Supplemental material 1

Model fitting

Prior to model fitting, covariates were scaled to zero mean and unit variance, after model fitting coefficients were unscaled to obtain regression coefficients on the original scale. In adaptive lasso, weights are applied to each of the covariates present in the lasso constraint, the weight vector has to be calculated before the adaptive lasso regression is performed. Due to multicollinearity between laboratory tests in the routine lab panel, weights in the adaptive lasso were based on ridge regression estimates ($\hat{\beta}_{ridge}$) as recommended by Zou. To obtain $\hat{\beta}_{ridge}$ the optimal penalty (λ) for the ridge regression was chosen using 10 fold cross-validation (CV) with area under the ROC curve (AUC) as the loss function. The λ corresponding to the maximum AUC was selected to obtain $\hat{\beta}_{ridge}$. The weight vector ($\widehat{w}$) was calculated by $\widehat{w} = 1/\left|\hat{\beta}_{ridge}\right|^2$. This weight vector was then used to fit an adaptive lasso regression where λ was chosen by the criterion ±1 SE of the maximum AUC.

Model intercept correction

The linear predictor for a patient *i* is calculated as follows: $lp_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_n x_{in}$ Where *n* is the number of variables in the final model, x_{in} are the observed predictor variables for subject *i* and $\beta_n$the model coefficients. The linear predictor can then be converted to a probability for patient *i* (*Pi*) by the logistic function: $P_i = \frac{1}{1+e^{-lp_i}}$

The intercept term β_0 is sensitive to the fraction of cases versus controls in the dataset/population. Since the model is fitted to a case-control dataset where the number cases is fixed (all patients tested positive for COVID-19) and the number of controls is randomly chosen (a 6-month period pre-COVID), the intercept term β_0 is a result of this choice and will likely not be generalizable to the real-world setting. Prior correction is a method to correct the estimate of the intercept based on the true fraction of positives in the population, τ (prevalence of COVID-19 in the ED) and the fraction of cases in the development dataset, $\bar{y}$. The intercept term β_0 can then be corrected to obtain $\beta_{0corrected}$ using the following formula:

$$\beta_{0corrected} = \beta_0 + \beta_{adj}$$

$$\beta_{adj} = -ln\left[\left(\frac{1-\tau}{\tau}\right)\left(\frac{\bar{y}}{1-\bar{y}}\right)\right]$$

In our dataset $\bar{y} = 0.02675$ therefore:

$$\beta_{adj} = -ln\left(\frac{1-\tau}{\tau}\right) + 3.594$$

An estimate $\bar{\tau}$ can be used for the prevalence τ to obtain $\bar{\beta}_{adj}$ which can be plugged in the original linear predictor formula to obtain calibrated probabilities:

$$lp_i(\tau) = \beta_0 - ln\left(\frac{1-\tau}{\tau}\right) + 3.594 + \beta_1 x_{i1} + \cdots + \beta_n x_{in}$$

CoLab-score

An alternative, which is the basis of the CoLab-score, is to choose a fixed probability P_i above which one considers a patient eligible for further testing. The probability can be expressed as a number needed to test. If one is willing to test 10 patients to find one positive, all patients with $P_i \geq 0.1$ should be considered positive. In this study a number needed to test of 15 is used, therefore all patients with a $P_i \geq 0.067$ should be considered positive. On the linear predictor scale this translates to $\text{logit}(0.067) = -2.639$. To determine the cutoffs for difference prevalence thresholds one solves the following equation:

$$\beta_0 + \beta_{adj} + \beta_1 x_{i1} + \cdots + \beta_n x_{in} \geq -2.639$$
$$\beta_0 + \beta_1 x_{i1} + \cdots + \beta_n x_{in} \geq -2.639 - \beta_{adj}$$
$$lp_i(\tau) \geq ln\left(\frac{1-\tau}{\tau}\right) - 6.233$$

Choosing values for τ yields the cutoffs for the CoLab score:

$lp_i(\tau = 0.4) \geq -5.83$ (CoLab-score = 1)
$lp_i(\tau = 0.1) \geq -4.03$ (CoLab-score = 2)
$lp_i(\tau = 0.05) \geq -3.29$ (CoLab-score = 3)
$lp_i(\tau = 0.02) \geq -2.34$ (CoLab-score = 4)
$lp_i(\tau = 0.01) \geq -1.64$ (CoLab-score = 5)

These thresholds correspond to CoLab-scores 0 to 5. The interpretation of these scores is as follows; if the prevalence is <1%, only CoLab-score 5 should be classified as positive and CoLab-score 0 till 4 as negative. If the prevalence is 1% – 2%, CoLab-score 4 and 5 should be classified as positive and 1 – 3 negative. Similarly, with a prevalence of 2 – 5% the split is between CoLab-score 2 and 3 and with prevalence of 5 – 10% between CoLab-score 1 – 2. If the prevalence is higher than 10% only CoLab-score 0 is classified as negative. Using the CoLab-score in this fashion, aims to preserve a number need to test of 15.

Relative importance of variables

Since the variables included in the model are on different scales, the magnitude of the unscaled coefficients cannot be used to compare the importance of variables to each other. To give some indication of the importance of the variables in predicting the outcome, the unscaled coefficients obtained from the adaptive lasso regression were used to calculate the relative importance. The variable with the highest unscaled coefficient was used as maximum ($\beta_{unscaled,max}$), and all other scaled coefficients were divided by this maximum and multiplied by 100 to obtain the relative importance in %: $\frac{\beta_{unscaled}}{\beta_{unscaled,max}} \cdot 100$.