN-2) | 0.71 (0.64 to 0.79) | 0.61 | 0.78 | 0.37 | 0.90 |
| Tree-based | | | | | |
| Random forest (RF) | 0.77 (0.71 to 0.83) | 0.75 | 0.71 | 0.37 | 0.93 |
| Gradient boosting (GB) | 0.73 (0.66 to 0.79) | 0.69 | 0.68 | 0.33 | 0.91 |
| Support vector machine (SVM) | | | | | |
| SVM-RBF | 0.77 (0.70 to 0.84) | 0.75 | 0.74 | 0.39 | 0.93 |
| Ensemble | | | | | |
| SuperLearner (SL) | 0.78 (0.72 to 0.84) | 0.71 | 0.77 | 0.41 | 0.92 |

*Average across five independent, 10-fold cross-validations.
†At an optimal cut-point found for each algorithm and iteration.
AUC, area under the curve; LASSO, least absolute shrinkage and selection operator; LR-S, logistic regression wih stepwise selection; NPV, negative predictive value; PPV, positive predictive value.

## Algorithm performance

### Discrimination

SL, the ensemble method that combines predictions from four different algorithms, demonstrated the best overall ability to discriminate APO from non-APO, with an AUC of 0.78 (table 1). LASSO, RF and SVM-RBF showed similar high discrimination performance, with each yielding an AUC of 0.77, followed by LR-S and NN-1 with AUCs of 0.74 for both. NN-2 (AUC=0.71) and GB (AUC=0.73) had the lowest AUCs and were significantly worse than SL (NN-2 vs SL: p=0.001; GB vs SL: p=0.01). The four methods with the highest AUCs in table 1: SL, LASSO, RF and SVM-RBF were statistically indistinguishable.

Sensitivity (proportion of correctly identified APOs) was highest with RF (0.75) and SVM-RBF (0.75), and lowest with the two regression methods (0.64–0.67). The opposite was observed with the results for specificity (the proportion of non-APOs that were correctly identified), which was highest with the regression-based methods (0.78–0.79) and lowest with GB (0.68), RF (0.71) and SVM-RBF (0.74). SL showed reasonable levels of both sensitivity (0.71) and specificity (0.77), which is expected since it had the highest overall AUC. A woman classified as 'high risk' by SL has a 41% predicted probability of having an APO (positive predictive value), while a woman classified as 'low risk' for an APO has a 92% predicted probability of not experiencing an APO (negative predictive value).

### Calibration

The four algorithms that showed the best discrimination performance (SL, LASSO, RF, SVM-RBF) were also the best calibrated based on the Brier and reliability scores (table 2). The ensemble SL method was again the top performer among all algorithms. Calibration was the poorest for standard logistic regression and the two NN approaches; these methods had the worst Brier scores (0.14–0.18), worst reliability (0.013–0.054) and a statistically significant lack of fit (p<0.0001 for each model).

The degree of calibration was further characterised for each algorithm by dividing subjects into four subgroups defined by model predicted APO risk level (≤25%, 26%–50%, 51%–75%, >75%) and evaluating whether the actual proportion of subjects who had an APO in that subgroup (ie, observed APO risk) was within the predicted range for that group. SL is well calibrated because the proportion who had an APO was consistent with the SL predicted risks in all four subgroups. In contrast, LR-S, NN-1 and NN-2 tended to overestimate the APO risk for the subgroup of patients with the higher predicted risks (51%–75%, >75%). In general, agreement between observed and predicted risk was closest for SL, LASSO, RF and SVM-RBF, which is consistent with the results using the other calibration metrics. In women classified as low predicted APO risk (≤25%), the observed rate of APO was consistently low across all algorithms (9%–13%).

### Model size

Among the 41 total predictor variables, LR-S retained an average of 14 variables while LASSO retained 22. The other algorithms considered do not automatically perform model selection and therefore generate APO predictions using all 41 variables.

