Predicting Serious Events Post COVID-19 Vaccination via Machine Learning: Comparing Applicability of Logistic LASSO Regression and Random Forest Algorithms

Abstract: The COVID-19 pandemic has been ongoing since the end of 2019, affecting millions of lives, and vaccines did not start rolling out until 2021. However, many are hesitant on getting vaccinated, leading to propagation of anti-vaccination sentiment. To combat these sentiments, a machine learning algorithm was developed to determine whether an adverse event is more likely to occur for individuals with certain characteristics who get the vaccine. Logistic LASSO regression and random forest were utilized on VAERS data to assess which algorithm would be better suited to determine if an individual could require additional monitoring and care. Our analyses showed that both logistic LASSO regression and random forest achieved overall respectable accuracies due to strong specificities. The results suggest that it is possible to use patient's medical information to predict whether getting the vaccine would require additional monitoring and care.

Introduction

Long-term management of COVID-19 is in part reliant on COVID-19 vaccinations alongside mask wearing and strategic social distancing measures. However, in and outside of the United States, experts acknowledge there has seen an increasing trend of overall vaccine hesitancy even prior to the COVID-19 pandemic[1]. Anti-vaccination sentiments have been propagated in more ways than one which has complicated countering fears of vaccines. For example, websites with misinformation about vaccines have and still play a vital role in cultivating vaccine fears[2]. Social media platforms like YouTube whose primary goal isn't to disseminate vaccine misinformation, also play a role in creating pipelines to vaccine hesitancy and anti-vaccination sentiments because of their search and recommendation algorithms[3]. Thus, scientists' and physicians' role in finding innovative ways to counter misinformation becomes imperative.

Plenty of research has been done on what can be done to increase the number of vaccinations taken by populations. Consistent government messaging, culturally targeted communication, vaccine development transparency, and a strong campaign for why a vaccine is necessary are just some of several factors associated with combating COVID-19 vaccine hesitancy globally [4,5]. In the United States, the COVID States Project surveyed 25,640 individuals across the 50-states and the District of Columbia regarding COVID-19 to investigate messaging strategies like a famous figure (politician, celebrity, or athlete) promoting COVID-19 vaccination or randomized messaging that centered around "patriotic duty" or endorsement from one's personal physician[6]. Findings showed that vaccine promotion via a famous figure did not always yield positive results as "partisan" figures catalyze resistance against COVID-19 vaccination[6]. Currently in pre-print is an article titled *Adverse effects of COVID-19 vaccination: machine learning and statistical approach to identify and classify incidences of morbidity and post-vaccination reactogenicity* by Ahamad et al. that implements machine learning[8]. Specifically, random forest (RF), support vector machine (SVM), and gradient boosting machine (GBM) was used for the purpose of identifying individuals that may need additional monitoring and care. Ahamad et al's analysis

flow was to first identify significant features and then determine an algorithm to distinguish between various groups of patients for various outcomes like mortality, COVID-19 positive patients, or hospitalization.

Here we use a binary variable labeled 'serious event' as our outcome, where 1 indicates whether any one of various possible serious events has occurred and 0 indicates no serious event has taken place post-vaccination. The purposes of this study were to: 1) develop and test the accuracy of a logistic LASSO regression (LLR) and RF algorithm to determine whether a serious event is likely to occur, and 2) to assess which of the two algorithms would be better applicable to determine whether an individual could require additional monitoring and care.

Methods

Data Source

We analyzed COVID-19 vaccine domestic reports received by VAERS from 12/14/2020, through 4/5/2021. This voluntary, national, passive surveillance system was established in 1990 and is operated jointly by the US Food and Drug Administration (FDA) and the Centers for Disease Control and Prevention (CDC). Our analysis included only US reports to VAERS, which usually have more complete information and more feasible follow-up review of medical records than foreign reports.

Outcome definition

Our primary outcome is serious event according to the FDA regulatory definition (21CFR§314.80) of post-marketing reporting of adverse drug event. The serious event outcome is a composite endpoint including death, life threatening event, hospitalization or prolongation of existing hospitalization, disability.

Predictors

Predictors include age, gender, COVID-19 vaccine type (MODERNA, PFIZER/BIONTECH, JANSSEN), prior drug, food, or other allergy, prior vaccine adverse event report, and baseline diseases extracted from the "CUR_ILL" variable, including hypertension, diabetes, chronic obstructive pulmonary disease, cancer, myocardial infarction, stroke, and heart failure.

