# COVID-19 Baseline Risk Score Analysis Report

## mock Study

USG COVID-19 Response Biostatistics Team

April 28, 2021

# Contents

# List of Tables

# List of Figures

# Chapter 1

# Baseline Risk Score (Proxy for SARS-CoV-2 Exposure)

Table 1.1: Variables considered for risk score analysis.

| Variable.Name | Definition | Total.missing.values | Comments |
|---|---|---|---|
| MinorityInd | Baseline covariate underrepresented minority status (1=minority, 0=non-minority) | 0/30000 (0.0%) | NA |
| EthnicityHispanic | Indicator ethnicity = Hispanic (0 = Non-Hispanic) | 0/30000 (0.0%) | NA |
| EthnicityNotreported | Indicator ethnicity = Not reported (0 = Non-Hispanic) | 0/30000 (0.0%) | NA |
| EthnicityUnknown | Indicator ethnicity = Unknown (0 = Non-Hispanic) | 0/30000 (0.0%) | NA |
| Black | Indicator race = Black (0 = White) | 0/30000 (0.0%) | NA |
| Asian | Indicator race = Asian (0 = White) | 0/30000 (0.0%) | NA |
| NatAmer | Indicator race = American Indian or Alaska Native (0 = White) | 0/30000 (0.0%) | NA |
| PacIsl | Indicator race = Native Hawaiian or Other Pacific Islander (0 = White) | 0/30000 (0.0%) | NA |
| Multiracial | Indicator race = Multiracial (0 = White) | 0/30000 (0.0%) | NA |
| Other | Indicator race = Other (0 = White) | 0/30000 (0.0%) | NA |
| Notreported | Indicator race = Not reported (0 = White) | 0/30000 (0.0%) | NA |
| Unknown | Indicator race = unknown (0 = White) | 0/30000 (0.0%) | NA |
| HighRiskInd | Baseline covariate high risk pre-existing condition (1=yes, 0=no) | 0/30000 (0.0%) | NA |
| Sex | Sex assigned at birth (1=female, 0=male) | 0/30000 (0.0%) | NA |
| Age | Age at enrollment in years, between 18 and 85 | 0/30000 (0.0%) | NA |
| BMI | BMI at enrollment (kg/m^2) | 0/30000 (0.0%) | NA |

Table 1.2: All learner-screen combinations (28 in total) used as input to the Superlearner.

| Learner | Screen* |
| --- | --- |
| SL.mean | all |
| SL.glm | all<br>glmnet<br>univar_logistic_pval<br>highcor_random |
| SL.glm.interaction | glmnet<br>univar_logistic_pval<br>highcor_random |
| SL.glmnet | all |
| SL.gam | glmnet<br>univar_logistic_pval<br>highcor_random |
| SL.xgboost | all |
| SL.ranger.imp | all |

*Note:*
*Screen details:
all: includes all variables
glmnet: includes variables with non-zero coefficients in the standard implementation of SL.glmnet that optimizes the lasso tuning parameter via cross-validation
univar_logistic_pval: Wald test 2-sided p-value in a logistic regression model $< 0.10$
highcor_random: if pairs of quantitative variables with Spearman rank correlation $> 0.90$, select one of the variables at random

