

COVID-19 Baseline Risk Score Analysis Report

MockENSEMBLE Study

USG COVID-19 Response Biostatistics Team

September 16, 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 | Binary input variable/s having ≤ 3 cases in the variable = 1 or 0 subgroup and dropped from analysis (sorted by number of cases in Variable = 1 subgroup). | 11 |
| 1.3 | All learner-screen combinations (14 in total) used as input to the Superlearner. | 12 |
| 1.4 | Weights assigned by Superlearner. | 16 |
| 1.5 | Learners assigned weight > 0.0 by Superlearner sorted by weight. Predictors within each learner are sorted by variable importance which is the absolute value in Coefficient (in case of learners like SL.glm, SL.gam, SL.glm.interaction), or Gain (in case of SL.xgboost) or Importance (in case of SL.ranger.imp). | 17 |
| 1.6 | Cases per treatment arm prior to risk score analysis. | 20 |
| 1.7 | Cases per treatment arm post risk score analysis. | 20 |

List of Figures

| | | |
|-----|--|----|
| 1.1 | Cross-validated AUC (95% CI) of algorithms for predicting COVID-19 disease status starting 7 days after Day 29. CV-AUCs were computed using only data from the placebo arm. | 13 |
| 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. CV-estimated predicted probabilities were computed using only data from the placebo arm. | 14 |
| 1.3 | ROC curves based off CV-estimated predicted probabilities for the top 2 learners, Superlearner and Discrete SL. CV-estimated predicted probabilities were computed using only data from the placebo arm. | 15 |
| 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) | 0/19268 (0.0%) | NA |
| Black | Indicator race = Black (1=Black, 0=complement) | 0/19268 (0.0%) | NA |
| Asian | Indicator race = Asian (1=Asian, 0=complement) | 0/19268 (0.0%) | NA |
| NatAmer | Indicator race = American Indian or Alaska Native (1=NatAmer, 0=complement) | 0/19268 (0.0%) | NA |
| Multiracial | Indicator race = Multiracial (1=Multiracial, 0=complement) | 0/19268 (0.0%) | NA |
| URMforsubcohortsampling | Indicator of under-represented minority (1=Yes, 0=No) | 0/19268 (0.0%) | NA |
| HighRiskInd | Baseline covariate indicating ≥ 1 Co-existing conditions (1=yes, 0=no, NA=missing) | 0/19268 (0.0%) | NA |
| HIVinfection | Indicator HIV infected at enrollment (1=infected, 0=not infected) | 0/19268 (0.0%) | NA |
| Sex | Sex assigned at birth (1=female, 0=male/undifferentiated/unknown) | 0/19268 (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 . | 0/19268 (0.0%) | NA |
| 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 | 0/19268 (0.0%) | NA |
| Country.X1 | Indicator country = Argentina (1 = Argentina, 0 = complement) | 0/19268 (0.0%) | NA |
| Country.X2 | Indicator country = Brazil (1 = Brazil, 0 = complement) | 0/19268 (0.0%) | NA |
| Country.X3 | Indicator country = Chile (1 = Chile, 0 = complement) | 0/19268 (0.0%) | NA |
| Country.X4 | Indicator country = Columbia (1 = Columbia, 0 = complement) | 0/19268 (0.0%) | NA |
| Country.X5 | Indicator country = Mexico (1 = Mexico, 0 = complement) | 0/19268 (0.0%) | NA |

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

| Variable.Name | Definition | Total.missing.values | Co |
|-------------------|--|----------------------|----|
| Country.X6 | Indicator country = Peru (1 = Peru, 0 = complement) | 0/19268 (0.0%) | NA |
| Country.X7 | Indicator country = South Africa (1 = South Africa, 0 = complement) | 0/19268 (0.0%) | NA |
| Region.X1 | Indicator region = Latin America (1 = Latin America, 0 = complement) | 0/19268 (0.0%) | NA |
| Region.X2 | Indicator country = Southern Africa (1 = Southern Africa, 0 = complement) | 0/19268 (0.0%) | NA |
| CalDtEnrollIND.X1 | Indicator variable representing enrollment occurring between 4-8 weeks periods of first subject enrolled (1 = Enrollment between 4-8 weeks, 0 = complement). | 0/19268 (0.0%) | NA |

Note:

1. Binary input variable/s PacIsl, Notreported, Unknown had ≤ 3 cases in the variable = 1 or 0 subgroup and dropped.
2. No input variable had more than 5% missing values.
3. No variable had less than 5% missing values to activate imputation.