Exclusion Criteria

To exclude potential reporting bias and confounding, we excluded the following patients: 1) had an adverse event onset date before vaccine; 2) had > 2 doses of a COVID-19 vaccine; 3) in hospice care; 4) had a missing COVID-19 vaccination type; 5) had a prior COVID-19 diagnosis before vaccination. Additionally, vaccination may not have been a necessary cause of all adverse events in the VAERS data. We therefore excluded 6,672 individuals (18% of those remaining) who had more than four days between their recorded vaccination and event date

Evaluation Criteria

R version 4.0.3 (R Foundation for Statistical Computing, Vienna, Austria) was the programming language used to conduct all analyses[9]. To assess the performance of algorithms, we used a 5-fold cross validation procedure where individuals vaccinated are randomly allocated to 1 of 5 independent folds of approximately equal size. Training and testing are iterated over such that each individual fold was treated as a testing dataset on its own once, while the remaining 4 folds are treated as the training dataset. Values used to each assess each model include accuracy, sensitivity, and specificity. Accuracy was calculated by summing the number of correctly predicted *No Serious Events* and *Serious Events* divided by the total number of events. Sensitivity was calculated by taking the total number of correctly predicted *Serious Events* and dividing by the total number of *Serious Events*. Meanwhile, specificity was calculated by taking the total number of correctly predicted *No Serious Events* and dividing by the total number of *No Serious Events*.

Random Forest

The RF algorithm was fit using the R package "caret" version 6.0-86[10]. A tuning grid consisting of the number of trees generated (*B;* 250, 500, 750) and the subset number of predictors considered at each split (*m;* 4, 6, 8, 12, or 16) was used. The combination of these parameters that minimized out-of-bag error rate was selected and used to train a final RF in the training set, before applying it to the final fold (testing set) to estimate algorithm performance.

Logistic LASSO Regression

The LLR algorithm was fit using the R package "glmnet" version 4.1-1[11]. The main effects of all predictors were included alongside all two-way interactions between predictors that were continuous or categorical with ≥150 observations at each level (i.e., age, vaccine manufacturer, number of doses administered, allergies, prior vaccine-related adverse event, and prevalent hypertension or diabetes). Interaction terms for categorical predictors with <150 observations in a given level were unable to be computed. Tuning was done using a 10-fold cross-validation procedures, such that the 4 training folds were combined and reallocated randomly to 10 folds to tune the functional form used to model age and the LASSO shrinkage penalty λ[12]. We tested linear/polynomial, B-spline, and cubic restricted forms for our only continuous variable age using degrees 1 through 4 for the polynomials and B-splines. 1 through 4 knots were considered for B-splines and the cubic restricted splines. The default tuning grid for λ provided by the "cv.glmnet" function was used[11]. We then used training set predicted probabilities to select a probability threshold for classifying *Serious* Event through Receiver Operator Curve (ROC) analysis using the "pROC" R package[13]. We selected the threshold that maximized the following weighted version of Youden's *J* statistic:

$$J = w_{sp} * \text{Specificity} + w_{se} * \text{Sensitivity} - 1$$

Where w_{sp} and w_{se} weight the specificity and sensitivity, respectively, and have a mean value of 1. Two thresholds were selected with this approach, and both were evaluated separately: the first with both weights set to 1 (LLR), and the second with a w_{sp} of 1.2 and w_{se} of 0.8 (LLR_{sp}). The former set of weights provide values equivalent to the original Youden's J statistic and the latter

set, where specificity was upweighted at the cost of sensitivity, is consistent with the costs attributed to false negatives versus false positives more generally with screening tools[14]. In this context, where serious events from vaccination are rare, increasing specificity can prevent a substantial number of false positives and attenuate the potential harms associated with vaccine hesitancy. Thus, LLR and LLR_{sp} only differ in how their probability thresholds are chosen.

Results

Data

After applying exclusion criteria, a total of 36812 patients are included in our analysis. To demonstrate the distribution of covariates, we stratify the population into 4 groups by the most important effect modifier, age: 1) 18 – 64 years old; 2) 65 – 74 years old; 3) 75-84 years old; 4) over 85 years old. Female patients are more likely to receive the vaccine, especially in the younger population group (80% female in patients under 64 years old vs 55% female patients over 75 years old). The younger population are more likely to have prior drug allergies and reported adverse events. Meanwhile, the older population are more likely to have a late onset of adverse events (51% in age under 64 years old vs 23% in age 85+ group had event on the same day as vaccination). The younger population are more likely to recover from the adverse events (43% in age under 65 years old vs 16% in age 85+ group) and the older population are more likely to experience a serious adverse event and have baseline diseases.