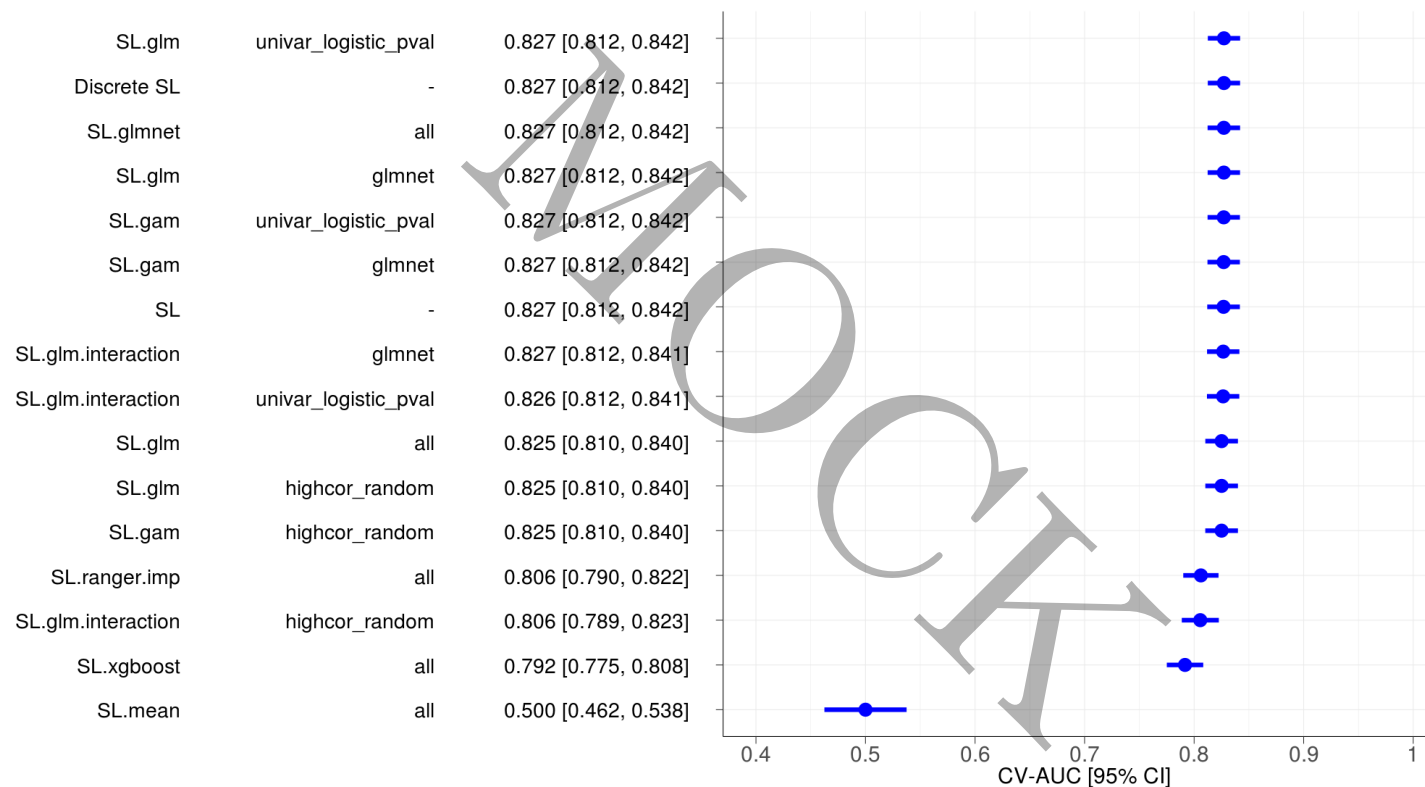| | | | |
|---|---|---|---|
| SL.glm | univar_logistic_pval | 0.827 [0.812, 0.842] | |
| Discrete SL | - | 0.827 [0.812, 0.842] | |
| SL.glmnet | all | 0.827 [0.812, 0.842] | |
| SL.glm | glmnet | 0.827 [0.812, 0.842] | |
| SL.gam | univar_logistic_pval | 0.827 [0.812, 0.842] | |
| SL.gam | glmnet | 0.827 [0.812, 0.842] | |
| SL | - | 0.827 [0.812, 0.842] | |
| SL.glm.interaction | glmnet | 0.827 [0.812, 0.841] | |
| SL.glm.interaction | univar_logistic_pval | 0.826 [0.812, 0.841] | |
| SL.glm | all | 0.825 [0.810, 0.840] | |
| SL.glm | highcor_random | 0.825 [0.810, 0.840] | |
| SL.gam | highcor_random | 0.825 [0.810, 0.840] | |
| SL.ranger.imp | all | 0.806 [0.790, 0.822] | |
| SL.glm.interaction | highcor_random | 0.806 [0.789, 0.823] | |
| SL.xgboost | all | 0.792 [0.775, 0.808] | |
| SL.mean | all | 0.500 [0.462, 0.538] | |

