Draft: Combining Rapid Antigen Testing and Syndromic Data Improves Sensitivity and Specificity in Real-World COVID-19 Detection

Contents

Abstract (288 words)

Background

The majority of the world's population live in low- and middle-income countries (LMICs) where access to gold-standard diagnostics like RT-PCR is often limited. Rapid antigen testing (RAT) and syndromic diagnosis are two alternative, inexpensive and easy-to-deploy surveillance methods but there are concerns that they lack the sensitivity and specificity to effectively guide practice.

Methods

Bangladesh's Institute of Epidemiology Disease Control And Research (IEDCR) identified potential COVID-19 patients in Dhaka using syndromic surveillance. A sample (n = 1172) of these patients was tested using RAT and syndromic data were collected. Models were fit to predict RT-PCR status using the RAT data, the syndromic data, and the two combined. Model performance was measured using predictive power and classification performance under three epidemiological scenarios: "Agnostic", "Rising Cases" and "Low-Level Cases".

Findings

Combined data models yielded equal or improved performance over syndromic- and RAT-only models across all three epidemiological scenarios and when compared as more generic prediction and classification engines. In the "Rising Cases" scenario, which most closely represents the current situation in many LMICs, the combined data model false negative rate is 26.2012339 percentage points lower that of the RAT only model.

Although the syndromic only model matches the combined models false negative rate, its false positive rate is 30.8213341 percentage points higher.

Interpretation

A few accurate tests may be less useful at the population level than many more imperfect ones. Small, scalable improvements in the accuracy of mass-deployed but imperfect tests can then make a very big difference for pandemic control.
We demonstrate that such improvements can be achieved by statistically utilising complementary strengths and weaknesses across two imperfect diagnostics, we can greatly improve the detection of COVID-19.

Funding

The Bill and Melinda Gates Foundation and the Wellcome Trust.

Introduction (1080 Words)

Identification and isolation of COVID-19 cases remains key to the pandemic response across the globe. The faster and more accurately we can identify cases, the more effectively we can provide clinical care, reduce transmission of infection and develop population-level interventions. RT-PCR testing has rapidly become the default, gold-standard test for COVID-19 in applied settings (although see [@drame2020should]) due to its high sensitivity and specificity for COVID-19 [@corman2020detection, @tahamtan2020real]. Most of the world's population, however, live in low- and middle-income countries (LMICs) where the laboratory facilities needed to carry out RT-PCR tests are often scarce and hard to reach [@chowdhury2020long, @vandenberg2021considerations], and the majority of patient diagnosis and support comes from telemedicine or community support teams (CSTs) composed of local volunteers with basic training. COVID-19 diagnosis worldwide, therefore, must be made accessible using inexpensive methods that can be carried out locally [@cash2020has, @olalekan2020covid].

An increasingly popular alternative to RT-PCR is rapid antigen testing (RAT) [@linares2020panbio]. Like RT-PCR, these tests have high specificity for COVID-19 while being less expensive, easier to implement, and faster [@boum2021performance]. For RT-PCR testing, patients must travel to a designated site or have officials visit their home in enhanced personal protective equipment. Then, invasive nasopharyngeal swabs must be taken. In contrast, RATs can be conducted on nasal swabs, completed in the home with minimal PPE, and results are available in 30 minutes. RATs can be taken by persons with limited training, thus decreasing the time and expense associated with identifying cases. Together, these traits make RATs an appealing alternative to RT-PCR, however, concerns have been raised that the lower sensitivity of RAT [@mak2020evaluation, muhi2021multi] leads to more false negative diagnoses.

Another diagnostic that has been used since the start of the pandemic is symptom-thresholding [@jin2020rapid]. Here, a patient presenting with a fever and one or more symptoms is treated as a COVID-19 positive patient. The main advantage of this approach is the ease of implementation. As with RAT, symptom-thresholding is faster, cheaper and less invasive than RT-PCR. Unlike RAT, symptom-thresholding can be scaled immediately at the onset of a pandemic, however, it is also reliant on thresholds developed then. These thresholds were necessarily drawn from clinical intuition, rather than data, often for different variants and populations than they are now applied to. Consequently, the relationship between the thresholds and the true COVID-19 status is often weak, with low specificity leading to a very large number of false positive diagnoses. A natural extension, therefore, is syndromic modelling. In this approach, rather than using a set of pre-determined thresholds, a range of symptomatic and risk factor data (such as age and gender) are collected and then a sub-sample of patients is tested using RT-PCR for validation [@sim2021utilizing]. These data are used to fit a model that allows more accurate prediction of how likely a patient is to have COVID-19 through the identification of COVID-19 syndromes [@undurraga2021covid, @wenham2020covid].

JMC Not an official term

It is worth highlighting at this point that in resource-limited settings there is very limited provision for testing of asymptomatic cases, despite their important role in disease transmission [@mayorga2020modelling] . Even while focusing solely on symptomatic patients, syndromic modelling is a complex and nuanced task. Disease syndromes can change between populations, when new variants emerge, and as other diseases become more

JMC Need to highlight that we are only dealing with symptomatic patients, this felt a good place to do or less common [@garry2020considerations]. These changes can make syndromic models generalise poorly. For example, if another disease for which loss of smell is a symptom becomes common, loss of smell is no longer strongly indicative of COVID-19. Similarly, if everyone who presents has a cough, regardless of their COVID-19 status, then coughing will show no relationship with COVID-19 (even if the two are strongly related in the general population). Furthermore, symptoms do not always occur in isolation, some, like loss of smell and loss of taste, are strongly related. Unfortunately, the majority of syndromic modelling methods currently used do not account for these complexities. Even where they can, at least partially, be accounted for, the many types of common respiratory disease generally means that syndromic modelling still tends to have quite low specificity [@garry2020considerations].

JMC query cross-sectional sampling

Moderate to poor sensitivity and specificity are problematic in diagnostics but may be tolerable depending on their scale and impact given the local situation. Low specificity means a patient is likely to be told they have COVID-19 when they do not (a high false positive rate), leading to patients unnecessarily self-isolating and receiving support. This is expensive to the individual and to local public health bodies, reducing available resources for those who need them [@surkova2020false]. Similarly, low sensitivity means more patients being told they do not have COVID-19 when they actually do (a high false negative rate), leading to the individual not getting appropriate support or taking action to prevent the disease spreading further [@west2020covid]. Although the default approach is generally to minimise both misclassification rates (our "Agnostic" scenario below), the true costs of these misclassifications will depend on local context. When the prevalence of the disease is low, false negatives will be correspondingly low and false positives may create local scepticism leading to poor adherence longer term. In this situation (our "Low-Level Cases" scenario), false positives might be more costly than false negatives [@surkova2020false]. If the disease is abundant or increasing rapidly then changes in the false negative rate might have an outsized impact on the pandemic trajectory and thus be more costly, as in our "Rising Cases" scenario. Often the situation will be even more nuanced and a different balance will need to be struck [@vandenberg2021considerations].

The "best" diagnostic, therefore, is not a single universal test. The two dominant testing methods available in LMICs when not adapted for the local situation are highly flawed. Relying solely on symptomatic diagnosis will likely overestimate the number of individuals with COVID-19 due to its lack of specificity. Conversely, RATs will give a false impression of control due to the number of positive cases that will be missed. In this paper, we demonstrate that by combining these two testing methods we can utilise their complementary strengths, ameliorate their respective weaknesses, and optimise them for different epidemiological scenarios. We aim to compare the performance of these two testing methods and the combined approach both in terms of general prediction and as diagnostics under three epidemiological scenarios with different misclassification requirements. We show that the optimised combined data models achieve equal-to-much-lower error rates than the next best method in all metrics. We then discuss the role of statistically integrating data from multiple imperfect testing methods in resource limited settings to improve the diagnosis of diseases, particularly COVID-19.

Methods (965 words)

Data Collection

Participants included in this study were identified for COVID-19 testing by community support teams (CSTs). Recruitment took place across Dhaka (the capital city of Bangladesh) between 19th May 2021 and 11th July 2021.

Patients were selected for further testing if they had a fever (>38°C) at the point of testing and one or more of 14 symptoms associated with COVID-19 (breathing problems, coughing, diarrhoea, fever (ongoing), a headache, loss of taste, loss of smell, muscle pain, red eyes, a runny nose, a sore throat, tiredness, vomiting or a wet cough). If selected, the CSTs collected the patient's age and gender, and took two nasal swabs.

One swab each was used for rapid antigen testing (RAT) and RT-PCR. The full questionnaire and testing protocols are provided in Supplementary 1. Participants provided written informed consent to sample collection and for their results to be analyzed in the study.