## Variable importance

Our previously reported APO prediction model using logistic regression with variable selection based on clinical and statistical consideration included LAC positivity,

**Table 2** Model calibration

| Model | Calibration measures* | | | APO predicted risk† | | | |
|---|---|---|---|---|---|---|---|
| | Brier‡ | Reliability‡ | P value§ | 1 (low) | 2 | 3 | 4 (high) |
| Regression models | | | | Actual APO rate* | | | |
| Stepwise-selection (LR-S) | 0.14 | 0.013 | <0.001 | 0.11 | 0.35 | 0.38 | 0.64 |
| Penalised (LASSO) | 0.13 | 0.007 | 0.13 | 0.11 | 0.40 | 0.39 | 0.74 |
| Neural networks (NN) | | | | | | | |
| One hidden layer (NN-1) | 0.16 | 0.033 | <0.001 | 0.12 | 0.27 | 0.40 | 0.49 |
| Two hidden layers (NN-2) | 0.18 | 0.054 | <0.001 | 0.11 | 0.29 | 0.32 | 0.43 |
| Tree-based | | | | | | | |
| Random forest (RF) | 0.13 | 0.004 | 0.55 | 0.09 | 0.35 | 0.47 | 1.00¶ |
| Gradient boosting (GB) | 0.14 | 0.009 | 0.19 | 0.13 | 0.33 | 0.38 | 0.83 |
| Support vector machine (SVM) | | | | | | | |
| SVM-RBF | 0.13 | 0.005 | 0.62 | 0.10 | 0.40 | 0.61 | 0.61 |
| Ensemble | | | | | | | |
| SuperLearner (SL) | 0.12 | 0.003 | 0.82 | 0.09 | 0.40 | 0.60 | 0.75 |

*Average across five independent, 10-fold cross-validations.
†APO predicted risk: 1: ≤25%, 2: 26%–50%, 3: 51%–75%, 4: >75%.
‡Agreement between predicted and observed APO; low scores indicate better agreement.
§P<0.05 indicates lack of fit using Spiegelhalter goodness-of-fit test.
¶Only one individual (who experienced an APO) had a prediction >75%.
APO, adverse pregnancy outcome; LASSO, least absolute shrinkage and selection operator; LR-S, logistic regression wih stepwise selection.

current antihypertensive use, Physician Global Assessment (PGA) score >1, decreasing platelet count and race/ethnicities other than non-Hispanic white as significant risk factors for APO.[1] The utility and stability of these patient characteristics as predictors of APO risk were largely confirmed here, with platelet count, LAC positivity, antihypertensive use and PGA score appearing in the majority, if not all, of the lists of top 10 most important variables identified by the best performing algorithms (figure 1). Although non-Hispanic white was not included among the top 10 for any of the methods, it was in the middle range of ranks for all methods. One of the most consistently important variables, LAC+ was prevalent in 8% of PROMISSE patients. For a 'typical' LAC+ patient (ie, with the other patient characteristics set to the mean values in the study population), the predicted APO risk using LASSO, RF, SVM and SL ranged from 0.46 to 0.54, compared with 0.03–0.11 for a LAC– patient. If both LAC+ and PGA >1 are present, the predicted APO risk across the four algorithms increases to 0.56–0.85. Diastolic blood pressure was identified as one of the top two most important predictors in all algorithms, with higher values associated with an increased risk of APO. PROMISSE patients with baseline diastolic blood pressure above the study median value (67 mm Hg) were observed to have an APO rate of 27%, compared with 9.5% in women with levels below the median. Higher diastolic and systolic values were also associated with antihypertensive use, a previously identified risk factor for APO.

LASSO was applied to analyse the expanded dataset that additionally included all possible pairwise interactions between predictor variables. LASSO was used for this analysis because it addresses the overfitting and multicollinearity that can occur with a large number of predictors by shrinking the regression coefficients of unimportant predictors to zero, thereby producing a simpler and more interpretable model with a reduced set of variables. After diastolic blood pressure, the LAC×aCL IgG interaction was identified as the second most important variable by LASSO, suggesting that the relationship of LAC status with APO risk is modified by aCL IgG status and vice versa. This interaction is also apparent by considering the observed APO rates in the four subgroups defined by LAC and aCL IgG status: 71.4% among LAC+/aCL IgG+, 33.3% among LAC+/aCL IgG–, 7.1% among LAC–/aCL IgG+ and 16.3% among LAC–/aCL IgG–. These results correspond to an OR for the association of LAC status with APO of 32.1 among the aCL IgG+, but 2.6 among the aCL IgG–, indicating that the relationship of LAC and APO, as quantified by the OR, is considerably stronger among the aCL IgG+. When considered alone, aCL IgG status was not among the 10 most important variables identified by any algorithm. In addition, the interactions between C3 and history of pregnancy morbidity, and hydroxychloroquine use and SSB (La) antibodies were among the top predictors identified in the expanded dataset. It is important to note that while using the expanded dataset did not improve the overall performance of LASSO to predict