CV-AUC [95% CI]

Figure 1.1: Cross-validated AUC (95% CI) of algorithms for predicting COVID-19 disease status starting 7 days after Day 57.
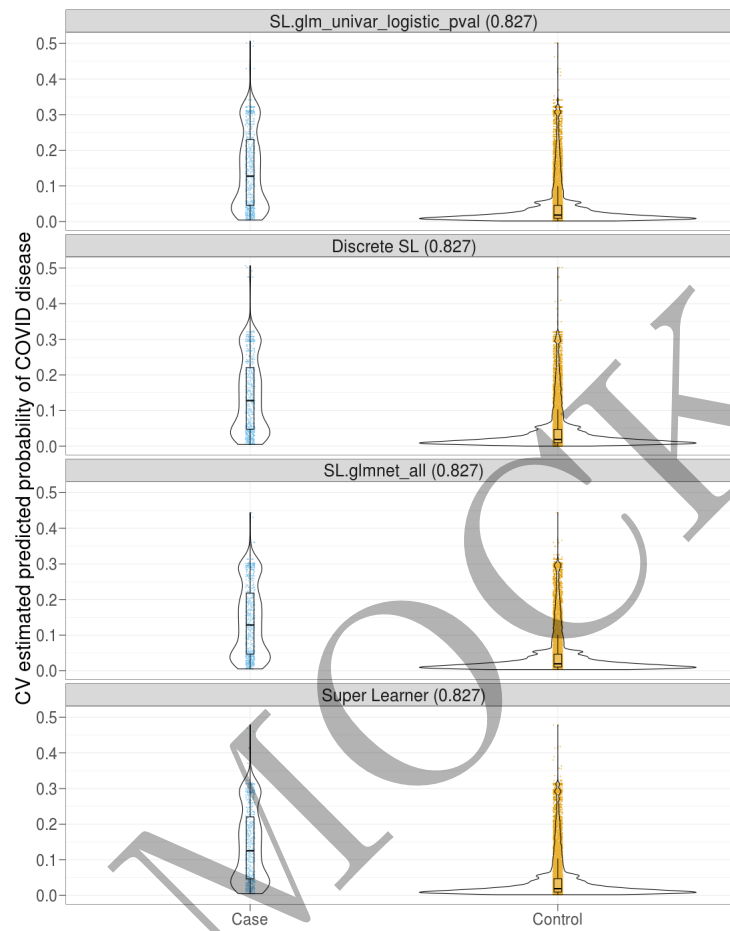
Figure 1.2: CV-estimated predicted probabilities of COVID-19 disease 7 days after Day 57 by case/control status for top 2 learners, SuperLearner and Discrete SL.
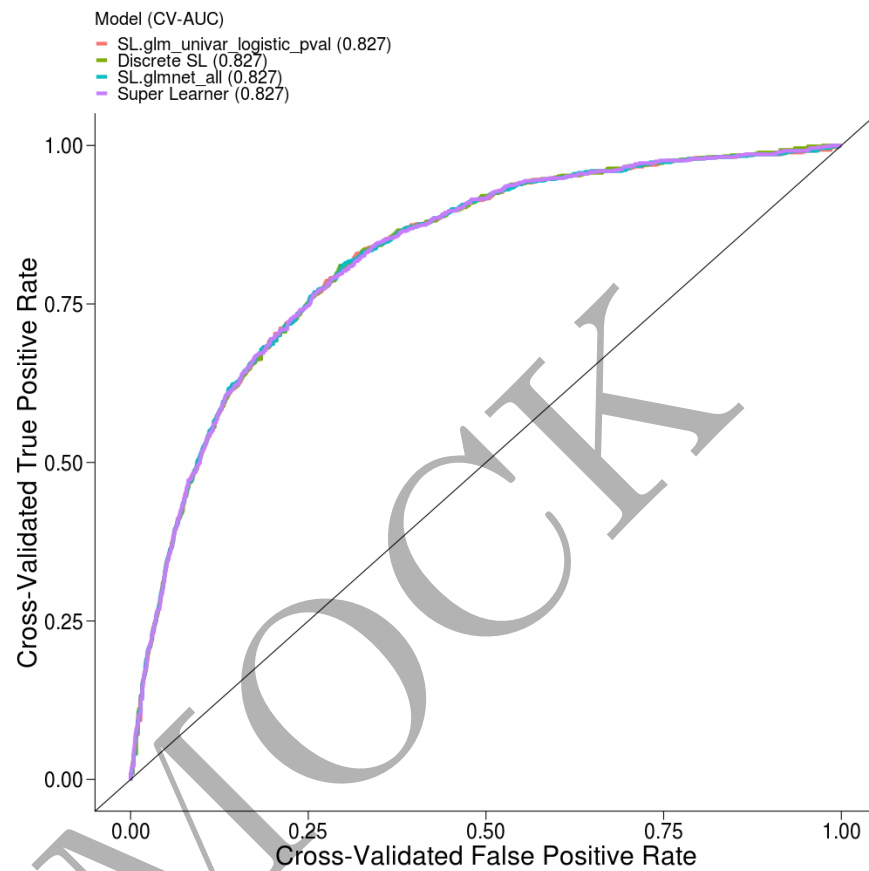
Figure 1.3: ROC curves based off CV-estimated predicted probabilities for the top 2 learners, Superlearner and Discrete SL.