Analysis

Table 2 and 3 show the results for the average accuracies, specificities, and sensitivities for each algorithm method obtained from the 5-fold cross validation and nested tuning of parameters of the analysis. The method of algorithms with the highest accuracy value was RF with a mean accuracy of 0.884 and an accuracy standard error of 0.019. However, when comparing the methods LLR, LLR_{sp}, and RF the average sensitivity was greater for the LLR method at 0.694 (standard error 0.012) opposed to the average sensitivities of the LLR_{sp} and RF method 0.620 (standard error 0.017) and 0.255 (standard error 0.022), respectively. The average specificities for the LLR, LLR_{sp}, and RF methods are all greater than 0.800.

Notably for each of the LLR and LLR_{sp} method folds the basic spline was chosen with degrees of freedom ranging from 5 to 8 and degrees ranging from 2 to 4. Furthermore, the mean shrinkage penalty λ was 5.40E-4. The probability thresholds for the equally weighted sensitivity and specificity averaged at 0.133. On the other hand, the probability thresholds for the upweighted specificity at the cost of sensitivity averaged at 0.191.

For the RF method folds the number of trees chosen to be generated included 250, 500, and 750, while the number of variables to be considered at each node were 4 and 6. Table 4 provides insight on which variables were most important among the RF algorithms. Notably the variable age was determined to be most important based on its error education weighted by the number observations at the node it appears at. The variable age is followed by sex and vaccine manufacturer in terms of importance. However, it should be noted that age is our only continuous variable and RF is inclined to inflate the relevance of continuous predictors. Figure 1

visualizes the accuracies, sensitivities, specificities, and corresponding standard errors by algorithm method. Figure 2 provides a visualization of variable importance averaged over the 5 folds.

Discussion

The public health importance of combating vaccine hesitancy has been reinforced multiple times in the past with events like the H1N1 pandemic, the U.S. Measles outbreak, and the ongoing anti-vaccination movement[15,16,17]. Research on methods to implement communication that improves vaccination uptakes has been done, however, vast number of websites and social media platforms have added a dimension of complication[1,2,3,4,5,6,7]. Thus, researchers should evolve their methods for reassuring individuals hesitant about getting their vaccine. Machine learning algorithms are one avenue that may be useful if the right algorithms can be developed and framed correctly to the public.

The study conducted here is one of a couple of studies to train machine learning algorithms on COVID-19 vaccination data and test their accuracy in predicting whether an individual getting their vaccine should be provided additional monitoring and care. Our results demonstrated that all algorithm methods (LLR, LLR_{sp}, and RF) achieved overall respectable accuracies largely because of strong specificities. Sensitivities, however, were not as strong and the averaged sensitivity for the RF method was particularly poor. Although not directly comparable due to a difference in outcomes, Ahamad et al found similar results regarding accuracy using RF and XGB methods when predicting patient *death status*, *SARS-CoV-2 test status*, and *hospital admission status*[8].

Although we have taken many precautions to ensure that the *Serious Event* outcome related back to the COVID-19 vaccine by using a subset of VAERS data that only took into consideration *Serious Events* at most 4 days after taking the COVID-19 vaccine, there is no guarantee there is an association. Furthermore, the nature of the VAERS data is such that reports are submitted voluntary and subject to reporting bias, so a limitation of the data is that it should only be used for hypothesis generation[18,19,20]. Other limitations of the VAERS data include incompleteness with regards to reports and a lack of an unvaccinated comparison group[18]. Limitations regarding the analyses is our outcome *Serious Event* which is an indicator based on an agglomeration of possible serious events that could occur post-vaccination. Thus, our algorithms are not specific to a particular event and consequently makes interpretation about the outcome unspecific as well.

In conclusion, the results of this study suggest that it is possible to use medical information to predict whether an individual getting their vaccine should get additional monitoring and care. Further investigation would need to be conducted using data that is not subject to reporting bias and is well defined with regards to what is reported. Ultimately, this analysis only represents a first step towards developing a tool to combat vaccine hesitancy by assuring individuals that they will be provided the right type of care. How to frame the algorithm as a helping tool that does not potentially reinforce vaccine hesitancy should also further be investigated.