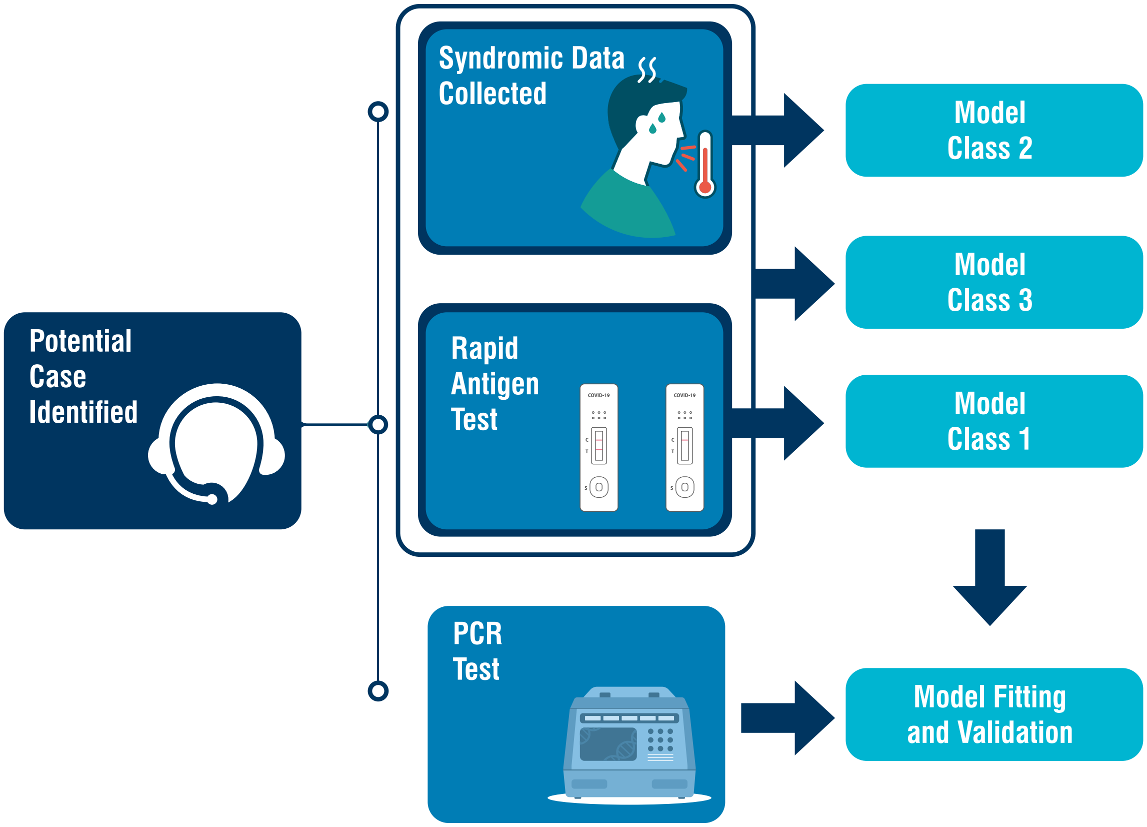


Figure 1: Schematic description of identification of likely COVID-19 patients by community support teams (CSTs), swab collection and model definitions. The teams collected syndromic data (age, gender and presence/absence of 14 predetermined symptoms), and two sets of naso-pharyngeal swabs (one each for Rapid Antigen Testing and RT-PCR). We then used rapid antigen testing (RAT) and syndromic data, two imperfect but inexpensive diagnostics, to generate three model classes: RAT result only in Model Class 1, syndromic data only in Model Class 2, and both RAT result and syndromic data in Model Class 3. The RT-PCR test result is used to train and test each model using temporal cross-validation.

Modelling

Structure

We examined the ability of the two imperfect identification methods, syndromic modelling and RAT, to predict the patient's COVID-19 status when used separately and together. These combinations define three model classes (Figure @ref(fig:data-flowchart)).

Find more ways to cite Figure ref(fig:data-flowchart) in this text.

Model Class 1 uses only the RAT result. It equates being RAT-positive with the patient being PCR-positive for COVID-19 (hereafter, PCR-positive), and being RAT-negative with PCR-negativity.

Model Class 2 uses only the syndromic data. For this model, we used a Bayesian multivariate probit model [@albert1993bayesian]. The multivariate probit structure allows the model to account for the binary and correlated nature of the symptoms while conditioning on the risk factors of age and gender. By using a Bayesian formulation, we are able to quantify uncertainty in the parameter estimates.

Model Class 3 combines the two data sources. We utilise the specificity of RAT by treating RAT-positive patients as PCR-positive patients. The RAT-negative patients are modelled using the sensitive syndromic approach using Model Class 2 to capture PCR-positive patients that are missed by the RAT. This approach leverages the potential different syndromic profiles of PCR-positve patients who are RAT-positive and -negative, allowing the model to adapt solely to the latter. The models were fitted to the data using Bayesian inference techniques based on Hamiltonian Monte Carlo in the Stan programming language [@carpenter2017stan].

Model Selection

We conducted backwards model selection (starting with the most complex, biologically plausible model) to identify a subset of models with the highest predictive power under temporal cross-validation (Figure @ref(fig:modsel-flowchart)). Reducing the number of possible models was necessary to reduce computational demand and reduce the risk of overfitting models to the test scenarios. The large number of symptoms corresponds to a high number of potential model configurations (>131 000 for 14 symptoms and two covariates) which might perform well on the test sets (even under the challenging conditions of temporal cross-validation) but lack transferability. By first using general predictive power to narrow down the number of candidate models and then testing those models, we are more likely to choose models which generalise well to new data. The number of candidate models used was not pre-determined but it was clear when fitting the models that there were "jumps" in performance (as defined below) between models containing five and four symptoms, so the models with one to four symptoms were used as the candidate models. Zero symptom models were not included in the analysis as they do not correspond to a feasible policy (with covariates they would require governments to ask individuals of a given gender and age as COVID-19 positive, and without covariates they would involve randomly assigning individuals as COVID-19 positive).

Predictive Performance

We scored the models' predictive power using cross-entropy. Cross-entropy measures the accuracy of models that generate probabilities of binary outcomes, rather than make binary classifications, similar in concept to a mean square error for normally-distributed data, but adapted for binary data [@gneiting2007strictly]. A cross-entropy value close to zero corresponds to high levels of accuracy, with larger values indicating lower accuracy. More details on the model structure and selection process, including code, are available in Supplementary 2.

Classification Performance

In applied settings, models must often be evaluated on their performance as classifiers rather than just as prediction engines (i.e. their ability to say a patient is COVID-19 positive or negative, not simply the probability the patient might be COVID-19 positive or negative). To generate a classification, a probability threshold must be chosen over which patients are classified as COVID-19 positive.

Classifier performance was compared both generically (using receiver operating characteristic (ROC) curves [@hoo2017roc]) and under three epidemiological scenarios (using error terms described in Table

