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

Table 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	NA	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	NA	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 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

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

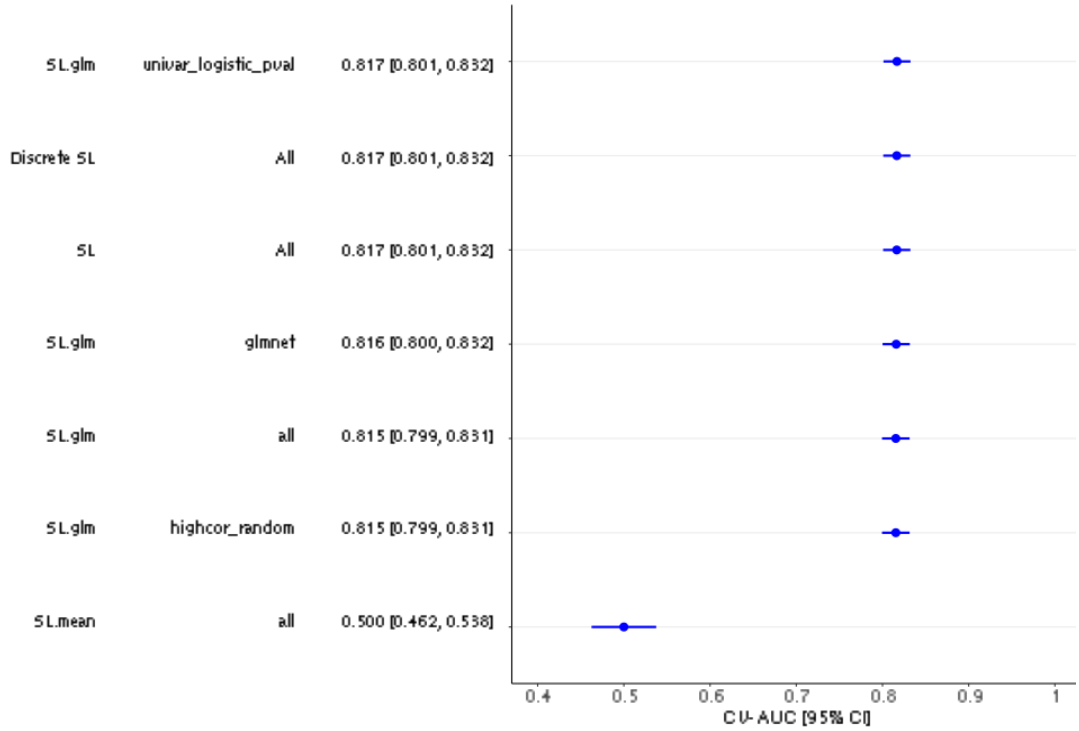

Figure 1: Risk score analysis with EventIndPrimaryD57 as the outcome: Plot shows CV-AUC point estimates and 95% confidence intervals for the Super Learner and all models trained to classify cases in placebo group defined by EventIndPrimaryD57. Learners are sorted by their CV-AUC point estimates.

