

COVID-19 Baseline Risk Score Analysis Report

mock Study

USG COVID-19 Response Biostatistics Team

May 19, 2021

Contents

| | | |
|----------|--|-----------|
| 1 | Baseline Risk Score (Proxy for SARS-CoV-2 Exposure) | 9 |
| 2 | Appendix | 21 |

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. | 18 |
| 1.5 | ROC curve based off Superlearner predicted probabilities in vaccinees. | 19 |

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.

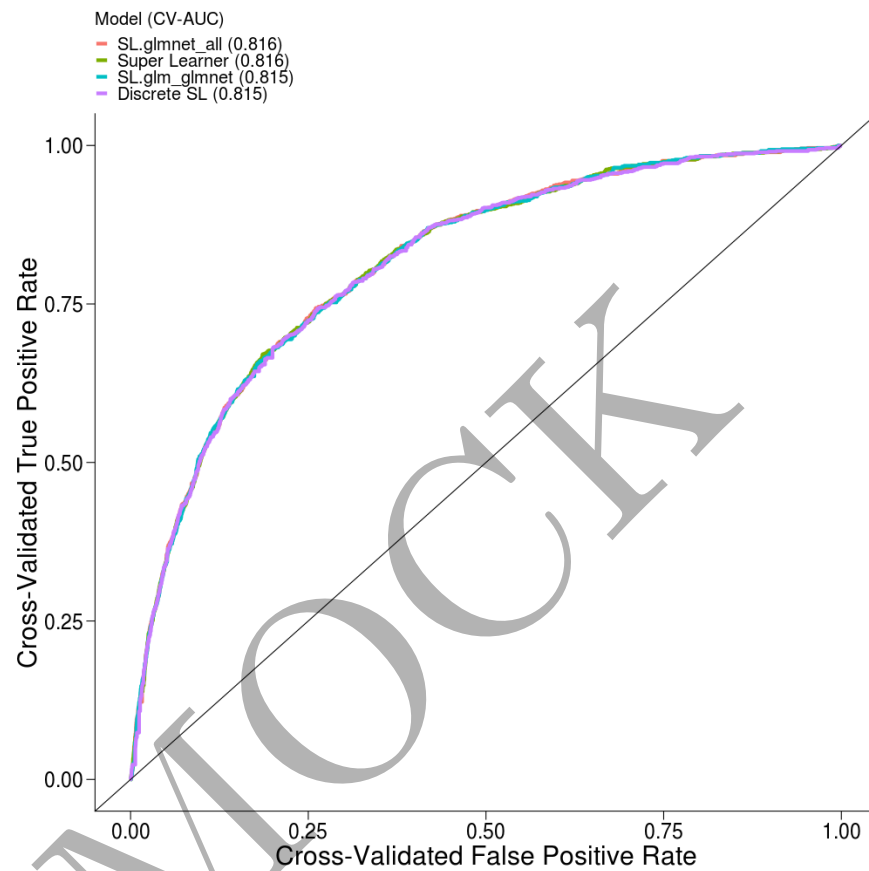


Figure 1.3: ROC curves based off CV-estimated predicted probabilities for the top 2 learners, Superlearner and Discrete SL.

Table 1.3: Weights assigned by Superlearner.