References

1. Dubé, E., Laberge, C., Guay, M., Bramadat, P., Roy, R., & Bettinger, J. A. (2013). Vaccine hesitancy: an overview. *Human vaccines & immunotherapeutics*, 9(8), 1763-1773.
2. Bean, S. J. (2011). Emerging and continuing trends in vaccine opposition website content. Vaccine, 29(10), 1874-1880.
3. Tang, L., Fujimoto, K., Amith, M. T., Cunningham, R., Costantini, R. A., York, F., ... & Tao, C. (2021). "Down the Rabbit Hole" of Vaccine Misinformation on YouTube: Network Exposure Study. *Journal of Medical Internet Research*, *23*(1), e23262.
4. Lazarus, J. V., Ratzan, S. C., Palayew, A., Gostin, L. O., Larson, H. J., Rabin, K., ... & El-Mohandes, A. (2021). A global survey of potential acceptance of a COVID-19 vaccine. *Nature medicine*, *27*(2), 225-228.
5. Thomson, A., Vallee-Tourangeau, G., & Suggs, L. S. (2018). Strategies to increase vaccine acceptance and uptake: From behavioral insights to context-specific, culturally-appropriate, evidence-based communications and interventions. *Vaccine*, *36*(44), 6457-6458.
6. Chou, W. Y. S., & Budenz, A. (2020). Considering Emotion in COVID-19 vaccine communication: addressing vaccine hesitancy and fostering vaccine confidence. *Health communication*, 35(14), 1718-1722.
7. Green, J., Lazer, D., Baum, M., Druckman, J., Uslu, A., Simonson, M., ... & Quintana, A. (2021). The COVID States Project# 36: Evaluation of COVID-19 vaccine communication strategies. *
8. Ahamad, M. M., Aktar, S., Uddin, M. J., Rashed-Al-Mahfuz, M., Azad, A. K. M., Uddin, S., ... & Moni, M. A. (2021). Adverse effects of COVID-19 vaccination: machine learning and statistical approach to identify and classify incidences of morbidity and post-vaccination reactogenicity. MedRxiv. *
9. R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.
10. Kuhn, M. (2020). caret: Classification and Regression Training. R package version 6.0-86. URL: https://CRAN.R-project.org/package=caret
11. Friedman, J., Hastie, T., Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. Journal of Statistical Software, 33(1), 1-22. URL https://www.jstatsoft.org/v33/i01/.
12. Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media.
13. Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J. C., & Müller, M. (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC bioinformatics*, *12*(1), 1-8.
14. Westreich, D. (2019). *Epidemiology by design: a causal approach to the health sciences*. Oxford University Press.
15. Larson, H. J., Jarrett, C., Schulz, W. S., Chaudhuri, M., Zhou, Y., Dube, E., ... & Wilson, R. (2015). Measuring vaccine hesitancy: the development of a survey tool. *Vaccine*, *33*(34), 4165-4175.
16. Peretti-Watel, P., Larson, H. J., Ward, J. K., Schulz, W. S., & Verger, P. (2015). Vaccine hesitancy: clarifying a theoretical framework for an ambiguous notion. *PLoS currents*, *7*.
17. Hussain, A., Ali, S., Ahmed, M., & Hussain, S. (2018). The anti-vaccination movement: a regression in modern medicine. *Cureus*, *10*(7).

18. Chen, R. T., Rastogi, S. C., Mullen, J. R., Hayes, S. W., Cochi, S. L., Donlon, J. A., & Wassilak, S. G. (1994). The vaccine adverse event reporting system (VAERS). Vaccine, 12(6), 542-550.
19. Shimabukuro, T. T., Nguyen, M., Martin, D., & DeStefano, F. (2015). Safety monitoring in the Vaccine Adverse Event Reporting System (VAERS). *Vaccine*, 33(36), 4398–4405. https://doi.org/10.1016/j.vaccine.2015.07.035
20. *VAERS Data: Guide to Interpreting VAERS Case Report Information.* Available at: https://vaers.hhs.gov/data/index. Accessed April 24, 2021.