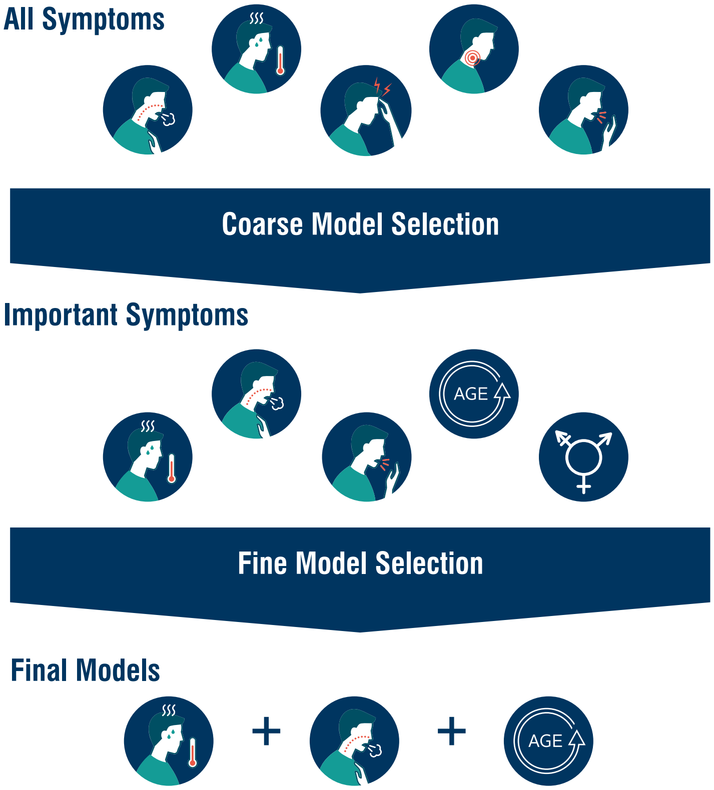


Figure 2: Schematic for rounds of model selection in the multivariate probit component of Model Classes 2 and 3. With 14 symptoms (only 5 shown here for demonstration purposes) and two covariates there are over 131000 possible model combinations. To make exploring these possible models computationally feasible and to reduce the risk of overfitting, we carried out two rounds of model selection. First, the data are divided into temporal cross-validation sets. The multivariate probit connects symptoms to the RT-PCR result through a correlation matrix. In the coarse model selection, the most complex feasible model (all symptoms and covariates) is fit to the training data. The estimated correlations between each symptom and the RT-PCR result are compared for each cross-validation set. The symptoms that have non-zero correlations in a systematic direction (i.e. all positively or all negatively correlated with RT-PCR result) are retained. The process is then repeated on each retained set of symptoms until the four symptoms in each model class with the strongest correlation to RT-PCR result. We then conduct a more exhaustive model selection on all the possible permutations of the four symptoms and two covariates. In this round, each model is fit to training data and used to predict for the test set, and the quality of those predictions is measured using cross-entropy scoring. The cross-entropy score is then used to select the best predictive model for each level of model complexity. Only these final models are then used for classification. This reduces the set of models tested as classifiers from >131 000 to just four per model class.

Table 1: For each epidemiological scenario there is a requirement and a performance criterion. The requirement refers to a base level of performance the model must achieve; in general this will be a maximum acceptable error rate of some kind. These requirements were determined in discussion with members of the Institute of Epidemiology, Disease Control and Research, Ministry of Health, Bangladesh (IEDCR). The requirement determines a probability threshold for each model which most closely meets that requirement. The performance criterion is then used to determine which model performs the 'best' given that the requirement has been met.

Scenario Name	Requirement	Performance Criterion (Error)
1 Agnostic	Maximise correct classification rates	Sum of error rates
2 Rising Cases	Max. 20% false negative rate	False negative rate
3 Low-Level Cases	Max. 20% false positive rate	False positive rate

@ref(tab:scenarios-tab)). We strongly emphasise that generic performance here is only used to show the flexibility of the model classes; the best model for a local situation can only be determined if the relative cost of false positives and false negatives is known.

In Scenario 1, we do not consider epidemiological context but simply minimise false negative and false positive rates equally. We do this by maximising the two correct classification rates both individually and in total, as measured by the harmonic mean (as opposed to the arithmetic mean which would only maximise the rates in total). Scenario 2 corresponds to the current situation in Bangladesh at time of writing (July 2021), with COVID-19 cases beginning to rapidly increase again. Under these circumstances, false negatives are extremely costly relative to false positives due to the exponential growth of the disease. In Scenario 3, the pandemic is not declining but maintaining a steady rate of cases. In this situation, policy-makers may be keen to keep false positive diagnoses low to prevent lockdown fatigue and to keep the workforce active. The requirements in Scenario 2 and 3 were developed in discussion with the Institute of Epidemiology, Disease Control and Research (IEDR), Bangladesh, for illustrative purposes.

Results (476 words)

Of 1241 subjects surveyed, a total of 1172 subjects had complete data available for the current analyses with the remainder removed due to duplication of barcodes or missing data. The mean age of women participants (47% of the sample) was 37 (SD = 14) years, and for men (53% of the sample) was 36 (SD = 14) years. Participants were identified by the community support teams (CSTs) and drawn from across Dhaka.

Model selection for both Model Class 2 (syndromic data only) and 3 (syndromic and RAT data) showed a marked decline in predictive power at more than 4 symptoms. The covariate gender was dropped for both model classes while age was dropped in Class 2 but retained in Class 3. The final four symptoms in order of importance (i.e. the most important symptom was retained in all of the final 4 models, the least important symptom was only retained in the 4 symptom model) were loss of taste, diarrhoea, vomit and fever for Model Class 2, and fever, wet cough, cough and loss of taste for Model Class 3.

In the comparison of model predictive performance, Model Class 1 (RAT only) performed worst with an out-of-sample cross-entropy of 3.24 (cross-entropy values further from zero correspond to worse predictive performance). The median cross-entropy values were between 2.53 and 2.59 for models in Class 2. Models in Class 3 performed best with cross-entropy values between 1.44 and 1.47 (see Figure @ref(fig:pred-perf)).

DHC out-of sample vs test set cross-entropy

Generic model classification performance is shown by their ROC curves (Figure @ref(fig:ROC-plot)).

Scenario specific classification performance is shown in Figure @ref(fig:scenario-plot). Across all three scenarios (defined in Table @ref(tab:scenarios-tab)), the best models in Class 3 performed equally well or better than the other two model classes. In Scenario 1 ("Agnostic"), models in Classes 1 and 3 performed equally well (overlapping posterior interquartile ranges) and distinctly better (no overlap in posterior interquartile range) than models in Class 2. In Scenario 2 ("Rising Cases"), Model Class 1 failed to meet the requirement and so was excluded, and Model Class 3 once again outperformed Class 2. In Scenario 3 ("Low-Level Cases"),

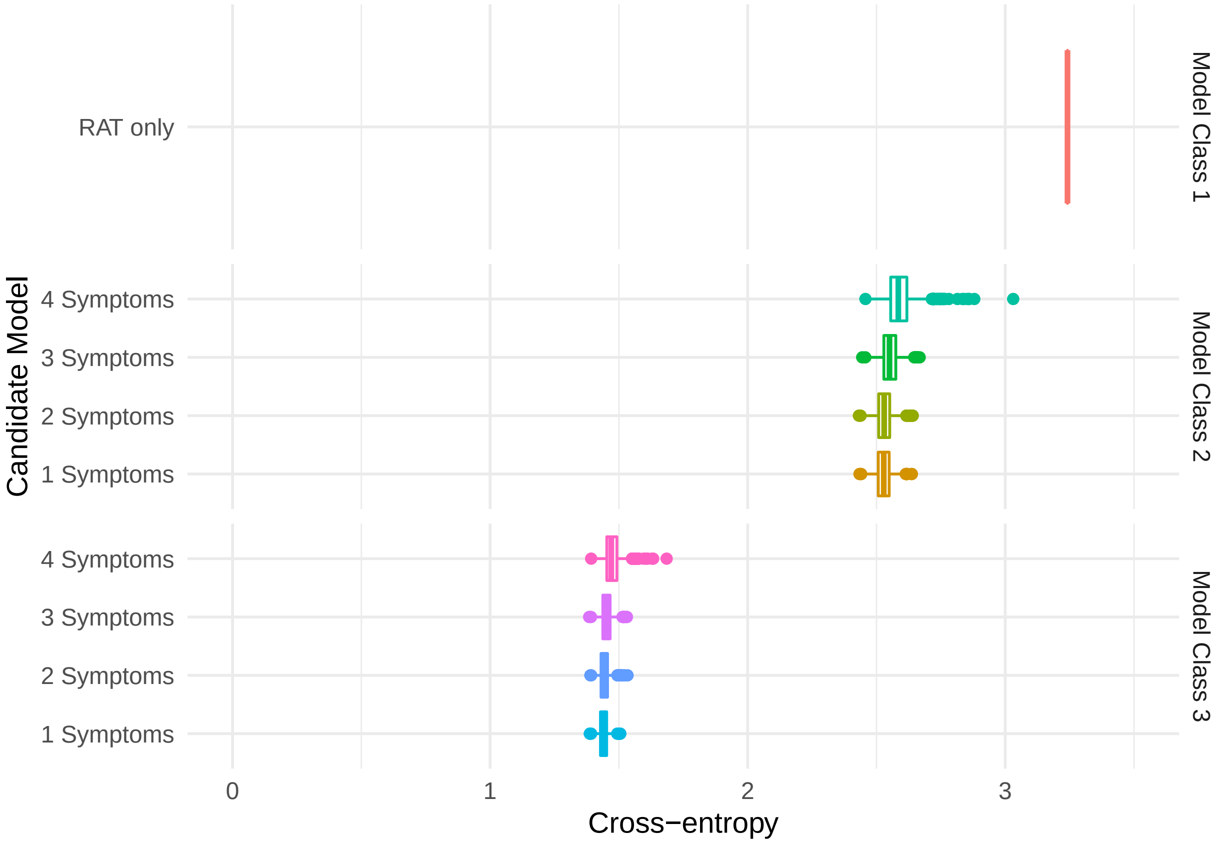


Figure 3: Predictive performance of candidate models. Interquartile ranges for the posterior cross-entropy of the best candidate models at each level of model complexity tested under temporal cross-validation. cross-entropy is a measure of distance from the truth, so values closer to zero indicate better models. The intermediate complexity models perform best at prediction, although performance is similar across all the models within each model class (1: rapid antigen testing (RAT) only; 2: syndromic data only; and 3: combined RAT and syndromic data).

