

COVID-19 Baseline Risk Score Analysis Report

MockENSEMBLE Study

USG COVID-19 Response Biostatistics Team

July 26, 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. | 9 |
| 1.2 | All learner-screen combinations (14 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 29. | 12 |
| 1.2 | CV-estimated predicted probabilities of COVID-19 disease 7 days after Day 29 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 29 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 |
|-------------------------|--|----------------------|----------|
| EthnicityHispanic | Indicator ethnicity = Hispanic (1 = Hispanic, 0 = complement) | 0/19501 (0.0%) | NA |
| EthnicityNotreported | Indicator ethnicity = Not reported (1 = Not reported, 0 = complement) | 0/19501 (0.0%) | NA |
| EthnicityUnknown | Indicator ethnicity = Unknown (1 = Unknown, 0 = complement) | 0/19501 (0.0%) | NA |
| Black | Indicator race = Black (1=Black, 0=complement) | 0/19501 (0.0%) | NA |
| Asian | Indicator race = Asian (1=Asian, 0=complement) | 0/19501 (0.0%) | NA |
| NatAmer | Indicator race = American Indian or Alaska Native (1=NatAmer, 0=complement) | 0/19501 (0.0%) | NA |
| Multiracial | Indicator race = Multiracial (1=Multiracial, 0=complement) | 0/19501 (0.0%) | NA |
| Notreported | Indicator race = Not reported (1=Notreported, 0=complement) | 0/19501 (0.0%) | NA |
| Unknown | Indicator race = unknown (1=Unknown, 0=complement) | 0/19501 (0.0%) | NA |
| URMforsubcohortsampling | Indicator of under-represented minority (1=Yes, 0=No) | 0/19501 (0.0%) | NA |
| HighRiskInd | Baseline covariate indicating ≥ 1 Co-existing conditions (1=yes, 0=no, NA=missing) | 0/19501 (0.0%) | NA |
| Sex | Sex assigned at birth (1=female, 0=male/undifferentiated/unknown) | 0/19501 (0.0%) | NA |
| Age | Age at enrollment in years (integer ≥ 18 , NA=missing). Note that the randomization strata included Age 18-59 vs. Age ≥ 60 . | 0/19501 (0.0%) | NA |

Table 1.1: Variables considered for risk score analysis. (*continued*)

| Variable.Name | Definition | Total.missing.values |
|------------------------|--|----------------------|
| BMI | BMI at enrollment (Ordered categorical 1,2, 3, 4, NA=missing); 1 = Underweight BMI < 18.5; 2 = Normal BMI 18.5 to < 25; 3 = Overweight BMI 25 to < 30; 4 = Obese BMI >= 30 | 0/19501 (0.0%) |
| Country | Country of the study site of enrollment (0=United States, 1=Argentina,2=Brazil, 3=Chile,4=Columbia, 5=Mexico, 6=Peru, 7=South Africa) | 0/19501 (0.0%) |
| HIVinfection | Indicator HIV infected at enrollment (1=infected, 0=not infected) | 0/19501 (0.0%) |
| CalendarDateEnrollment | Date variable (used to control for calendar time trends in COVID incidence). Coded as number of days since first person enrolled until the ppt is enrolled. | 0/19501 (0.0%) |

Note:

Variables with more than 5% missing values were dropped from analysis; missing values for other variables were dropped. Indicator variables not meeting the threshold, such that under the null of not a risk factor there were less than 5% in the subgroup with value 1 (or 0), were dropped from analysis.

Table 1.2: All learner-screen combinations (14 in total) used as input to the Superlearner.

| Learner | Screen* |
|---------|---|
| SL.mean | all |
| SL.glm | all glmnet univar_logistic_pval highcor_random |

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 29.