Table 1.3: Weights assigned by Superlearner.

| Learner | Screen | Weight |
|---|---|---|
| SL.glmnet | screen_all | 0.606 |
| SL.glm.interaction | screen_glmnet | 0.294 |
| SL.glm | screen_univariate_logistic_pval | 0.086 |
| SL.ranger.imp | screen_all | 0.014 |
| SL.mean | screen_all | 0.000 |
| SL.glm | screen_all | 0.000 |
| SL.xgboost | screen_all | 0.000 |
| SL.glm | screen_glmnet | 0.000 |
| SL.glm | screen_highcor_random | 0.000 |
| SL.glm.interaction | screen_univariate_logistic_pval | 0.000 |
| SL.glm.interaction | screen_highcor_random | 0.000 |
| SL.gam | screen_glmnet | 0.000 |
| SL.gam | screen_univariate_logistic_pval | 0.000 |
| SL.gam | screen_highcor_random | 0.000 |

Table 1.4: Predictors in learners assigned weight > 0.0 by Superlearner.

| Learner | Screen | Weight | Predictors | Coefficient | Odds.Ratio | Importance |
|---------|--------|--------|-----------|-------------|------------|------------|
| SL.glmnet | screen_all | 0.606 | (Intercept) | -3.279 | 0.038 | NA |
| SL.glmnet | screen_all | 0.606 | MinorityInd | 0.000 | 1.000 | NA |
| SL.glmnet | screen_all | 0.606 | EthnicityHispanic | 0.000 | 1.000 | NA |
| SL.glmnet | screen_all | 0.606 | EthnicityNotreported | 0.000 | 1.000 | NA |
| SL.glmnet | screen_all | 0.606 | EthnicityUnknown | 0.000 | 1.000 | NA |
| SL.glmnet | screen_all | 0.606 | Black | 0.000 | 1.000 | NA |
| SL.glmnet | screen_all | 0.606 | Asian | 0.000 | 1.000 | NA |
| SL.glmnet | screen_all | 0.606 | NatAmer | 0.000 | 1.000 | NA |
| SL.glmnet | screen_all | 0.606 | PacIsl | 0.000 | 1.000 | NA |
| SL.glmnet | screen_all | 0.606 | Multiracial | 0.000 | 1.000 | NA |
| SL.glmnet | screen_all | 0.606 | Other | 0.000 | 1.000 | NA |
| SL.glmnet | screen_all | 0.606 | Notreported | 0.000 | 1.000 | NA |
| SL.glmnet | screen_all | 0.606 | Unknown | 0.000 | 1.000 | NA |
| SL.glmnet | screen_all | 0.606 | HighRiskInd | 0.647 | 1.910 | NA |
| SL.glmnet | screen_all | 0.606 | Sex | 0.000 | 1.000 | NA |
| SL.glmnet | screen_all | 0.606 | Age | 0.394 | 1.482 | NA |
| SL.glmnet | screen_all | 0.606 | BMI | 0.000 | 1.000 | NA |