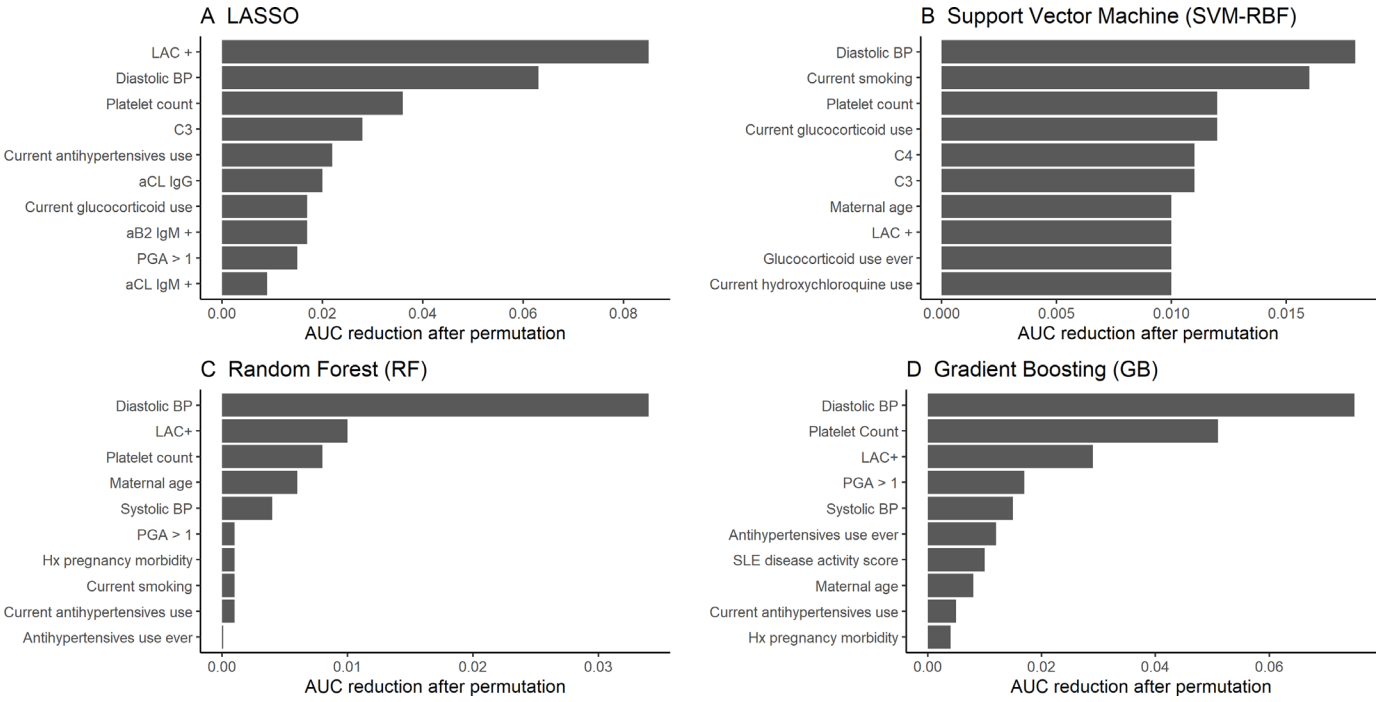


**Figure 1** Bar graph of the top 10 predictors of adverse pregnancy outcomes (APO) using the PROMISSE dataset for (A) penalised regression (least absolute shrinkage and selection operator (LASSO)), (B) support vector machine (SVM-RBF), (C) random forest (RF) and (D) gradient boosting (GB). Variables are each ranked by the average reduction in area under the receiver operating characteristic curve (AUC) after 10 permutation iterations. aCL, anticardiolipin; BP, blood pressure; Hx, history; LAC, lupus anticoagulant.

APO (AUC=0.74, 95% CI: (0.67 to 0.81)) compared with the original dataset, clinically interesting and novel interactions were discovered when these terms were explicitly included in the model.