Tables and Figures

TABLES

TABLE 1

Variable	N	**18-64**, N = 29,158[1]	**65-74**, N = 3,925[1]	**75-84**, N = 2,325[1]	**85+**, N = 1,421[1]	**p-value**[2]
Age (yrs)	36,829	42 (12)	69 (3)	79 (3)	90 (4)	<0.001
Sex	36,829					<0.001
Female		23,207 (80%)	2,530 (64%)	1,289 (55%)	785 (55%)	
Male		5,813 (20%)	1,382 (35%)	1,023 (44%)	625 (44%)	
Unknown		138 (0.5%)	13 (0.3%)	13 (0.6%)	11 (0.8%)	
COVID19 vaccine type	36,829					<0.001
MODERNA		11,799 (40%)	1,880 (48%)	1,141 (49%)	709 (50%)	
PFIZER\BIONTECH		13,858 (48%)	1,697 (43%)	1,108 (48%)	688 (48%)	
JANSSEN		3,501 (12%)	348 (8.9%)	76 (3.3%)	24 (1.7%)	
Prior drug, food, or other allergy	36,829	11,850 (41%)	1,615 (41%)	885 (38%)	504 (35%)	<0.001
Prior Vax Adverse Event	36,829	1,540 (5.3%)	234 (6.0%)	60 (2.6%)	27 (1.9%)	<0.001
Covid-19 prior to Vaccine	36,829	0 (0%)	0 (0%)	0 (0%)	0 (0%)	
Number of symptom latent days (Onset date - Vaccine date)	36,829					<0.001
0		14,769 (51%)	1,480 (38%)	730 (31%)	330 (23%)	
1		7,052 (24%)	926 (24%)	551 (24%)	320 (23%)	
2		1,428 (4.9%)	312 (7.9%)	170 (7.3%)	140 (9.9%)	
>=3		5,909 (20%)	1,207 (31%)	874 (38%)	631 (44%)	

Anaphylactic reaction	36,829	221 (0.8%)	17 (0.4%)	6 (0.3%)	3 (0.2%)	<0.001
Anaphylaixs event with prior allergy	36,829	115 (0.4%)	7 (0.2%)	2 (<0.1%)	2 (0.1%)	0.006
Emergency room or doctor visit	36,829					0.5
N		29,138 (100%)	3,924 (100%)	2,325 (100%)	1,420 (100%)	
Y		20 (<0.1%)	1 (<0.1%)	0 (0%)	1 (<0.1%)	
Recovered	36,829					<0.001
Yes		12,405 (43%)	1,355 (35%)	619 (27%)	232 (16%)	
No		10,269 (35%)	1,692 (43%)	1,071 (46%)	765 (54%)	
Unknown		6,484 (22%)	878 (22%)	635 (27%)	424 (30%)	
Composite endpoint: serious event	36,829	2,486 (8.5%)	1,248 (32%)	1,251 (54%)	1,022 (72%)	<0.001
Died	36,829					<0.001
N		28,900 (99%)	3,566 (91%)	1,878 (81%)	841 (59%)	
Y		258 (0.9%)	359 (9.1%)	447 (19%)	580 (41%)	
Life-threatening illness	36,829					<0.001
N		28,490 (98%)	3,698 (94%)	2,157 (93%)	1,325 (93%)	
Y		668 (2.3%)	227 (5.8%)	168 (7.2%)	96 (6.8%)	
Hospitalized	36,829					<0.001
N		27,522 (94%)	3,138 (80%)	1,489 (64%)	902 (63%)	
Y		1,636 (5.6%)	787 (20%)	836 (36%)	519 (37%)	

Prolongation of existing hospitalization	36,829					<0.001
N		29,135 (100%)	3,915 (100%)	2,319 (100%)	1,417 (100%)	
Y		23 (<0.1%)	10 (0.3%)	6 (0.3%)	4 (0.3%)	
Disability	36,829					<0.001
N		28,735 (99%)	3,748 (95%)	2,213 (95%)	1,361 (96%)	
Y		423 (1.5%)	177 (4.5%)	112 (4.8%)	60 (4.2%)	
Number of days hospitalized	2,644	3 (3)	4 (4)	4 (4)	4 (3)	<0.001
Unknown		27,927	3,381	1,773	1,104	
Congenital anomaly/birth defect	36,829	25 (<0.1%)	0 (0%)	0 (0%)	0 (0%)	0.11
Prevalent HTN	36,829	158 (0.5%)	65 (1.7%)	56 (2.4%)	60 (4.2%)	<0.001
Prevalent diabetes	36,829	147 (0.5%)	69 (1.8%)	62 (2.7%)	39 (2.7%)	<0.001
Prevalent COPD	36,829	20 (<0.1%)	33 (0.8%)	32 (1.4%)	24 (1.7%)	<0.001
Prevalent cancer	36,829	31 (0.1%)	30 (0.8%)	25 (1.1%)	19 (1.3%)	<0.001
History of MI	36,829	0 (0%)	1 (<0.1%)	0 (0%)	4 (0.3%)	<0.001
History of stroke	36,829	5 (<0.1%)	9 (0.2%)	12 (0.5%)	4 (0.3%)	<0.001
Prevalent heart failure	36,829	8 (<0.1%)	14 (0.4%)	24 (1.0%)	32 (2.3%)	<0.001