Model Class 2 once again performed worst, and Model Class 3 achieved the lowest error, with Model Class 1 falling in between the two (closer to Class 3 than 2). For Classes 2 and 3 across all the scenarios the number of symptoms made relatively little difference within the final four candidate models in terms of median performance, although the more complex models have higher precision. It should be noted that the candidate models are chosen as a result of a selection process and performed much better than more complex models (i.e. those with 5 or more symptoms) or simpler models (with no symptoms but an intercept and covariates) in terms of cross-entropy and ROC, indicating they would likely also perform worse in these scenarios.

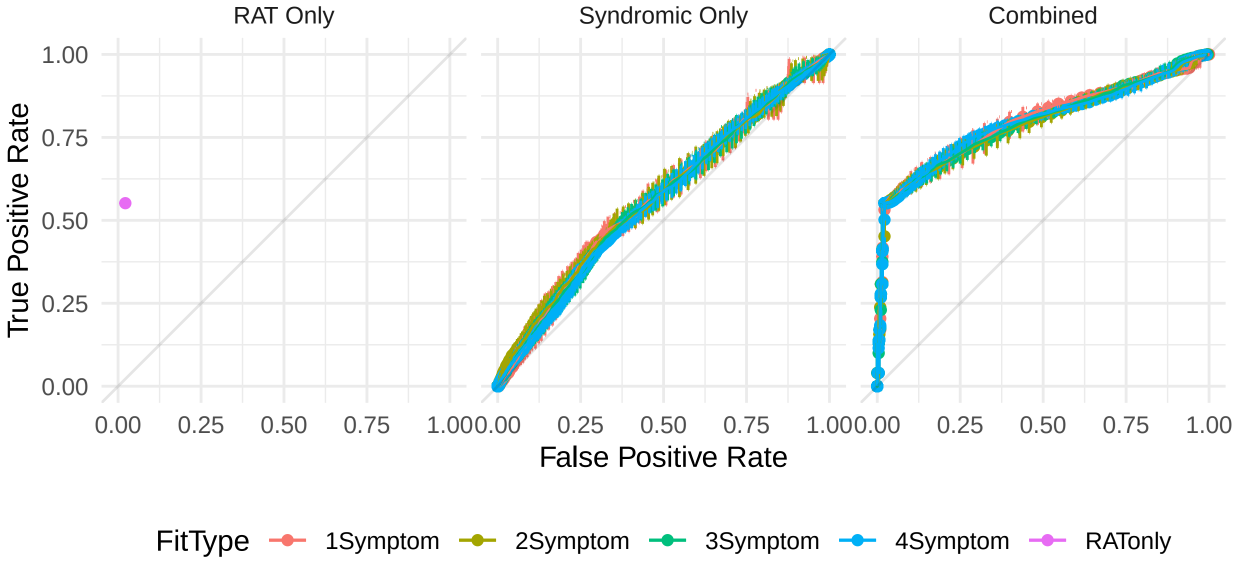


Figure 4: Receiver operating characteristics for rapid antigen testing (RAT) only approach (Model Class 1) and posterior median and interquartile range ROC for Class 2 (syndromic data only) and 3 (syndromic and RAT data) models. These curves demonstrate the performance of the model for any hypothetical scenario as defined by the axes (as opposed to Figure 5 which demonstrates model performance in specific epidemiological scenarios which are realisations of a single point in this space).

Discussion (815 words)

We have demonstrated that combining rapid antigen tests (RATs) with syndromic modelling yields better identification of COVID-19 cases than either diagnostic in isolation. These gains in performance are mirrored across metrics of prediction, generic classification and scenario-specific classification. The biggest improvement is seen in Scenario 2 ("Rising Cases") which was developed around the current situation in Bangladesh (see Table @ref(tab:scenarios-tab) where the pandemic is once again accelerating. In this scenario, the combined data model (Model Class 3) false negative rate is 26 percentage points lower that of the RAT only model (Model Class 1). Although the syndromic only model (Model Class 2) matches the combined models false negative rate, its false positive rate is 31 percentage points higher.

In a country where there are currently 15 000 new cases being identified every day, these improvements are non-trivial, representing tens of thousands of daily cases that would otherwise be missed. Futhermore, this boost in diagnostic performance is achieved with data that are already being collected in Bangladesh and other low- and middle- income countries (LMICs). Outwith developing and rerunning the models presented in this paper, these improvements are essentially cost-free and eminently scalable.

The pattern is similar in epidemiological Scenarios 1 ("Agnostic") and 3 ("Low-Level Cases"), with the combined model class performing performing equally well or better than the other two classes (Figure @ref(fig:scenario-plot)). These three scenarios only offer snapshots of performance. An indication of how these models will perform under any condition can be obtained by comparing the more generic model performance metrics for prediction and classification (Figures @ref(fig:pred-perf) and @ref(fig:ROC-plot), respectively).

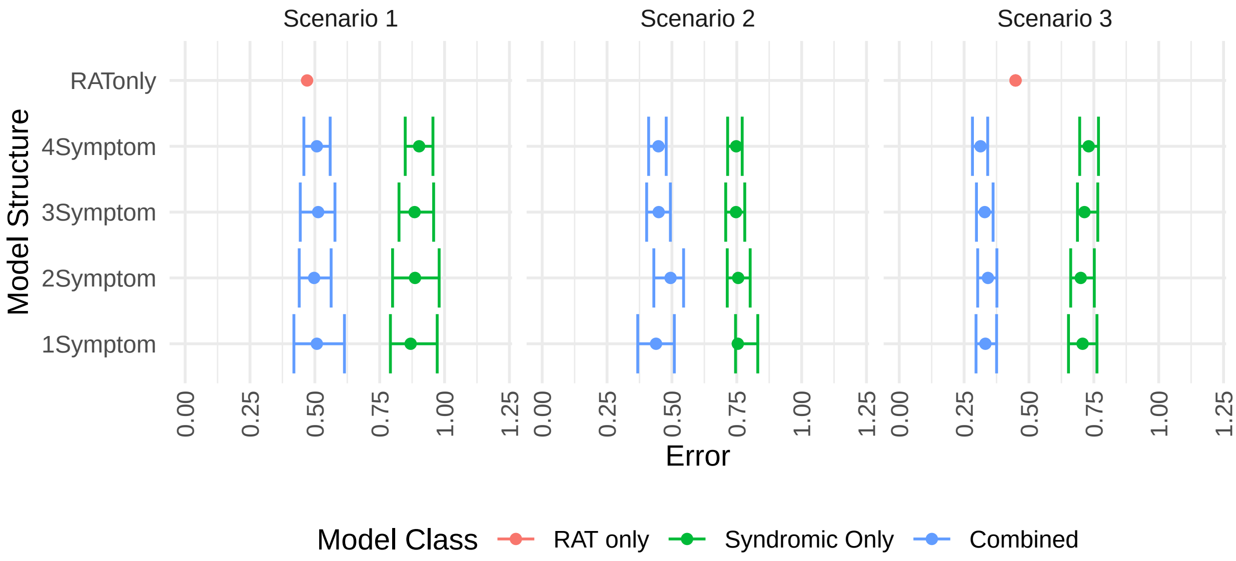


Figure 5: Performance of models under each scenario measured by posterior median and interquartile range for errors defined in Table 1. Low errors correspond to better model performance. There is no error rate defined for the Model Class 1 (RAT only model) in Scenario 2 as the model failed to meet the requirement for that scenario (making the error functionally infinite).

These figures demonstrate both the added flexibility of the more complex model classes that allow them to be tailored to specific needs and the need to combine the high-quality but inflexible RAT results with the more flexible but lower quality syndromic data.

The final symptoms chosen through model selection should be interpreted cautiously. These models were developed for prediction and classification in a unique sub-population: CST-identified, symptomatic patients. The symptoms and risk factors retained in the model classes differed, despite these data being collected over a short time period from the same population. These differences may point to mechanisms by which CST-identified and RAT-positive patients differ from other groups. Of particular interest is whether individuals that are missed by RAT are less infectious, which could be explored by using viral load measured as Threshold Cycle (Ct) values from the RT-PCR [@albert2021field].