## DISCUSSION

Development of accurate and validated models for predicting health outcomes in individual patients is of increasing interest in many disease areas because of their potential use for devising more effective treatment strategies tailored to specific risk profiles, and for identifying high-risk patients who would benefit from enrolment in clinical trials of new therapies. Different statistical and ML approaches for predicting clinical outcomes have been proposed and applied in various settings, but no method has shown to perform consistently better than others in all circumstances. In addition, the methods have different pros and cons with respect to interpretability of results, ease of implementation of the algorithm and information generated about the underlying relationships between predictors and outcomes.

Our primary question here was whether APO prediction in patients with lupus using data obtained in the first trimester and standard logistic regression could be improved by applying more flexible ML methods. The usual logistic regression approach was found to substantially overestimate the risk of APO for certain subgroups of patients. This lack of calibration could be an important weakness if the intended use of the model is to accurately estimate the probability that an individual patient will develop APO, not just to differentiate between patients who will and will not have the event. LASSO and ML methods showed superior calibration and discrimination compared with standard logistic regression.

However, despite the good performance of several of the ML approaches (RF, SVM-RBF, SL), a major disadvantage is that unlike regression-based methods, that clearly specify the mathematical relationship between the predictor variable and the outcome, these black box ML approaches do not provide any quantitative measures of the magnitude of the associations of individual factors with APO risk. While the factors can be ranked by variable importance using the permutation-based approach we applied here, estimates of the actual degree to which a predictor increases or decreases risk are essential if a goal is to better understand the processes and mechanisms underlying the development of APO and to assess the potential efficacy of risk mitigation strategies targeting those risk factors. ML methods also do not perform variable selection automatically and require the complete set of 41 patient variables to generate predictions of risk without further modification, which makes them less practical to implement in the real world, for example, as an online risk calculator.

LASSO, a regression-based approach that uses 'shrinkage' to reduce model complexity and prevent overfitting, performed better than standard logistic regression with stepwise selection and similarly to ML

methods in predicting APO. Our findings are consistent with recent[20–23] and past studies,[24] demonstrating that in many clinical applications, the risk of the outcome can be adequately specified as a simple additive function of clinical features using a regression framework; more complicated modelling approaches are not necessarily better at finding novel relationships, fitting the data at hand or generalising to different patient populations.[25] The reasons for this may be related to pathological mechanism or specific characteristics of the datasets, such as sample size. Regardless, the strong relative performance of LASSO compared with the other ML methods is notable, given the advantages of former approach: greater interpretability and more insights about important predictor variables, potentially fewer input variables needed to generate future predictions and potentially more practical implementation as a risk calculator.

A limitation of this study is that PROMISSE study examined singleton pregnancy outcomes in women with mild or moderate SLE disease at conception and without hypertension and diabetes; predictive performances of the algorithms considered may not generalise to a higher risk population of women with more severe disease or comorbidities. Our sample size is also relatively small, limiting the ability to detect complex patterns in the data even with the powerful ML algorithms, and to precisely estimate model performance. An important next step is to externally validate these results using data from independent SLE pregnancy cohorts that we are in the process of obtaining. Finally, we defined APO as a composite of several different outcomes; future studies with a larger number of APO events would enable exploration of event-specific risk factors and prediction.

We were able to predict APO events with reasonable accuracy using SLE patient characteristics that are routinely assessed prior to the 12th week of pregnancy. The most important predictors consistently observed across the majority of algorithms included LAC positivity, PGA score, diastolic blood pressure, current antihypertensives use and platelet count, which both confirms our previously reported APO risk factors and identifies additional risk factors that may also be important for APO prediction. Subtle increases in first trimester diastolic blood pressure and/or the requirement of antihypertensives may reflect endothelial dysfunction and increased vulnerability to antiangiogenic factors produced by the placenta. We also obtained preliminary findings about interactions between predictor variables that should be confirmed in future studies. The interaction between LAC+ and aCL IgG is particularly interesting, because retrospective studies in patients with primary APS have shown that positivity for multiple APLs (LAC, aCL and anti-β2GPI) is associated with higher risk for APO than single positive.[26–29] Identification of an interaction effect using LASSO suggests enhanced APO risk due to LAC when aCL IgG is also present. That LAC and another APL increases probability of pregnancy morbidity beyond that of LAC alone is supported by the EUREKA algorithm.[30] EUREKA revealed that LAC and anti-β2GPI IgG provided the highest risk for APO. In contrast to the current report, EUREKA included pregnancy loss <10 weeks as an outcome and only 17% of the patients had SLE.