Figure 1.2: CV-estimated predicted probabilities of COVID-19 disease 7 days after Day 29 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 | 0.833 |
| SL.glm | screen_glmnet | 0.167 |
| SL.mean | screen_all | 0.000 |
| SL.glm | screen_univariate_logistic_pval | 0.000 |
| SL.glm | 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 | 0.833 | (Intercept) | -19.397 | 0 |
| SL.glm | screen_all | 0.833 | EthnicityHispanic | -0.037 | 0.964 |
| SL.glm | screen_all | 0.833 | EthnicityNotreported | 0.013 | 1.013 |
| SL.glm | screen_all | 0.833 | EthnicityUnknown | 0.05 | 1.051 |
| SL.glm | screen_all | 0.833 | Black | -0.16 | 0.852 |
| SL.glm | screen_all | 0.833 | Asian | -0.066 | 0.936 |
| SL.glm | screen_all | 0.833 | NatAmer | -0.056 | 0.945 |
| SL.glm | screen_all | 0.833 | Multiracial | -0.039 | 0.961 |
| SL.glm | screen_all | 0.833 | Notreported | -1.403 | 0.246 |
| SL.glm | screen_all | 0.833 | Unknown | -0.056 | 0.946 |
| SL.glm | screen_all | 0.833 | URMforsubcohortsampling | 0.063 | 1.065 |
| SL.glm | screen_all | 0.833 | HighRiskInd | 0.379 | 1.461 |
| SL.glm | screen_all | 0.833 | Sex | 12.071 | 174754.678 |
| SL.glm | screen_all | 0.833 | Age | 0.354 | 1.425 |
| SL.glm | screen_all | 0.833 | BMI | -0.046 | 0.955 |
| SL.glm | screen_all | 0.833 | Country | -0.017 | 0.983 |
| SL.glm | screen_all | 0.833 | HIVinfection | 0.011 | 1.011 |
| SL.glm | screen_all | 0.833 | CalendarDateEnrollment | 0.027 | 1.027 |
| SL.glm | screen_glmnet | 0.167 | (Intercept) | -19.395 | 0 |
| SL.glm | screen_glmnet | 0.167 | EthnicityNotreported | 0.015 | 1.015 |
| SL.glm | screen_glmnet | 0.167 | EthnicityUnknown | 0.052 | 1.053 |
| SL.glm | screen_glmnet | 0.167 | Black | -0.136 | 0.873 |
| SL.glm | screen_glmnet | 0.167 | Asian | -0.067 | 0.936 |
| SL.glm | screen_glmnet | 0.167 | NatAmer | -0.055 | 0.947 |
| SL.glm | screen_glmnet | 0.167 | Multiracial | -0.04 | 0.961 |
| SL.glm | screen_glmnet | 0.167 | Notreported | -1.403 | 0.246 |
| SL.glm | screen_glmnet | 0.167 | Unknown | -0.056 | 0.946 |
| SL.glm | screen_glmnet | 0.167 | URMforsubcohortsampling | 0.03 | 1.031 |
| SL.glm | screen_glmnet | 0.167 | HighRiskInd | 0.379 | 1.461 |
| SL.glm | screen_glmnet | 0.167 | Sex | 12.07 | 174555.444 |
| SL.glm | screen_glmnet | 0.167 | Age | 0.354 | 1.425 |
| SL.glm | screen_glmnet | 0.167 | BMI | -0.046 | 0.955 |
| SL.glm | screen_glmnet | 0.167 | Country | -0.013 | 0.987 |
| SL.glm | screen_glmnet | 0.167 | HIVinfection | 0.011 | 1.011 |
| SL.glm | screen_glmnet | 0.167 | CalendarDateEnrollment | 0.026 | 1.027 |
| SL.glm | screen_glmnet | 0.167 | (Intercept) | -19.395 | 0 |
| SL.glm | screen_glmnet | 0.167 | EthnicityNotreported | 0.015 | 1.015 |
| SL.glm | screen_glmnet | 0.167 | EthnicityUnknown | 0.052 | 1.053 |
| SL.glm | screen_glmnet | 0.167 | Black | -0.136 | 0.873 |
| SL.glm | screen_glmnet | 0.167 | Asian | -0.067 | 0.936 |
| SL.glm | screen_glmnet | 0.167 | NatAmer | -0.055 | 0.947 |
| SL.glm | screen_glmnet | 0.167 | Multiracial | -0.04 | 0.961 |
| SL.glm | screen_glmnet | 0.167 | Notreported | -1.403 | 0.246 |
| SL.glm | screen_glmnet | 0.167 | Unknown | -0.056 | 0.946 |
| SL.glm | screen_glmnet | 0.167 | URMforsubcohortsampling | 0.03 | 1.031 |

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

| Learner | Screen | Weight | Predictors | Coefficient | Odds.Ratio |
|---------|---------------|--------|------------------------|-------------|------------|
| SL.glm | screen_glmnet | 0.167 | HighRiskInd | 0.379 | 1.461 |
| SL.glm | screen_glmnet | 0.167 | Sex | 12.07 | 174555.444 |
| SL.glm | screen_glmnet | 0.167 | Age | 0.354 | 1.425 |
| SL.glm | screen_glmnet | 0.167 | BMI | -0.046 | 0.955 |
| SL.glm | screen_glmnet | 0.167 | Country | -0.013 | 0.987 |
| SL.glm | screen_glmnet | 0.167 | HIVinfection | 0.011 | 1.011 |
| SL.glm | screen_glmnet | 0.167 | CalendarDateEnrollment | 0.026 | 1.027 |



Figure 1.4: Superlearner predicted probabilities of COVID-19 disease in vaccinees 7 days after Day 29 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/commit/dc78cfe664d840e1ab59dedfe1bda41025af2f13) repository with commit hash dc78cfe664d840e1ab59dedfe1bda41025af2f13. A diff of the changes introduced by that commit may be viewed at https://github.com/CoVPN/correlates_reporting/commit/dc78cfe664d840e1ab59dedfe1bda41025af2f13
- The sha256 hash sum of the raw input file, “COVID_ENSEMBLE_practicedata.csv”:
186c6fefe0d7ee781c3b0bf5cedee3686398eb67ddc3da19dfa0c9bf3ad88fcc
- The sha256 hash sum of the processed file, “janssen_pooled_mock_data_processed.csv”:
49456667a8b5fd8daf3e3a644a1e9ef5db530fb94e73d3d357e4a76b30962b8a