| SL.glm.interaction | screen_glmnet | 0.294 | (Intercept) | -3.921 | 0.020 | NA |
| SL.glm.interaction | screen_glmnet | 0.294 | Other | -1.430 | 0.239 | NA |
| SL.glm.interaction | screen_glmnet | 0.294 | Unknown | 0.074 | 1.077 | NA |
| SL.glm.interaction | screen_glmnet | 0.294 | HighRiskInd | 1.053 | 2.866 | NA |
| SL.glm.interaction | screen_glmnet | 0.294 | Age | 0.819 | 2.268 | NA |
| SL.glm.interaction | screen_glmnet | 0.294 | Other:Unknown | NA | NA | NA |
| SL.glm.interaction | screen_glmnet | 0.294 | Other:HighRiskInd | 0.791 | 2.205 | NA |
| SL.glm.interaction | screen_glmnet | 0.294 | Other:Age | 0.025 | 1.025 | NA |
| SL.glm.interaction | screen_glmnet | 0.294 | Unknown:HighRiskInd | 0.006 | 1.006 | NA |
| SL.glm.interaction | screen_glmnet | 0.294 | Unknown:Age | 0.006 | 1.006 | NA |
| SL.glm.interaction | screen_glmnet | 0.294 | HighRiskInd:Age | -0.078 | 0.925 | NA |
| SL.glm | screen_univariate_logistic_pval | 0.086 | (Intercept) | -3.690 | 0.025 | NA |
| SL.glm | screen_univariate_logistic_pval | 0.086 | Other | -0.104 | 0.901 | NA |
| SL.glm | screen_univariate_logistic_pval | 0.086 | Unknown | 0.083 | 1.087 | NA |
| SL.glm | screen_univariate_logistic_pval | 0.086 | HighRiskInd | 0.897 | 2.453 | NA |
| SL.glm | screen_univariate_logistic_pval | 0.086 | Age | 0.738 | 2.092 | NA |
| SL.ranger.imp | screen_all | 0.014 | MinorityInd | NA | NA | 7.977 |
| SL.ranger.imp | screen_all | 0.014 | EthnicityHispanic | NA | NA | 7.681 |
| SL.ranger.imp | screen_all | 0.014 | EthnicityNotreported | NA | NA | 7.226 |
| SL.ranger.imp | screen_all | 0.014 | EthnicityUnknown | NA | NA | 6.967 |
| SL.ranger.imp | screen_all | 0.014 | Black | NA | NA | 8.158 |
| SL.ranger.imp | screen_all | 0.014 | Asian | NA | NA | 6.157 |
| SL.ranger.imp | screen_all | 0.014 | NatAmer | NA | NA | 3.761 |
| SL.ranger.imp | screen_all | 0.014 | PacIsl | NA | NA | 3.201 |
| SL.ranger.imp | screen_all | 0.014 | Multiracial | NA | NA | 5.162 |
| SL.ranger.imp | screen_all | 0.014 | Other | NA | NA | 2.423 |
| SL.ranger.imp | screen_all | 0.014 | Notreported | NA | NA | 6.122 |
| SL.ranger.imp | screen_all | 0.014 | Unknown | NA | NA | 4.650 |