| Learner | Screen | Weight |
|--------------------|---------------------------------|--------|
| SL.glmnet | screen_all | 0.654 |
| SL.ranger.imp | screen_all | 0.125 |
| SL.glm | screen_glmnet | 0.098 |
| SL.glm | screen_univariate_logistic_pval | 0.066 |
| SL.glm | screen_all | 0.057 |
| SL.mean | screen_all | 0.000 |
| SL.xgboost | screen_all | 0.000 |
| SL.glm | screen_highcor_random | 0.000 |
| SL.glm.interaction | screen_glmnet | 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 | 0.654 | (Intercept) | -2.868 | 0.057 | NA |
| SL.glmnet | screen_all | 0.654 | MinorityInd | 0.000 | 1.000 | NA |
| SL.glmnet | screen_all | 0.654 | EthnicityHispanic | 0.000 | 1.000 | NA |
| SL.glmnet | screen_all | 0.654 | EthnicityNotreported | 0.000 | 1.000 | NA |
| SL.glmnet | screen_all | 0.654 | EthnicityUnknown | 0.000 | 1.000 | NA |
| SL.glmnet | screen_all | 0.654 | Black | 0.000 | 1.000 | NA |
| SL.glmnet | screen_all | 0.654 | Asian | 0.000 | 1.000 | NA |
| SL.glmnet | screen_all | 0.654 | NatAmer | 0.000 | 1.000 | NA |
| SL.glmnet | screen_all | 0.654 | PacIsl | 0.000 | 1.000 | NA |
| SL.glmnet | screen_all | 0.654 | Multiracial | 0.000 | 1.000 | NA |
| SL.glmnet | screen_all | 0.654 | Other | 0.000 | 1.000 | NA |
| SL.glmnet | screen_all | 0.654 | Notreported | 0.000 | 1.000 | NA |
| SL.glmnet | screen_all | 0.654 | Unknown | 0.000 | 1.000 | NA |
| SL.glmnet | screen_all | 0.654 | HighRiskInd | 0.615 | 1.850 | NA |
| SL.glmnet | screen_all | 0.654 | Sex | 0.000 | 1.000 | NA |
| SL.glmnet | screen_all | 0.654 | Age | 0.431 | 1.539 | NA |
| SL.glmnet | screen_all | 0.654 | BMI | 0.000 | 1.000 | NA |
| SL.ranger.imp | screen_all | 0.125 | MinorityInd | NA | NA | 11.065 |
| SL.ranger.imp | screen_all | 0.125 | EthnicityHispanic | NA | NA | 10.195 |
| SL.ranger.imp | screen_all | 0.125 | EthnicityNotreported | NA | NA | 11.820 |
| SL.ranger.imp | screen_all | 0.125 | EthnicityUnknown | NA | NA | 9.214 |
| SL.ranger.imp | screen_all | 0.125 | Black | NA | NA | 10.111 |
| SL.ranger.imp | screen_all | 0.125 | Asian | NA | NA | 8.396 |
| SL.ranger.imp | screen_all | 0.125 | NatAmer | NA | NA | 5.002 |
| SL.ranger.imp | screen_all | 0.125 | PacIsl | NA | NA | 4.260 |
| SL.ranger.imp | screen_all | 0.125 | Multiracial | NA | NA | 7.162 |
| SL.ranger.imp | screen_all | 0.125 | Other | NA | NA | 4.091 |
| SL.ranger.imp | screen_all | 0.125 | Notreported | NA | NA | 9.353 |
| SL.ranger.imp | screen_all | 0.125 | Unknown | NA | NA | 4.914 |
| SL.ranger.imp | screen_all | 0.125 | HighRiskInd | NA | NA | 144.498 |
| SL.ranger.imp | screen_all | 0.125 | Sex | NA | NA | 16.469 |
| SL.ranger.imp | screen_all | 0.125 | Age | NA | NA | 198.616 |
| SL.ranger.imp | screen_all | 0.125 | BMI | NA | NA | 221.076 |
| SL.glm | screen_glmnet | 0.098 | (Intercept) | -3.234 | 0.039 | NA |
| SL.glm | screen_glmnet | 0.098 | MinorityInd | -0.044 | 0.957 | NA |
| SL.glm | screen_glmnet | 0.098 | Other | -0.089 | 0.914 | NA |
| SL.glm | screen_glmnet | 0.098 | Notreported | 0.036 | 1.036 | NA |
| SL.glm | screen_glmnet | 0.098 | HighRiskInd | 0.862 | 2.367 | NA |
| SL.glm | screen_glmnet | 0.098 | Sex | 0.080 | 1.084 | NA |
| SL.glm | screen_glmnet | 0.098 | Age | 0.763 | 2.145 | NA |
| SL.glm | screen_glmnet | 0.098 | BMI | -0.047 | 0.954 | NA |
| SL.glm | screen_univariate_logistic_pval | 0.066 | (Intercept) | -3.232 | 0.039 | NA |
| SL.glm | screen_univariate_logistic_pval | 0.066 | Other | -0.099 | 0.906 | NA |
| SL.glm | screen_univariate_logistic_pval | 0.066 | HighRiskInd | 0.861 | 2.366 | NA |
| SL.glm | screen_univariate_logistic_pval | 0.066 | Sex | 0.080 | 1.083 | NA |