Finally, it should be emphasised that the PROMISSE dataset contains all currently known real-world patient characteristics and risk factors associated with adverse pregnancy outcomes in patients with SLE. Using these features as input variables, the maximum AUC that was achieved with the algorithms we evaluated was nearly 0.8, which is generally considered moderate to excellent performance for discrimination.[31] Furthermore, this AUC is similar in magnitude to reported AUCs of other published predictive models for various outcomes in patients with SLE.[2 3 32–34] In our view, further improvement beyond this level in the ability to identify patients with SLE at high risk of APO will likely be realised not through increasingly complicated and opaque algorithms or more complex mathematical formulations of existing predictors, but rather by conducting additional research to discover new biomarkers and risk factors for APO.

**Acknowledgements** We are grateful to PROMISSE investigators for recruiting and caring for study patients. Portions of this manuscript were presented at ACR Convergence (November 2021; virtual): Fazzari M, Guerra M, Salmon J, Kim M. Predicting adverse pregnancy outcomes in women with systemic lupus erythematosus: a comparison of machine learning methods (abstract). *Arthritis Rheumatol* 2021; 73 (suppl 10).

**Contributors** MYK, MJF and JS conceptualised the study questions. JS and MMG acquired the data. MJF and MYK carried out the statistical analysis and wrote the initial manuscript draft. MYK supervised all stages of this study and is the guarantor of this project. All authors provided key contribution to the study and manuscript and approved the final draft of this manuscript.

**Funding** This work was supported by the National Institute of Arthritis and Musculoskeletal and Skin Diseases at the National Institutes of Health under award number AR076612 to JS and MYK, and a Lupus Foundation of America research grant to JS.

**Competing interests** JS is an Associate Editor for Lupus Science & Medicine.

**Patient and public involvement** Patients and/or the public were not involved in the design, or conduct, or reporting, or dissemination plans of this research.

**Patient consent for publication** Not applicable.

**Ethics approval** Institutional review boards approved protocols and consent forms, and written, informed consent was obtained from all patients.

**Provenance and peer review** Not commissioned; externally peer reviewed.

**Data availability statement** Data are available on reasonable request. Clinical datasets from PROMISSE will be made available to the community pursuant to written request in the form of a one page protocol as well as an NIH CV. Such inquiries will be reviewed by the Steering Committee. If the majority of the Steering Committee considers the project meritorious, approved investigators will be provided with de-identified data in Excel-compatible and SAS-compatible formats. Investigators with whom data are shared will be encouraged to acknowledge the source of the data.