Table 1.4: Predictors in learners assigned weight > 0.0 by Superlearner. *(continued)*

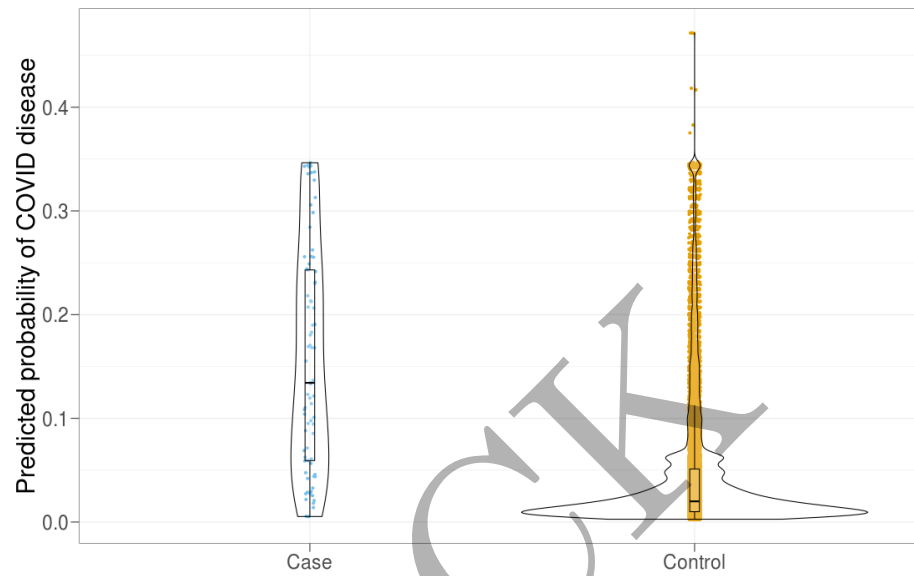| Learner | Screen | Weight | Predictors | Coefficient | Odds.Ratio | Importance |
|---|---|---|---|---|---|---|
| SL.ranger.imp | screen_all | 0.014 | HighRiskInd | NA | NA | 87.399 |
| SL.ranger.imp | screen_all | 0.014 | Sex | NA | NA | 13.377 |
| SL.ranger.imp | screen_all | 0.014 | Age | NA | NA | 129.163 |
| SL.ranger.imp | screen_all | 0.014 | BMI | NA | NA | 177.933 |
| SL.ranger.imp | screen_all | 0.014 | MinorityInd | NA | NA | 7.977 |
| SL.ranger.imp | screen_all | 0.014 | EthnicityHispanic | NA | NA | 7.681 |
| SL.ranger.imp | screen_all | 0.014 | EthnicityNotreported | NA | NA | 7.226 |
| SL.ranger.imp | screen_all | 0.014 | EthnicityUnknown | NA | NA | 6.967 |
| SL.ranger.imp | screen_all | 0.014 | Black | NA | NA | 8.158 |
| SL.ranger.imp | screen_all | 0.014 | Asian | NA | NA | 6.157 |
| SL.ranger.imp | screen_all | 0.014 | NatAmer | NA | NA | 3.761 |
| SL.ranger.imp | screen_all | 0.014 | PacIsl | NA | NA | 3.201 |
| SL.ranger.imp | screen_all | 0.014 | Multiracial | NA | NA | 5.162 |
| SL.ranger.imp | screen_all | 0.014 | Other | NA | NA | 2.423 |
| SL.ranger.imp | screen_all | 0.014 | Notreported | NA | NA | 6.122 |
| SL.ranger.imp | screen_all | 0.014 | Unknown | NA | NA | 4.650 |
| SL.ranger.imp | screen_all | 0.014 | HighRiskInd | NA | NA | 87.399 |
| SL.ranger.imp | screen_all | 0.014 | Sex | NA | NA | 13.377 |
| SL.ranger.imp | screen_all | 0.014 | Age | NA | NA | 129.163 |
| SL.ranger.imp | screen_all | 0.014 | BMI | NA | NA | 177.933 |