Our methodology has been developed using a large sample size drawn under field-realistic conditions and has thus developed with the practicalities of mass deployment in mind. Improving case identification using statistical methods allows us to update our diagnostic process in real-time, allowing rapid adaptation to new variants or even new diseases. The modelling frameworks we have used are also sufficiently flexible to accommodate new data sources (such as background case numbers) or changes in the local relative costs of false positives and false negatives.

Naturally, these strengths have complementary limitations. Our models require updating in real-time and can only achieve good performance if the validation data are of good quality. Similarly, targeting misdiagnosis rates is only sensible if those rates properly reflect local conditions which can be challenging. While these limitations should be seriously considered, we believe that the alternatives simply hide these problems. We choose to make these decisions explicitly to allow them to be more readily challenged, researched and improved upon. These challenges represent promising new avenues for impactful research that improve our understanding of estimating misdiagnosis rate trade offs and how to translate sample population findings to target populations.

We believe that our combined syndromic and rapid antigen testing approach is the most promising method for large-scale testing in LMICs for COVID-19 at present. We have demonstrated that these improvements can be impressive in real-world scenarios, and will have a large impact when scaled to the population sizes in LMICs. The framework we outline above is readily adapted for other diagnostic problems. Malaria, schistosomiasis, rabies and many other diseases are all currently monitored either sparsely with gold-standard methods (such as RT-PCR, autopsies, fluorescent antibody testing) or at a large scale with more error-prone methods (RATs, blood smears, egg counts, differential diagnosis).

The management of global pandemics can only be done with global testing. While the quest to achieve this using only gold-standard diagnostic methods is laudable, it is also often impractical. Imperfect diagnostics are frequently imperfect in different ways, and these differences are ripe for statistical treatment. What is more, these approaches are often more agile than gold-standard diagnostics in situations of flux, for example, in the early stages of new pandemics or disease strains, when fast responses are essential.
By investing in understanding how to utilise the complementary strengths of imperfect testing and deploy the limited gold-standard testing available for validation, we can provide good quality testing at the scale needed to fight infectious diseases.

Funding (26 words)

The Bill and Melinda Gates Foundation funded work by FAO (INV-022851), and University of Glasgow reports funding from Wellcome (207569/Z/17/Z). The authors declare no competing interests.

Acknowledgements (69 words)

We would like to thank members of the community support teams in Bangladesh who have provided essential services throughout the pandemic. Earlier drafts of this mansucript benefited from the input of Paul Johnson, Daniel Haydon, Anne-Sophie Bonnet-Lebrun, Luca Nelli, Crinan Jarrett, Rita Claudia Cardoso Ribeiro, Halfan Ngowo, Heather McDevitt and Gina Bertolacci. The University of Glasgow COVID-19 in LMICs Group provided the environment in which to develop this work.

References (Max 30)

Below we have extended the modelling description provided in the main text to include more technical detail. The code used to implement these tasks is available at https://github.com/fergusjchadwick/COVID19_SyndromicRATDiagnosis.

Modelling

Structure

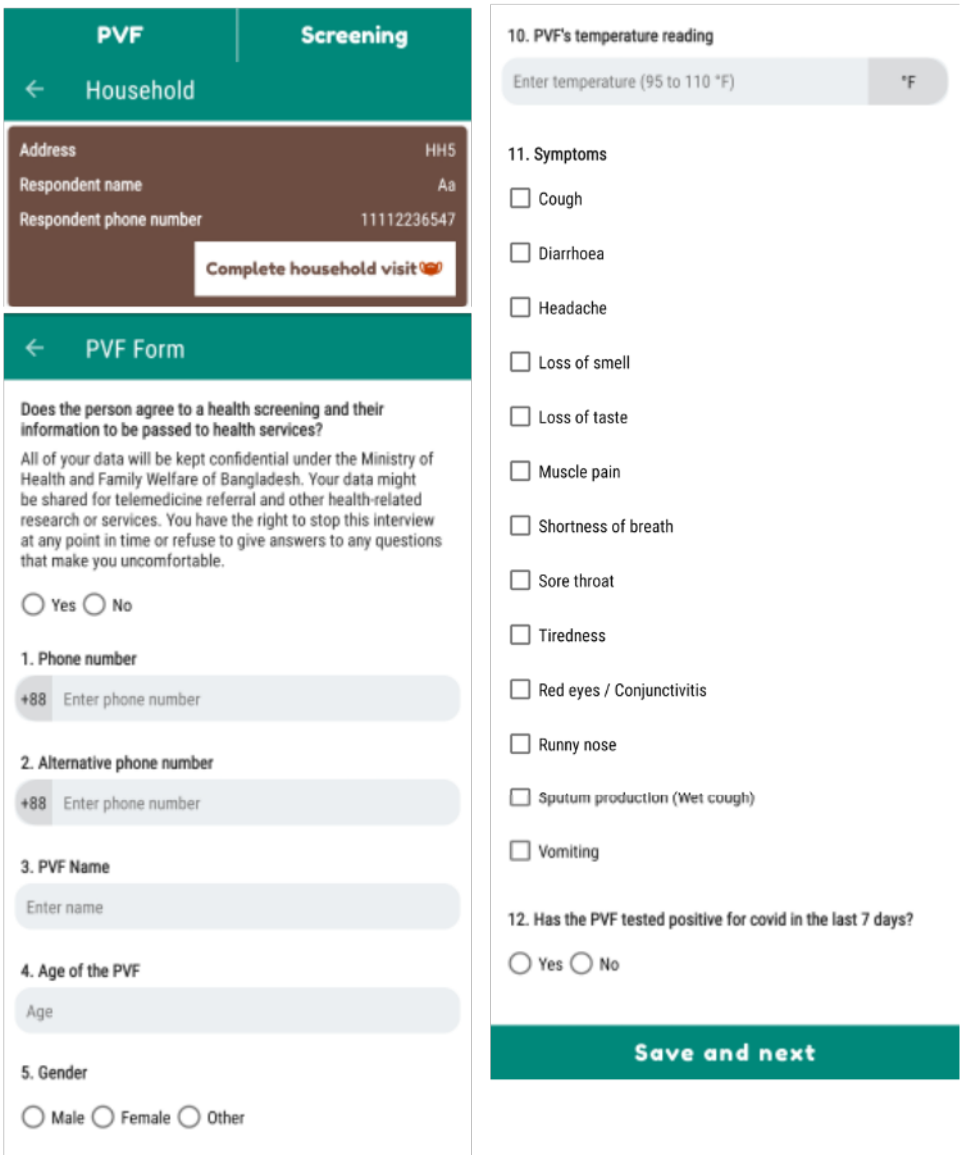
We examined the ability of the two imperfect identification methods, syndromic modelling and RAT, to predict the patient's COVID-19 status when used separately and together. These combinations define three model classes (Figure @ref(fig:data-flowchart2)).

Find more ways to cite Figure ref(fig:data-flowchart2) in this text.

Model Class 1 uses only the RAT result. It equates being RAT-positive with the patient being PCR-positive for COVID-19 (hereafter, PCR-positive), and being RAT-negative with PCR-negativity. We can write this formally for the ith individual as:

$$\begin{gathered}\mathrm{PCR}_i = \mathrm{RAT}_i \\ \mathrm{PCR} \in \{0,1\} \\ \mathrm{RAT} \in \{0,1\}\end{gathered} \tag{1}$$

Model Class 2 uses only the syndromic data. For this model, we used a Bayesian multivariate probit model [@albert1993bayesian]. The multivariate probit structures the outcomes of the PCR test and symptoms presence/absence as a D-dimensional vector of binary outcomes ($\boldsymbol{y}_i = (y_{i1}, y_{i2}, \ldots, y_{id})$, $y_{ij} \in \{0,1\}$). These outcomes are determined by an indicator function which takes a D-dimensional vector of *continuous latent* variables ($\boldsymbol{z}_i = (z_{i1}, z_{i2}, \ldots, z_{id})$, $z_{ij} \in \mathbb{R}$). These latent continuous variables then covary as realisations of



PVF | Screening

← Household

Address HH5

Respondent name Aa

Respondent phone number 11112236547

Complete household visit

← PVF Form

Does the person agree to a health screening and their information to be passed to health services?

All of your data will be kept confidential under the Ministry of Health and Family Welfare of Bangladesh. Your data might be shared for telemedicine referral and other health-related research or services. You have the right to stop this interview at any point in time or refuse to give answers to any questions that make you uncomfortable.