[1] Mean (SD); n (%)

[2] Kruskal-Wallis rank sum test; Pearson's Chi-squared test; Fisher's exact test

TABLE 2 Averaged accuracy, specificity, and sensitivity across 5-folds

Method	Accuracy	Specificity	Sensitivity
LASSO	0.789 (0.019)	0.804 (0.021)	0.694 (0.012)
$LASSO_{sp}$	0.834 (0.013)	0.867 (0.015)	0.620 (0.017)
Random Forest	0.884 (0.003)	0.979 (0.004)	0.255 (0.022)

TABLE 3 Averaged per-class accuracies and overall accuracies across 5-folds

Method	Event	N	N Correct	N Incorrect	% Correct (SE)	% Overall accuracy (SE)
LASSO	No	26,187	21,045	5,142	80.4 (2.1)	78.9 (1.9)
LASSO	Yes	3,970	2,754	1,216	69.4 (1.2)	
$LASSO_{sp}$	No	26,187	22,703	3,484	86.7 (1.5)	83.4 (1.3)
$LASSO_{sp}$	Yes	3,970	2,457	1,513	62.0 (1.7)	
Random Forest	No	26,187	25,638	549	97.9 (0.4)	88.4 (0.3)
Random Forest	Yes	3,970	1,014	2,956	25.5 (2.2)	

TABLE 4 Average and fold-specific variable importance in random forest

Predictor	Mean (SE)	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Age (years)	100.00 (0.00)	100.00	100.00	100.00	100.00	100.00
Sex	12.45 (0.87)	13.86	12.31	12.46	12.11	11.52
Vaccine manufacturer	12.23 (0.35)	11.98	12.50	12.16	12.67	11.84
Second vaccine dose	9.17 (0.74)	9.50	8.59	9.60	9.96	8.21
Prior allergy	4.29 (0.59)	3.79	4.96	3.82	3.97	4.91
Prevalent HTN	2.85 (0.35)	3.17	2.51	3.28	2.74	2.57
Prevalent diabetes	2.64 (0.25)	2.35	2.53	3.03	2.68	2.64
Prior vaccine adverse event	1.82 (0.47)	1.55	2.27	1.42	1.48	2.39
Prevalent COPD	1.82 (0.29)	1.76	1.69	2.30	1.80	1.53
Prevalent heart failure	1.80 (0.32)	2.11	1.49	2.06	1.92	1.43
Prevalent cancer	1.36 (0.27)	1.16	1.72	1.13	1.23	1.59
On dialysis	0.84 (0.13)	0.95	0.77	0.89	0.95	0.64
Prevalent CKD	0.55 (0.13)	0.62	0.34	0.69	0.55	0.53
History of stroke	0.45 (0.11)	0.37	0.40	0.44	0.65	0.41
History of MI	0.00 (0.00)	0.00	0.00	0.00	0.00	0.00

FIGURES

FIGURE 1: Mean (standard error) LASSO, $LASSO_{sp}$, and random forest predictive performance during five-fold cross-validation