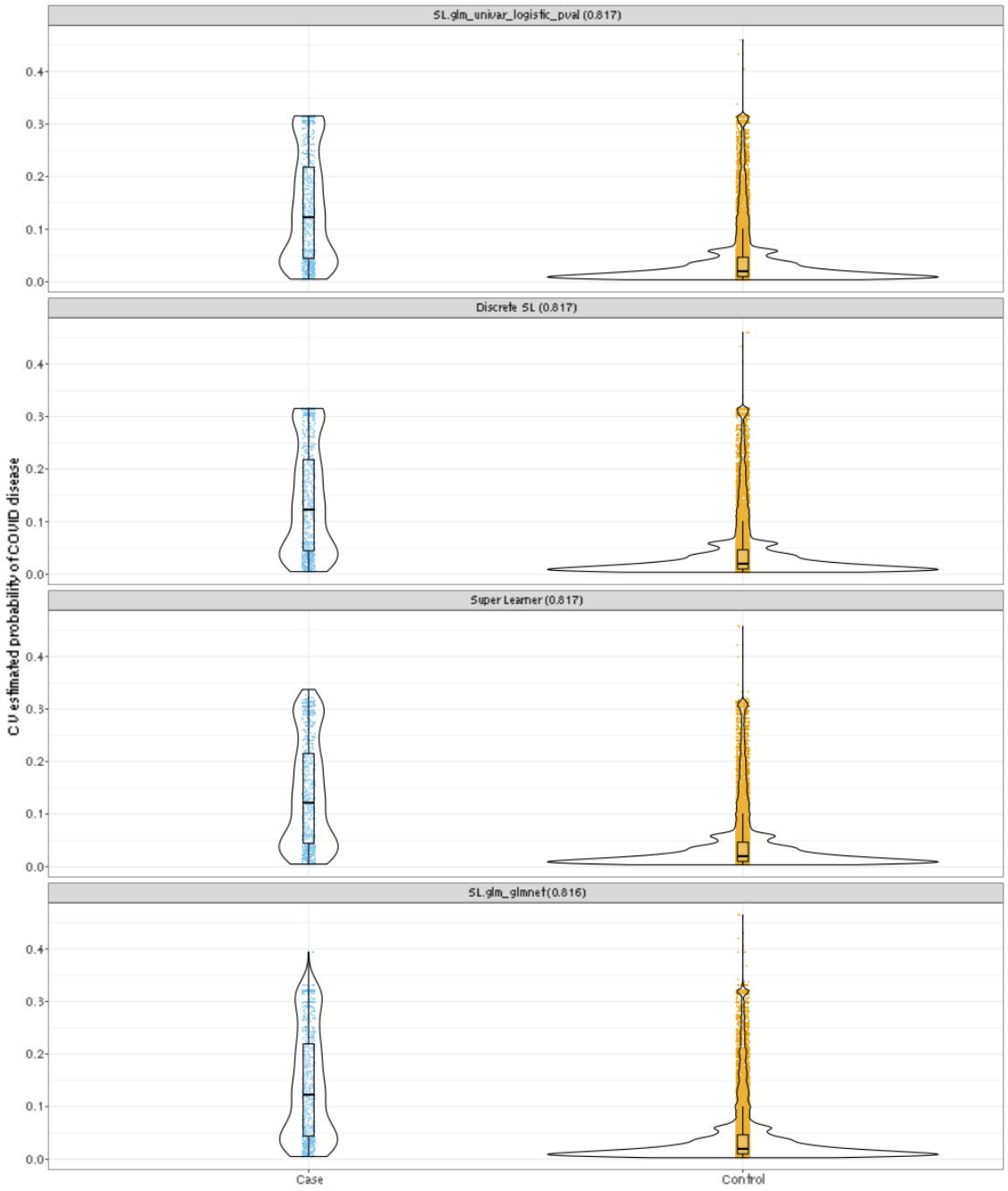

Figure 2: Plots showing CV estimated probabilities of COVID disease split by cases and controls based off EventIndPrimaryD57 for the top 2 learners, SuperLearner and Discrete SL.

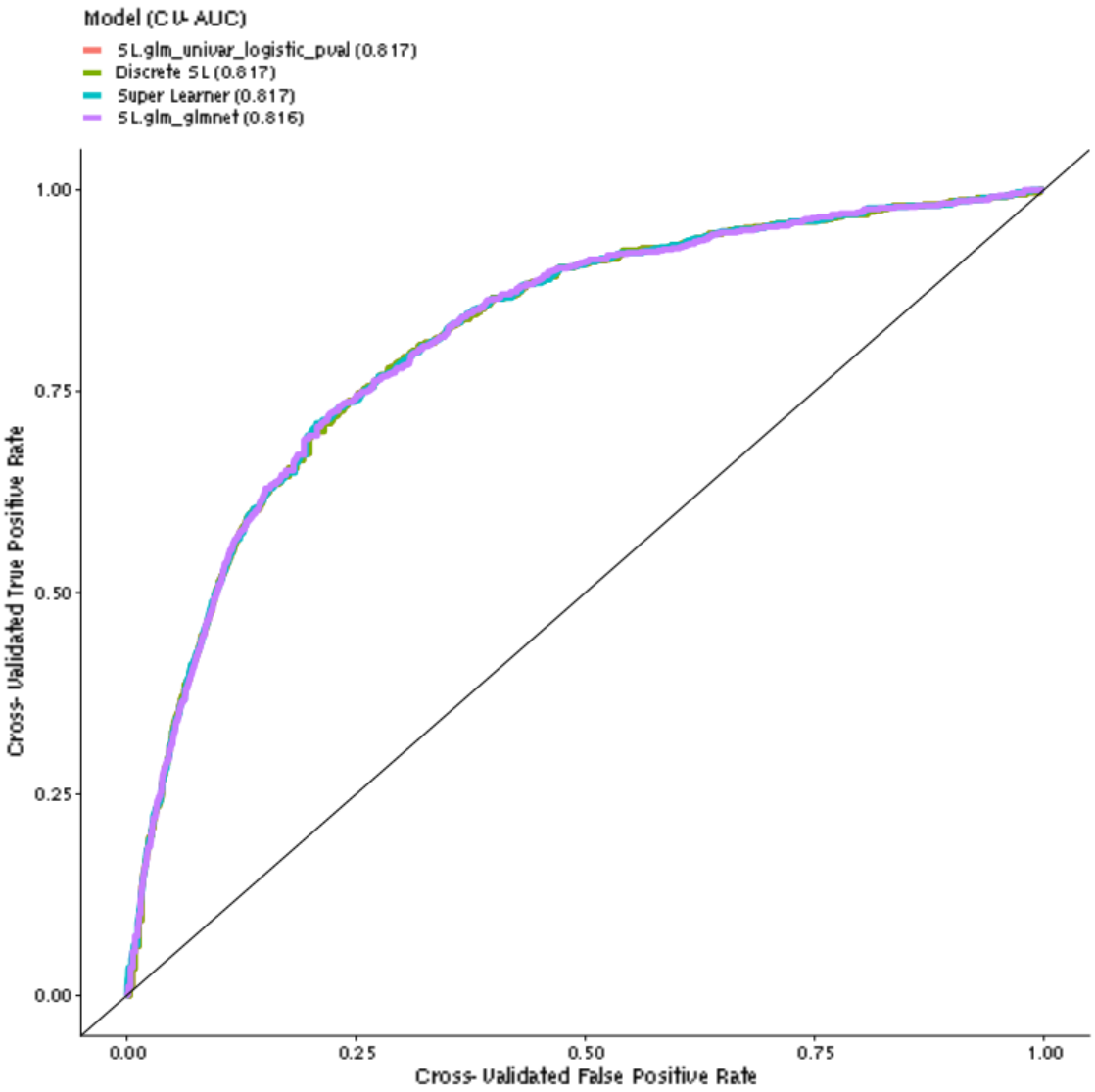

Figure 3: ROC curves for the top 2 learners, SuperLearner and Discrete SL.

Table 3: Weights assigned by Superlearner (risk score analysis).

Learner	**Weights**
screen_univariate_logistic_pval_SL.glm	0.993
screen_all_SL.mean	0.007
screen_all_SL.glm	0.000
screen_glmnet_SL.glm	0.000
screen_highcor_random_SL.glm	0.000