○ Yes ○ No

1. Phone number

+88 Enter phone number

2. Alternative phone number

+88 Enter phone number

3. PVF Name

Enter name

4. Age of the PVF

Age

5. Gender

○ Male ○ Female ○ Other

10. PVF's temperature reading

Enter temperature (95 to 110 °F) °F

11. Symptoms

□ Cough

□ Diarrhoea

□ Headache

□ Loss of smell

□ Loss of taste

□ Muscle pain

□ Shortness of breath

□ Sore throat

□ Tiredness

□ Red eyes / Conjunctivitis

□ Runny nose

□ Sputum production (Wet cough)

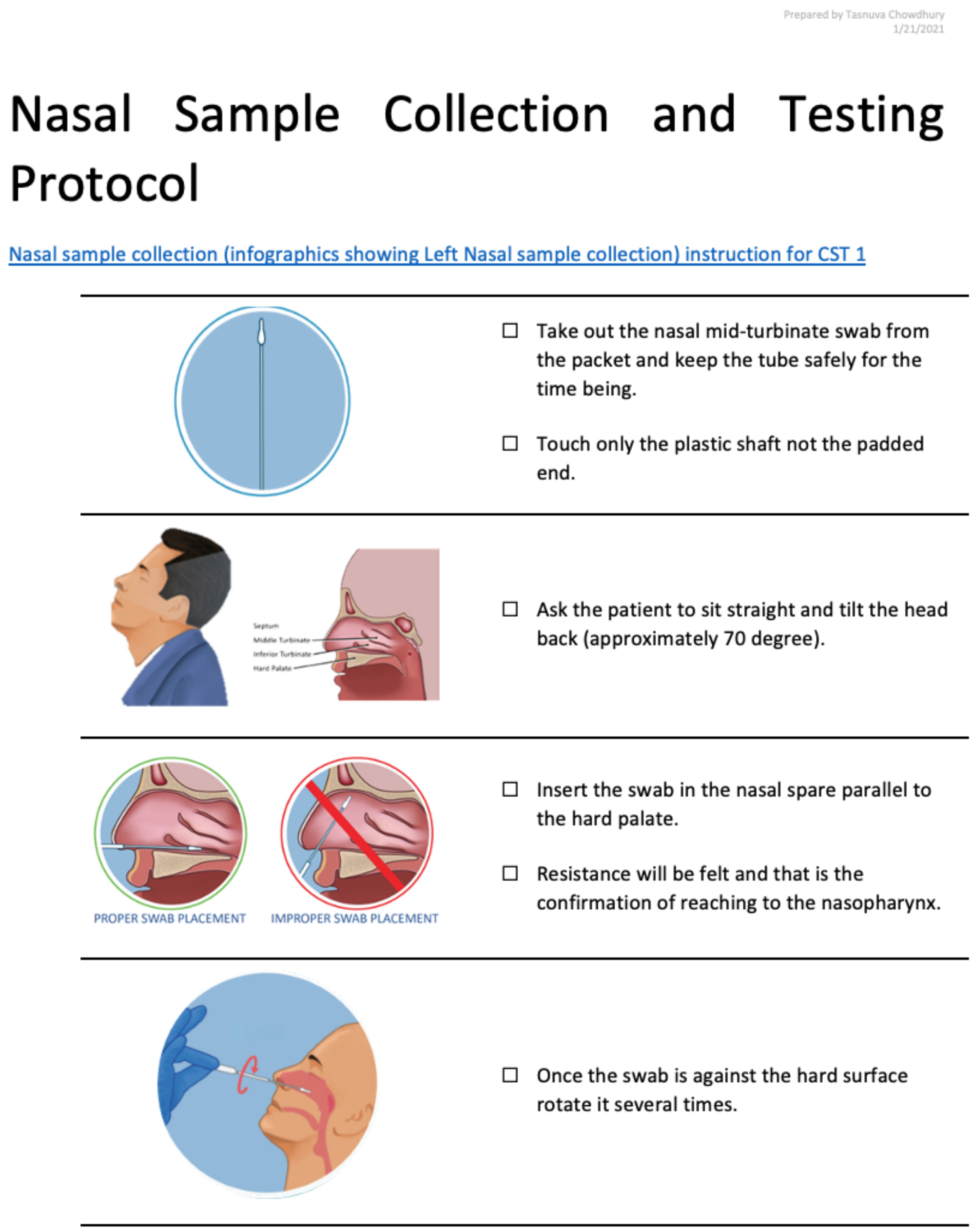
□ Vomiting

12. Has the PVF tested positive for covid in the last 7 days?

○ Yes ○ No

Save and next

Figure 6: Data collection by the community support teams (CSTs) is implemented using a mobile phone application, screenshots presented here.



Prepared by Tasnuva Chowdhury
1/21/2021

Nasal Sample Collection and Testing Protocol

Nasal sample collection (infographics showing Left Nasal sample collection) instruction for CST 1

□ Take out the nasal mid-turbinate swab from the packet and keep the tube safely for the time being.

□ Touch only the plastic shaft not the padded end.

Septum
Middle Turbinate
Inferior Turbinate
Hard Palate

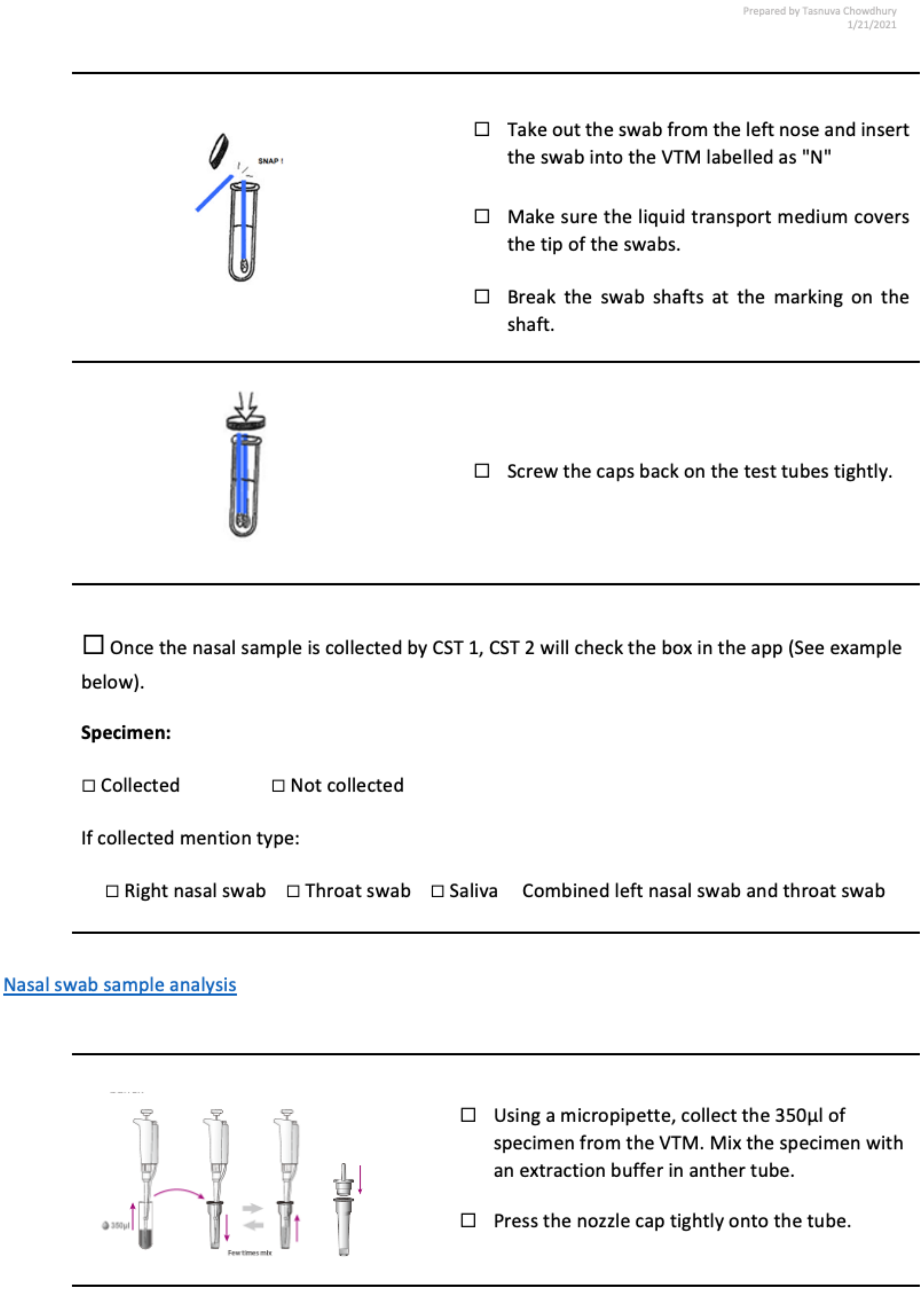
□ Ask the patient to sit straight and tilt the head back (approximately 70 degree).

