

COVID-19 Baseline Risk Score Analysis Report

mock Study

USG COVID-19 Response Biostatistics Team

May 12, 2021

Contents

| | | |
|----------|--|-----------|
| 1 | Baseline Risk Score (Proxy for SARS-CoV-2 Exposure) | 9 |
| 2 | Appendix | 19 |

List of Tables

| | | |
|-----|---|----|
| 1.1 | Variables considered for risk score analysis. | 10 |
| 1.2 | All learner-screen combinations (28 in total) used as input to the Superlearner. | 11 |
| 1.3 | Weights assigned by Superlearner. | 15 |
| 1.4 | Predictors in learners assigned weight > 0.0 by Superlearner. . . | 16 |

List of Figures

| | | |
|-----|--|----|
| 1.1 | Cross-validated AUC (95% CI) of algorithms for predicting COVID-19 disease status starting 7 days after Day 57. | 12 |
| 1.2 | CV-estimated predicted probabilities of COVID-19 disease 7 days after Day 57 by case/control status for top 2 learners, Super-Learner and Discrete SL. | 13 |
| 1.3 | ROC curves based off CV-estimated predicted probabilities for the top 2 learners, Superlearner and Discrete SL. | 14 |
| 1.4 | Superlearner predicted probabilities of COVID-19 disease in vaccinees 7 days after Day 57 by case/control status. | 17 |
| 1.5 | ROC curve based off Superlearner predicted probabilities in vaccinees. | 18 |

MOCK

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 glmnet univar_logistic_pval highcor_random |
| SL.glm.interaction | glmnet univar_logistic_pval highcor_random |
| SL.glmnet | all |
| SL.gam | glmnet univar_logistic_pval 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



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.

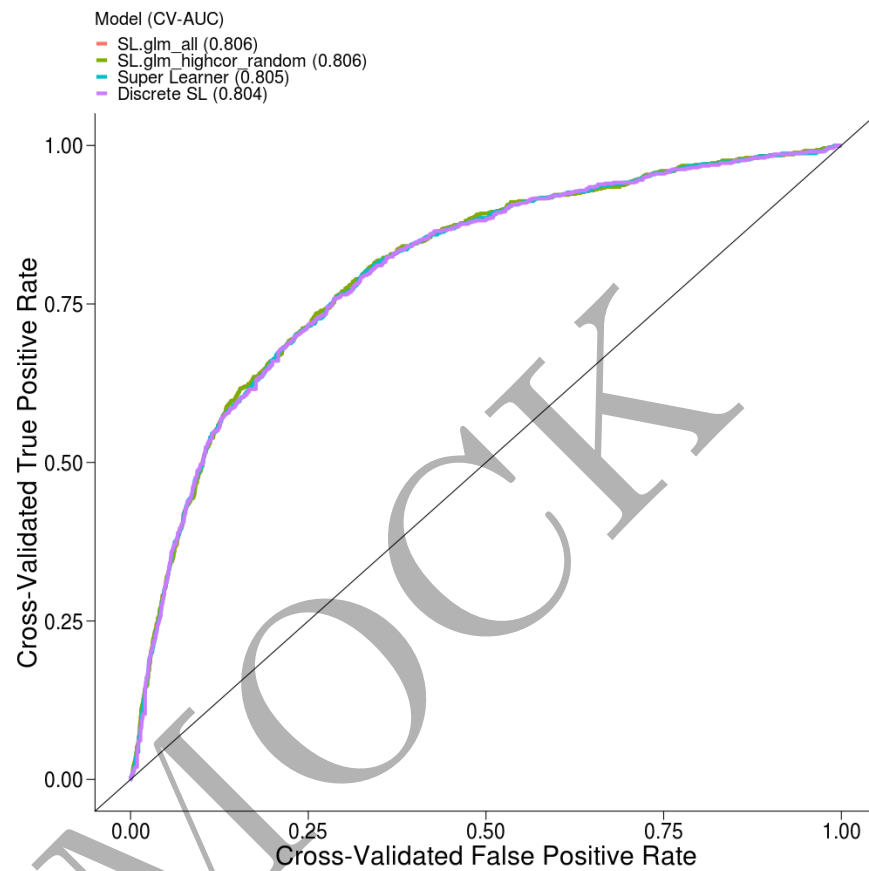


Figure 1.3: ROC curves based off CV-estimated predicted probabilities for the top 2 learners, Superlearner and Discrete SL.

Table 1.3: Weights assigned by Superlearner.