Figure 1.4: Superlearner predicted probabilities of COVID-19 disease in vaccinees 7 days after Day 57 by case/control status.
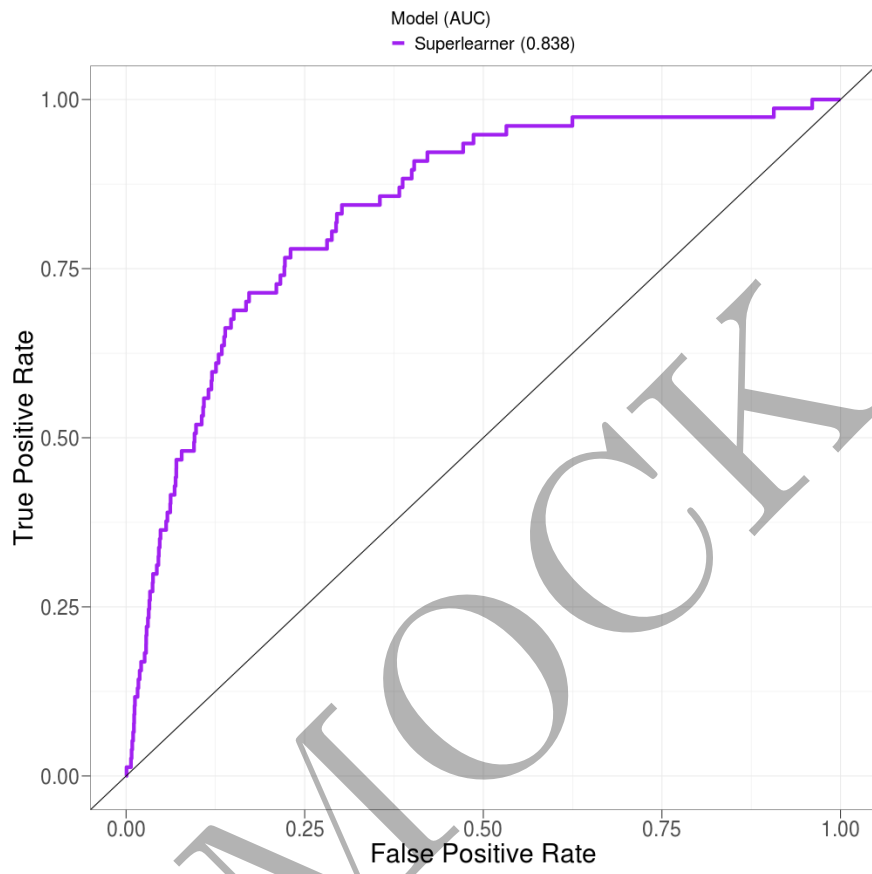
Figure 1.5: ROC curve based off Superlearner predicted probabilities in vaccinees.

# Chapter 2

# Appendix

The commit hash this report was built on from `CoVPN/correlates_reporting` repository:

7b7c6f05e9c1f8520ac1ca1a4bd938ac92059fcf

The sha256 hash sum of the raw input file, "COVID_VEtrial_practicedata_primarystage1.csv":

5ad1ea260155f243f8c988ded6cc3c2a35610379144393cd041ceadbcfb76a06

The sha256 hash sum of the processed file, "practice_data.csv":

6ec258268f7dcd7b3f4aaf954581d355093918f181aa410b24069277d5a0aed2