□ Insert the swab in the nasal spare parallel to the hard palate.

□ Resistance will be felt and that is the confirmation of reaching to the nasopharynx.

□ Once the swab is against the hard surface rotate it several times.

Figure 7: Page 1 sample collection protocols developed by Tasnuva Chowdhury.


SNAP !
350µl
Few times mix

Prepared by Tasnuva Chowdhury
1/21/2021

- ☐ Take out the swab from the left nose and insert the swab into the VTM labelled as "N"
- ☐ Make sure the liquid transport medium covers the tip of the swabs.
- ☐ Break the swab shafts at the marking on the shaft.

- ☐ Screw the caps back on the test tubes tightly.

☐ Once the nasal sample is collected by CST 1, CST 2 will check the box in the app (See example below).

Specimen:

☐ Collected ☐ Not collected

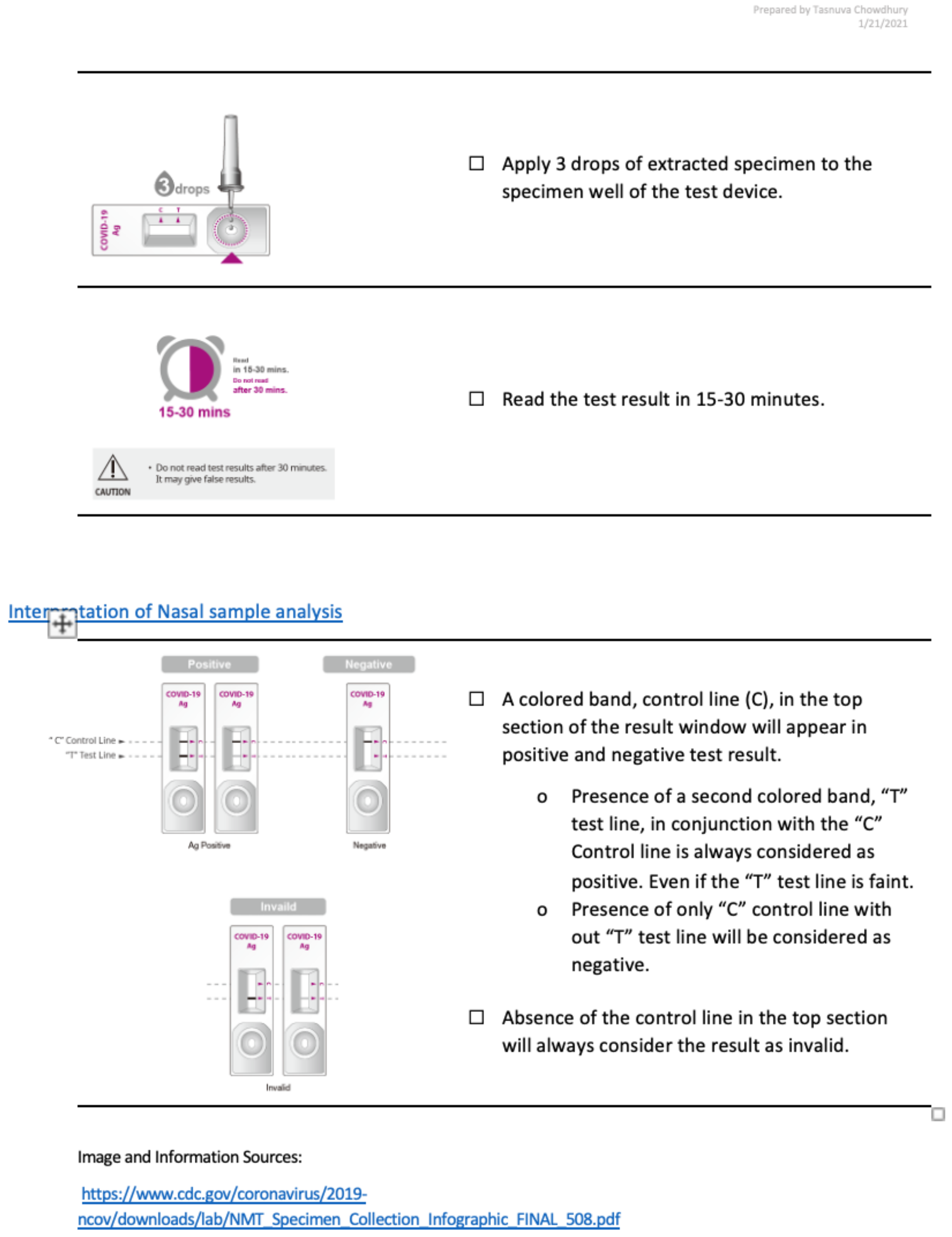
If collected mention type:

☐ Right nasal swab ☐ Throat swab ☐ Saliva Combined left nasal swab and throat swab

Nasal swab sample analysis

- ☐ Using a micropipette, collect the 350µl of specimen from the VTM. Mix the specimen with an extraction buffer in anther tube.
- ☐ Press the nozzle cap tightly onto the tube.

Figure 8: Page 2 sample collection protocols developed by Tasnuva Chowdhury.



Prepared by Tasnuva Chowdhury
1/21/2021

□ Apply 3 drops of extracted specimen to the specimen well of the test device.

□ Read the test result in 15-30 minutes.

CAUTION

- Do not read test results after 30 minutes. It may give false results.

Interpretation of Nasal sample analysis

□ A colored band, control line (C), in the top section of the result window will appear in positive and negative test result.

  - o Presence of a second colored band, "T" test line, in conjunction with the "C" Control line is always considered as positive. Even if the "T" test line is faint.
  - o Presence of only "C" control line with out "T" test line will be considered as negative.

□ Absence of the control line in the top section will always consider the result as invalid.

Image and Information Sources:

https://www.cdc.gov/coronavirus/2019-ncov/downloads/lab/NMT_Specimen_Collection_Infographic_FINAL_508.pdf

Figure 9: Page 3 sample collection protocols developed by Tasnuva Chowdhury.

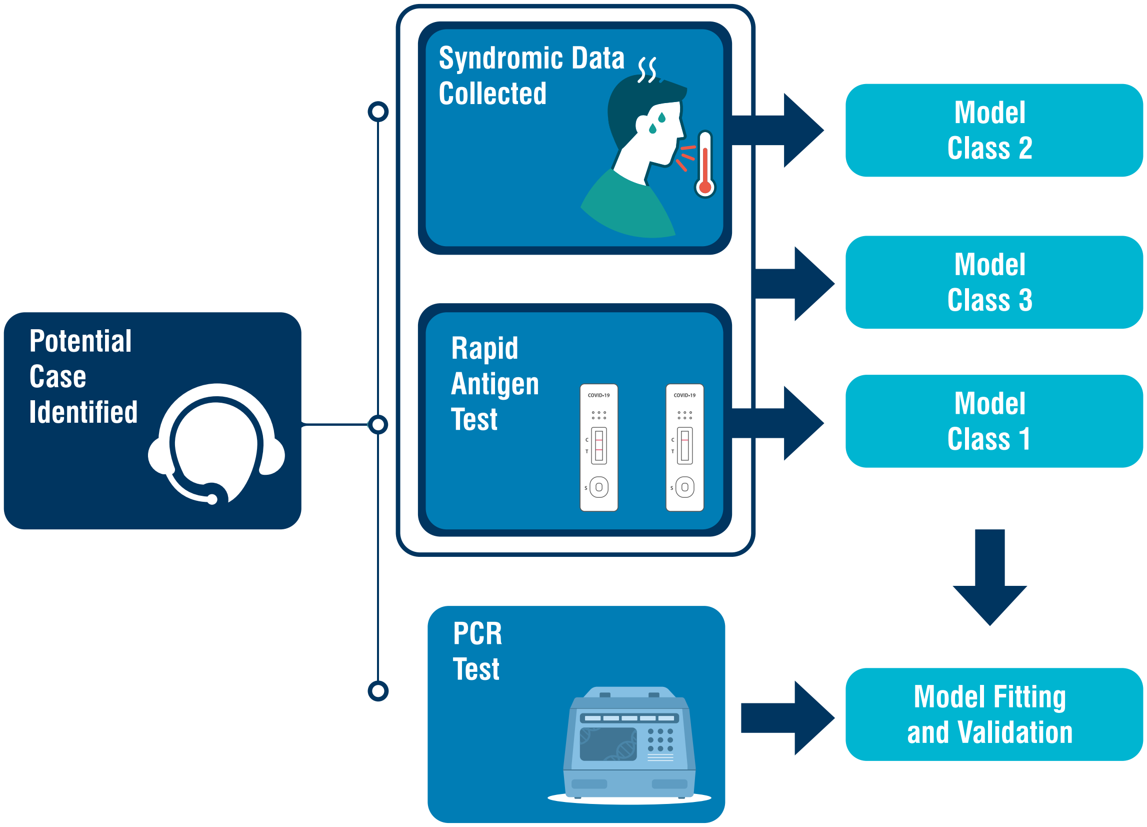


Figure 10: Schematic description of identification of likely COVID-19 patients by community support teams (CSTs), swab collection and model definitions. The teams collected syndromic data (age, gender and presence/absence of 14 predetermined symptoms), and two sets of naso-pharyngeal swabs (one each for Rapid Antigen Testing and RT-PCR). We then used rapid antigen testing (RAT) and syndromic data, two imperfect but inexpensive diagnostics, to generate three model classes: RAT result only in Model Class 1, syndromic data only in Model Class 2, and both RAT result and syndromic data in Model Class 3. The RT-PCR test result is used to train and test each model using temporal cross-validation.