**ORCID iD**

Melissa J Fazzari http://orcid.org/0000-0002-5674-3589

## REFERENCES

1 Buyon JP, Kim MY, Guerra MM, *et al*. Predictors of pregnancy outcomes in patients with lupus: a cohort study. *Ann Intern Med* 2015;163:153–63.
2 Kegerreis B, Catalina MD, Bachali P, *et al*. Machine learning approaches to predict lupus disease activity from gene expression data. *Sci Rep* 2019;9:9617.
3 Chen Y, Huang S, Chen T, *et al*. Machine learning for prediction and risk stratification of lupus nephritis renal flare. *Am J Nephrol* 2021;52:152–60.
4 Hochberg MC. Updating the American College of rheumatology revised criteria for the classification of systemic lupus erythematosus. *Arthritis Rheum* 1997;40:40.
5 Lockshin MD, Kim M, Laskin CA, *et al*. Prediction of adverse pregnancy outcome by the presence of lupus anticoagulant, but not anticardiolipin antibody, in patients with antiphospholipid antibodies. *Arthritis Rheum* 2012;64:2311–8.
6 Buuren Svan, Groothuis-Oudshoorn K. mice : Multivariate Imputation by Chained Equations in *R*. *J Stat Softw* 2011;45:1–67.
7 Miller AJ. *Subset selection in regression*. London: Chapman and Hall, 1990.
8 Tibshirani R. Regression shrinkage and selection via the LASSO. *J R Stat Soc Series B Stat Methodol* 1996;58:267–88.
9 Ripley BD. *Pattern recognition and neural networks*. Cambridge university press, 2007.
10 Cristianini N, Shawe-Taylor J. *An introduction to support vector machines and other kernel-based learning methods*. Cambrigde: Cambridge University Press, 2000. www.support-vector.net
11 Breiman L. Random forests. *Mach Learn* 2001;45:5–32.
12 Friedman JH. Greedy function approximation: a gradient boosting machine. *Ann Stat* 2001;29:1189–232.
13 van der Laan MJ, Polley EC, Hubbard AE. Super learner. *Stat Appl Genet Mol Biol* 2008;6.
14 R Core Team. R: a language and environment for statistical computing. Vienna, Austria R Foundation for Statistical Computing; 2021. https://www.R-project.org/
15 edHastie T, Tibshirani R, Jerome F. The Elements of Statistical Learning: Data Mining, Inference, and Prediction.. In: *2Nd ED*. New York: Springer, 2009.
16 Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 1982;143:29–36.
17 Steyerberg EW. *Clinical prediction models*. New York: Springer, 2009.
18 Spiegelhalter DJ. Probabilistic prediction in patient management and clinical trials. *Stat Med* 1986;5:421–33.
19 Biecek P. DALEX: Explainers for complex predictive models in R. *JMLR* 2018;19:1–5 https://jmlr.org/papers/v19/18-416.html
20 Engelhard MM, Navar AM, Pencina MJ. Incremental benefits of machine Learning-When do we need a better Mousetrap? *JAMA Cardiol* 2021;6:621–3.
21 Khera R, Haimovich J, Hurley NC, *et al*. Use of machine learning models to predict death after acute myocardial infarction. *JAMA Cardiol* 2021;6:633–41.
22 Roberts M, Driggs D, Thorpe M, *et al*. Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. *Nat Mach Intell* 2021;3:199–217.
23 Christodoulou E, Ma J, Collins GS, *et al*. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *J Clin Epidemiol* 2019;110:12–22.
24 Dudoit S, Fridlyand J, Speed TP. Comparison of discrimination methods for the classification of tumors using gene expression data. *J Am Stat Assoc* 2002;97:77–87.
25 Rudin C, Radin J. Why are we using black box models in AI when we don't need to? A lesson from an explainable AI competition. *Harv Data Sci Rev* 2019;1.
26 Saccone G, Berghella V, Maruotti GM, *et al*. Antiphospholipid antibody profile based obstetric outcomes of primary antiphospholipid syndrome: the PREGNANTS study. *Am J Obstet Gynecol* 2017;216:525.e1–525.e12.
27 Lazzaroni M-G, Fredi M, Andreoli L, *et al*. Triple antiphospholipid (aPL) antibodies positivity is associated with pregnancy complications in aPL carriers: a multicenter study on 62 pregnancies. *Front Immunol* 2019;10:1948.
28 Pengo V, Ruffatti A, Del Ross T, *et al*. Confirmation of initial antiphospholipid antibody positivity depends on the antiphospholipid antibody profile. *J Thromb Haemost* 2013;11:1527–31.
29 Galli M, Luciani D, Bertolini G, *et al*. Lupus anticoagulants are stronger risk factors for thrombosis than anticardiolipin antibodies in the antiphospholipid syndrome: a systematic review of the literature. *Blood* 2003;101:1827–32.
30 Pregnolato F, Gerosa M, Raimondo MG, *et al*. Eureka algorithm predicts obstetric risk and response to treatment in women with different subsets of anti-phospholipid antibodies. *Rheumatology* 2021;60:1114–24.
31 Hosmer WD, Lemeshow S, Sturdivant RX. *Applied logistic regression*. 398. John Wiley & Sons, 2013.
32 Ceccarelli F, Sciandrone M, Perricone C. Prediction of chronic damage in systemic lupus erythematosus by using machine-learning models. *PLoS One* 201;7:e0174200.
33 Reddy BK, Delen D. Predicting hospital readmission for lupus patients: an RNN-LSTM-based deep-learning methodology. *Comput Biol Med* 2018;101:199–209.
34 Huang T, Liu S, Huang J, *et al*. Prediction and associated factors of hypothyroidism in systemic lupus erythematosus: a cross-sectional study based on multiple machine learning algorithms. *Curr Med Res Opin* 2022;38:229–35.