Table 1.2: Binary input variable/s having ≤ 3 cases in the variable = 1 or 0 subgroup and dropped from analysis (sorted by number of cases in Variable = 1 subgroup).

| Variable Name | Definition | Variable = 0 subgroup N (cases) | Variable = 1 subgroup N (cases) |
|---------------|---|------------------------------------|------------------------------------|
| Unknown | Indicator race = unknown (1=Unknown, 0=complement) | 19150 (513) | 118 (3) |
| PacIsl | Indicator race = Native Hawaiian or Other Pacific Islander (1=PacIsl, 0=complement) | 19232 (514) | 36 (2) |
| Notreported | Indicator race = Not reported (1=Notreported, 0=complement) | 19151 (515) | 117 (1) |

Table 1.3: 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. CV-AUCs were computed using only data from the placebo arm.



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. CV-estimated predicted probabilities were computed using only data from the placebo arm.



Figure 1.3: ROC curves based off CV-estimated predicted probabilities for the top 2 learners, Superlearner and Discrete SL. CV-estimated predicted probabilities were computed using only data from the placebo arm.

Table 1.4: Weights assigned by Superlearner.

| Learner | Screen | Weight |
|----------------|---------------------------------|---------------|
| SL.glm | screen_univariate_logistic_pval | 0.935 |
| SL.mean | screen_all | 0.065 |
| SL.glm | screen_all | 0.000 |
| SL.glm | screen_glmnet | 0.000 |
| SL.glm | screen_highcor_random | 0.000 |

Table 1.5: Learners assigned weight > 0.0 by Superlearner sorted by weight. Predictors within each learner are sorted by variable importance which is the absolute value in Coefficient (in case of learners like SL.glm, SL.gam, SL.glm.interaction), or Gain (in case of SL.xgboost) or Importance (in case of SL.ranger.imp).

| Learner | Screen | Weight | Predictors | Coefficient | Odds.Ratio |
|---------|---------------------------------|--------|-------------|-------------|------------|
| SL.glm | screen_univariate_logistic_pval | 0.935 | (Intercept) | -3.738 | 0.024 |
| SL.glm | screen_univariate_logistic_pval | 0.935 | HighRiskInd | 0.424 | 1.528 |
| SL.glm | screen_univariate_logistic_pval | 0.935 | Age | 0.344 | 1.410 |
| SL.glm | screen_univariate_logistic_pval | 0.935 | Sex | -0.084 | 0.919 |



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.

Table 1.6: Cases per treatment arm prior to risk score analysis.

| Study-Arm | Non-Cases | Post-Day 29-Cases |
|-----------|-----------|-------------------|
| Placebo | 18752 | 516 |
| Vaccine | 19077 | 167 |

Table 1.7: Cases per treatment arm post risk score analysis.

| Study-Arm | Non-Cases | Post-Day 29-Cases |
|-----------|-----------|-------------------|
| Placebo | 18752 | 516 |
| Vaccine | 19077 | 167 |

```
#> [1] "running references ~~~~~"
```

Chapter 2

Appendix

- This report was built from the [CoVPN/correlates_reporting](https://github.com/CoVPN/correlates_reporting) repository with commit hash 2066a047403ded9f46304c03353d37ddd93dfb5d. A diff of the changes introduced by that commit may be viewed at https://github.com/CoVPN/correlates_reporting/commit/2066a047403ded9f46304c03353d37ddd93dfb5d
- The sha256 hash sum of the raw input file, “COVID_ENSEMBLE_practicedata.csv”:
c5c374aafab433f963f8b9a6426b1ff1b94a81450990e6cf9e574b9f08a48187
- The sha256 hash sum of the processed file, “janssen_pooled_mock_data_processed.csv”:
5d6af1d6b6307d64f61e32e01a297faedb5e41c17bcdabc7807f46f4f8200c75