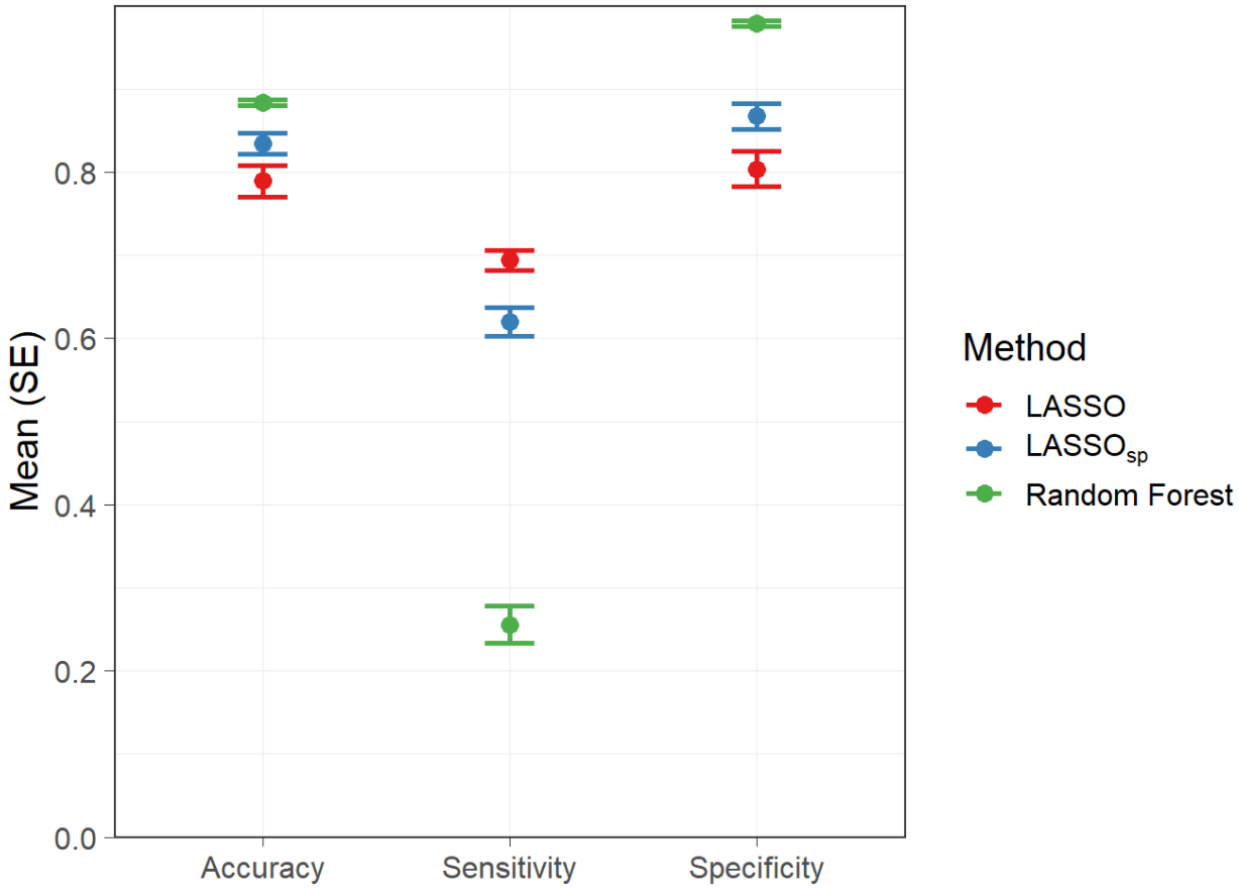
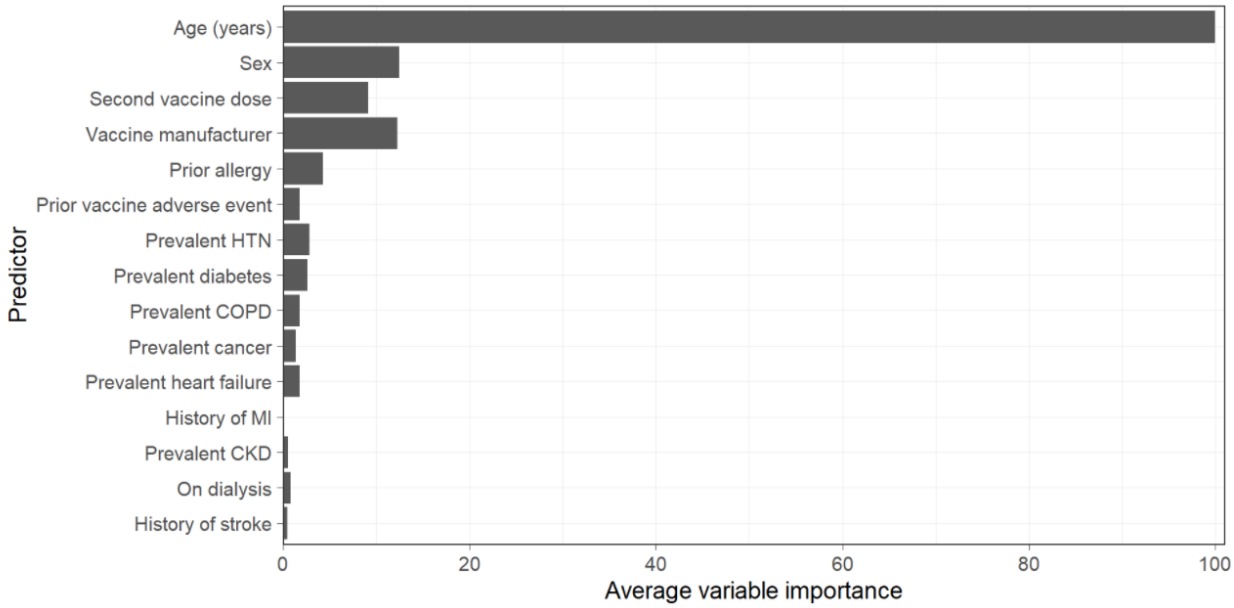
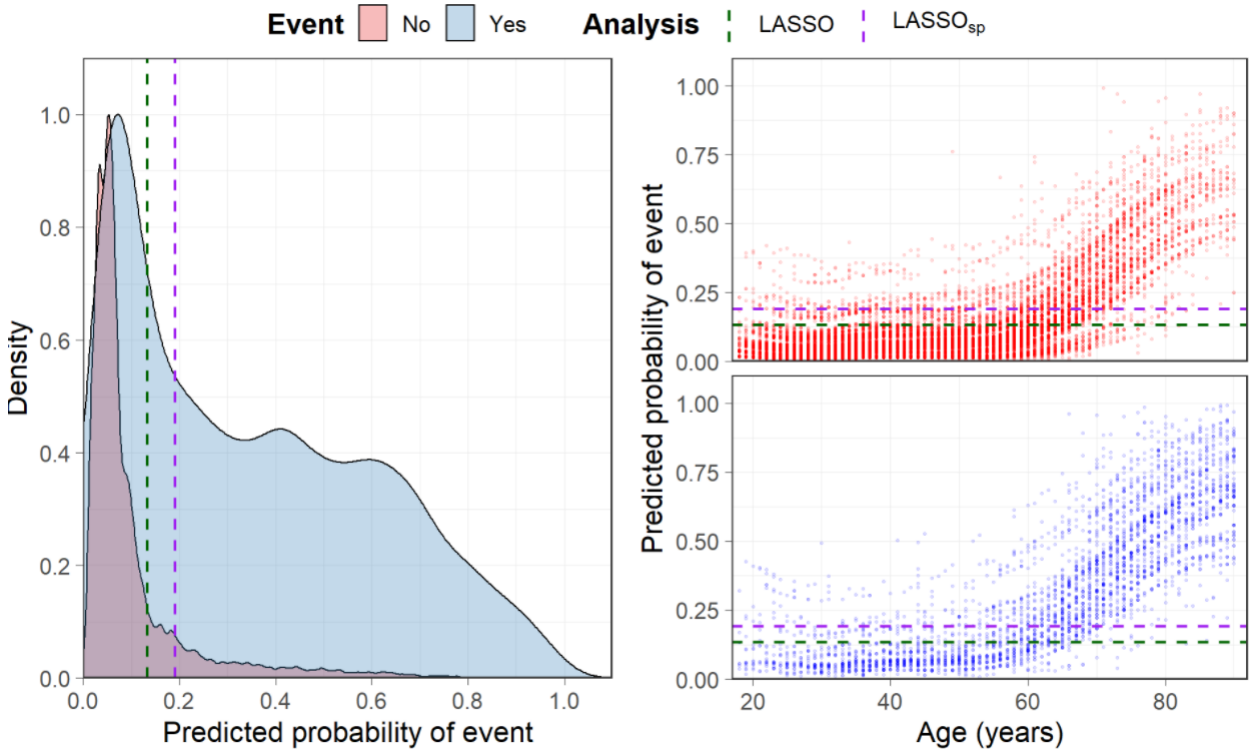


FIGURE 2: Random forest variable importance average over folds



SUPPLEMENTARY

FIGURE S1: LASSO predicted probabilities by observed event status



NOTE: Dashed line represents average threshold selected for predicting an event from ROC analyses.

FIGURE S2: Random forest number of predictors selected for each node versus out-of-bag error