| Learner | Screen | Weight |
|--------------------|---------------------------------|--------|
| SL.glm | screen_all | 0.714 |
| SL.glm.interaction | screen_univariate_logistic_pval | 0.271 |
| SL.mean | screen_all | 0.016 |
| SL.glmnet | screen_all | 0.000 |
| SL.xgboost | screen_all | 0.000 |
| SL.ranger.imp | screen_all | 0.000 |
| SL.glm | screen_glmnet | 0.000 |
| SL.glm | screen_univariate_logistic_pval | 0.000 |
| SL.glm | screen_highcor_random | 0.000 |
| SL.glm.interaction | screen_glmnet | 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 |
|--------------------|---------------------------------|--------|----------------------|-------------|------------|
| SL.glm | screen_all | 0.714 | (Intercept) | -3.592 | 0.028 |
| SL.glm | screen_all | 0.714 | MinorityInd | 0.109 | 1.115 |
| SL.glm | screen_all | 0.714 | EthnicityHispanic | 0.016 | 1.017 |
| SL.glm | screen_all | 0.714 | EthnicityNotreported | 0.064 | 1.066 |
| SL.glm | screen_all | 0.714 | EthnicityUnknown | 0.055 | 1.057 |
| SL.glm | screen_all | 0.714 | Black | -0.127 | 0.881 |
| SL.glm | screen_all | 0.714 | Asian | -0.051 | 0.950 |
| SL.glm | screen_all | 0.714 | NatAmer | -0.022 | 0.978 |
| SL.glm | screen_all | 0.714 | PacIsl | 0.014 | 1.014 |
| SL.glm | screen_all | 0.714 | Multiracial | -0.106 | 0.900 |
| SL.glm | screen_all | 0.714 | Other | 0.049 | 1.051 |
| SL.glm | screen_all | 0.714 | Notreported | 0.064 | 1.066 |
| SL.glm | screen_all | 0.714 | Unknown | 0.056 | 1.058 |
| SL.glm | screen_all | 0.714 | HighRiskInd | 0.793 | 2.210 |
| SL.glm | screen_all | 0.714 | Sex | -0.033 | 0.967 |
| SL.glm | screen_all | 0.714 | Age | 0.743 | 2.102 |
| SL.glm | screen_all | 0.714 | BMI | -0.001 | 0.999 |
| SL.glm.interaction | screen_univariate_logistic_pval | 0.271 | (Intercept) | -3.604 | 0.027 |
| SL.glm.interaction | screen_univariate_logistic_pval | 0.271 | PacIsl | 0.060 | 1.062 |
| SL.glm.interaction | screen_univariate_logistic_pval | 0.271 | Other | 0.112 | 1.119 |
| SL.glm.interaction | screen_univariate_logistic_pval | 0.271 | HighRiskInd | 0.806 | 2.239 |
| SL.glm.interaction | screen_univariate_logistic_pval | 0.271 | Age | 0.767 | 2.154 |
| SL.glm.interaction | screen_univariate_logistic_pval | 0.271 | PacIsl:Other | NA | NA |
| SL.glm.interaction | screen_univariate_logistic_pval | 0.271 | PacIsl:HighRiskInd | 0.020 | 1.020 |
| SL.glm.interaction | screen_univariate_logistic_pval | 0.271 | PacIsl:Age | -0.083 | 0.921 |
| SL.glm.interaction | screen_univariate_logistic_pval | 0.271 | Other:HighRiskInd | -0.015 | 0.985 |
| SL.glm.interaction | screen_univariate_logistic_pval | 0.271 | Other:Age | -0.050 | 0.952 |
| SL.glm.interaction | screen_univariate_logistic_pval | 0.271 | HighRiskInd:Age | -0.021 | 0.979 |
| SL.glm.interaction | screen_univariate_logistic_pval | 0.271 | (Intercept) | -3.604 | 0.027 |
| SL.glm.interaction | screen_univariate_logistic_pval | 0.271 | PacIsl | 0.060 | 1.062 |
| SL.glm.interaction | screen_univariate_logistic_pval | 0.271 | Other | 0.112 | 1.119 |
| SL.glm.interaction | screen_univariate_logistic_pval | 0.271 | HighRiskInd | 0.806 | 2.239 |
| SL.glm.interaction | screen_univariate_logistic_pval | 0.271 | Age | 0.767 | 2.154 |
| SL.glm.interaction | screen_univariate_logistic_pval | 0.271 | PacIsl:Other | NA | NA |
| SL.glm.interaction | screen_univariate_logistic_pval | 0.271 | PacIsl:HighRiskInd | 0.020 | 1.020 |
| SL.glm.interaction | screen_univariate_logistic_pval | 0.271 | PacIsl:Age | -0.083 | 0.921 |
| SL.glm.interaction | screen_univariate_logistic_pval | 0.271 | Other:HighRiskInd | -0.015 | 0.985 |
| SL.glm.interaction | screen_univariate_logistic_pval | 0.271 | Other:Age | -0.050 | 0.952 |
| SL.glm.interaction | screen_univariate_logistic_pval | 0.271 | HighRiskInd:Age | -0.021 | 0.979 |



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

- This report was built from the [CoVPN/correlates_reporting](https://github.com/CoVPN/correlates_reporting) repository with commit hash b17cf6b09b9cf8d496a76d3b35656e84af2c7f99. A diff of the changes introduced by that commit may be viewed at https://github.com/CoVPN/correlates_reporting/commit/b17cf6b09b9cf8d496a76d3b35656e84af2c7f99
- The sha256 hash sum of the raw input file, “COVID_VEtrial_practicedata_primarystage1.csv”:
2353971c2e14399ede55ef6ba0d4e624626433dc15ec507c2482bb886210019a
- The sha256 hash sum of the processed file, “practice_data.csv”:
139bb66a1fa4cdcb6987bd18f5f3bb83ee7ecc49934e31a0af0431a15f1b2593