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

| Learner | Screen | Weight | Predictors | Coefficient | Odds.Ratio | Importance |
|---------|---------------------------------|--------|----------------------|-------------|------------|------------|
| SL.glm | screen_univariate_logistic_pval | 0.066 | Age | 0.763 | 2.144 | NA |
| SL.glm | screen_all | 0.057 | (Intercept) | -3.236 | 0.039 | NA |
| SL.glm | screen_all | 0.057 | MinorityInd | -0.102 | 0.903 | NA |
| SL.glm | screen_all | 0.057 | EthnicityHispanic | 0.060 | 1.062 | NA |
| SL.glm | screen_all | 0.057 | EthnicityNotreported | 0.018 | 1.018 | NA |
| SL.glm | screen_all | 0.057 | EthnicityUnknown | -0.022 | 0.978 | NA |
| SL.glm | screen_all | 0.057 | Black | 0.023 | 1.023 | NA |
| SL.glm | screen_all | 0.057 | Asian | 0.025 | 1.025 | NA |
| SL.glm | screen_all | 0.057 | NatAmer | -0.013 | 0.987 | NA |
| SL.glm | screen_all | 0.057 | PacIsl | 0.018 | 1.018 | NA |
| SL.glm | screen_all | 0.057 | Multiracial | 0.022 | 1.022 | NA |
| SL.glm | screen_all | 0.057 | Other | -0.091 | 0.913 | NA |
| SL.glm | screen_all | 0.057 | Notreported | 0.036 | 1.037 | NA |
| SL.glm | screen_all | 0.057 | Unknown | 0.027 | 1.027 | NA |
| SL.glm | screen_all | 0.057 | HighRiskInd | 0.862 | 2.368 | NA |
| SL.glm | screen_all | 0.057 | Sex | 0.080 | 1.084 | NA |
| SL.glm | screen_all | 0.057 | Age | 0.765 | 2.149 | NA |
| SL.glm | screen_all | 0.057 | BMI | -0.047 | 0.954 | NA |
| SL.glm | screen_all | 0.057 | (Intercept) | -3.236 | 0.039 | NA |
| SL.glm | screen_all | 0.057 | MinorityInd | -0.102 | 0.903 | NA |
| SL.glm | screen_all | 0.057 | EthnicityHispanic | 0.060 | 1.062 | NA |
| SL.glm | screen_all | 0.057 | EthnicityNotreported | 0.018 | 1.018 | NA |
| SL.glm | screen_all | 0.057 | EthnicityUnknown | -0.022 | 0.978 | NA |
| SL.glm | screen_all | 0.057 | Black | 0.023 | 1.023 | NA |
| SL.glm | screen_all | 0.057 | Asian | 0.025 | 1.025 | NA |
| SL.glm | screen_all | 0.057 | NatAmer | -0.013 | 0.987 | NA |
| SL.glm | screen_all | 0.057 | PacIsl | 0.018 | 1.018 | NA |
| SL.glm | screen_all | 0.057 | Multiracial | 0.022 | 1.022 | NA |
| SL.glm | screen_all | 0.057 | Other | -0.091 | 0.913 | NA |
| SL.glm | screen_all | 0.057 | Notreported | 0.036 | 1.037 | NA |
| SL.glm | screen_all | 0.057 | Unknown | 0.027 | 1.027 | NA |
| SL.glm | screen_all | 0.057 | HighRiskInd | 0.862 | 2.368 | NA |
| SL.glm | screen_all | 0.057 | Sex | 0.080 | 1.084 | NA |
| SL.glm | screen_all | 0.057 | Age | 0.765 | 2.149 | NA |
| SL.glm | screen_all | 0.057 | BMI | -0.047 | 0.954 | NA |



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.

MOCK

Chapter 2

Appendix

- This report was built from the [CoVPN/correlates_reporting](https://github.com/CoVPN/correlates_reporting) repository with commit hash `ec058661281464915e45e8e87a269d17a5f3f7af`. A diff of the changes introduced by that commit may be viewed at https://github.com/CoVPN/correlates_reporting/commit/ec058661281464915e45e8e87a269d17a5f3f7af
- The sha256 hash sum of the raw input file, “COVID_VEtrial_practicedata_primarystage1.csv”:
`45ff85033ffbc717462d678b41bc4060a12c7bc60952e2cb72297bb5500b97b9`
- The sha256 hash sum of the processed file, “practice_data.csv”:
`8de0bf2f66901eb123908b42ec8dd87cf9304412ca180aff8a88acfeb4c777e4`