a D-dimensional multivariate normal, with the mean of the error structure informed by a linear predictor, $\sum_{j=1}^{J} x_{ij}\beta_{jd} + \epsilon_{id}$, and a covariance between dimensions, Σ. The linear predictor allows us to condition the outcomes on risk factor variables (here, age and gender) while the covariance structure allows us to account for the correlated nature of the symptoms with each other and the outcome. As Albert and Chib [@albert1993bayesian] identify in their paper, the covariance matrix formulation is not identifiable, with the variance, $diag(\Sigma)$ and means of the latent variables, $\boldsymbol{z}_i$ trading off against each other. For this reason, we use a correlation matrix, Ω, formulation with the variance set to 1. A correlation based framework also makes communication with clinicians and other practitioners smoother as correlations are more familiar.

$$\begin{aligned}
y_{id} &= \mathbb{I}(z_{id} > 0) \\
\boldsymbol{z}_i &= \boldsymbol{x}_i\boldsymbol{\beta} + \boldsymbol{\epsilon}_i \\
z_{id} &= \sum_{j=1}^{J} x_{ij}\beta_{jd} + \epsilon_{id} \\
\boldsymbol{\epsilon}_i &\sim N(\boldsymbol{0}, \boldsymbol{\Omega}) \\
diag(\Omega) &= 1 \\
\beta &\sim N(0,1) \\
\boldsymbol{\Omega} &\sim \text{LKJ}(1)(\#eq:ModelClass2)
\end{aligned} \tag{2}$$

Model Class 3 combines the two data sources. We utilise the specificity of RAT by treating RAT-positive patients as PCR-positive patients. The RAT-negative patients are modelled using the sensitive syndromic approach using Model Class 2 to capture PCR-positive patients that are missed by the RAT. This approach leverages the potential different syndromic profiles of PCR-positve patients who are RAT-positive and -negative, allowing the model to adapt solely to the latter. Structurally, the model combines Equations @ref(eq:ModelClass1) and @ref(eq:ModelClass2), with RAT-positive patients being modelled using Equation @ref(eq:ModelClass1), and Equation @ref(eq:ModelClass2).

$$\begin{aligned}
y_{i1} &= \begin{cases} 1, & \text{if } \text{RAT}_i = 1 \\ & \text{otherwise} \end{cases} \\
y_{id} &= \mathbb{I}(z_{id} > 0) \\
\boldsymbol{z}_i &= \boldsymbol{x}_i\boldsymbol{\beta} + \boldsymbol{\epsilon}_i \\
z_{id} &= \sum_{j=1}^{J} x_{ij}\beta_{jd} + \epsilon_{id} \\
\boldsymbol{\epsilon}_i &\sim N(\boldsymbol{0}, \boldsymbol{\Omega}) \\
diag(\Omega) &= 1\beta & \sim N(0,1) \\
\boldsymbol{\Omega} &\sim \text{LKJ}(1)(\#eq:ModelClass3)
\end{aligned} \tag{3}$$

This isn't quite right - need to demonstrate subsetting of i's, otherwise it implies that z is calculated for all data points

By using a Bayesian formulation, we generate full posteriors for our parameter estimates, allowing natural quantification of uncertainty. Bayesian methods also facilitate the use of more informative priors. While we used minimally informative priors here (standard normals in the probit scale for betas and an LKJ correlation prior with minimal shrinkage, $\eta = 1$ [@lewandowski2009generating]), more informative priors that incorporate spatio-temporal effects, for instance, would be natural extensions. The models were fitted to the data using Bayesian inference techniques based on Hamiltonian Monte Carlo in the Stan programming language [@carpenter2017stan]. The models all converged with zero divergent transitions and large effective sample sizes.

Model Selection

We conducted backwards model selection (starting with the most complex, biologically plausible model) to identify a subset of models with the highest predictive power under temporal cross-validation (Figure

@ref(fig:modsel-flowchart2)). For the cross-validation, we divided the data into 5 folds of equal sizes in time order (i.e. the first fold is formed of the chronologically first $\frac{N}{K}$ patients, where N is the number of patients and K is the number of folds, the second fold by the next $\frac{N}{K}$ etc.) To test the sensitivity of this cross-validation structure, we also did a strict temporal division (i.e. the first $\frac{T}{K}$ days where T is the number of days samples were taken on). The results did not change qualitatively between these approaches.

The coarse round of model selection (Figure @ref(fig:modsel-flowchart2)) selected candidate symptoms based on whether they had a strong and consistent correlation with PCR as estimated according to Equations @ref(eq:ModelClass2) and @ref(eq:ModelClass3). The models were fit with both covariates throughout the coarse round and symptoms were compared in nested models. In the fine round of model selection, these candidate symptoms and the covariate combinations (age and gender, age, gender and no covariates) were permuted to more exhaustively explore the model space. Reducing the number of possible models using the two stages of model selection was necessary to reduce computational demand and reduce the risk of overfitting models to the test scenarios. The large number of symptoms corresponds to a high number of potential model configurations (>131 000 for 14 symptoms and two covariates) which might perform well on the test sets (even under the challenging conditions of temporal cross-validation) but lack transferability.

By using general predictive power to narrow down the number of candidate models and then testing those models, we are more likely to choose models which generalise well to new data. The number of candidate models used was not pre-determined but it was clear when fitting the models that there were "jumps" in performance (as defined below) between models containing five and four symptoms, so the models with one to four symptoms were used as the candidate models. Zero symptom models were not included in the analysis as they do not correspond to a feasible policy (with covariates they would require governments to ask individuals of a given gender and age as COVID-19 positive, and without covariates they would involve randomly assigning individuals as COVID-19 positive).

Model selection carried out in this way is quite robust. However, if more time were available for model development we would advocate more sophisticated methods of variable selection, such as penalised complexity priors [@simpson2017penalising] or auxiliary variable selection [@held2006bayesian]. These methods, once developed for the use-case, are superior but require significant testing that would prevent getting this work out in a policy-relevant timeframe. We also believe that these approaches may be harder for non-statisticians to adapt.

Predictive Performance

We scored the models' predictive power using binary cross-entropy (hereafter, cross-entropy). Cross-entropy measures the accuracy of models that generate probabilities of binary outcomes, rather than make binary classifications, similar in concept to a mean square error for normally-distributed data, but adapted for binary data [@gneiting2007strictly]. A cross-entropy value close to zero corresponds to high levels of accuracy, with larger values indicating lower accuracy. More specifically, the metric allows us to compare a binary vector, $\boldsymbol{y} \in [0, 1]$, with a vector of probablistic predictions ($p(\boldsymbol{y}) \in (0, 1)$) as follows:

$$\boldsymbol{H}_p(q) = \frac{1}{N}\sum_{i=1}^{N} y_i \cdot log(p(y_i)) + (1 - y_i) \cdot log(1 - p(y_i))(\#eq : CrossEntropy) \tag{4}$$

The resulting score is comparable across all methods for assigning predictions where the same test data are used, allowing us to compare predictions from Model Classes 1-3. $H_p(q) \in 0, \boldsymbol{R}_+$ with zero indicating perfect prediction (assigning probabilities of ones and zeroes to outcomes of ones and zeros exactly) and larger values indicating worse predictions.

Classification Performance

In applied settings, models must often be evaluated on their performance as classifiers rather than just as prediction engines (i.e. their ability to say a patient is COVID-19 positive or negative, not simply the

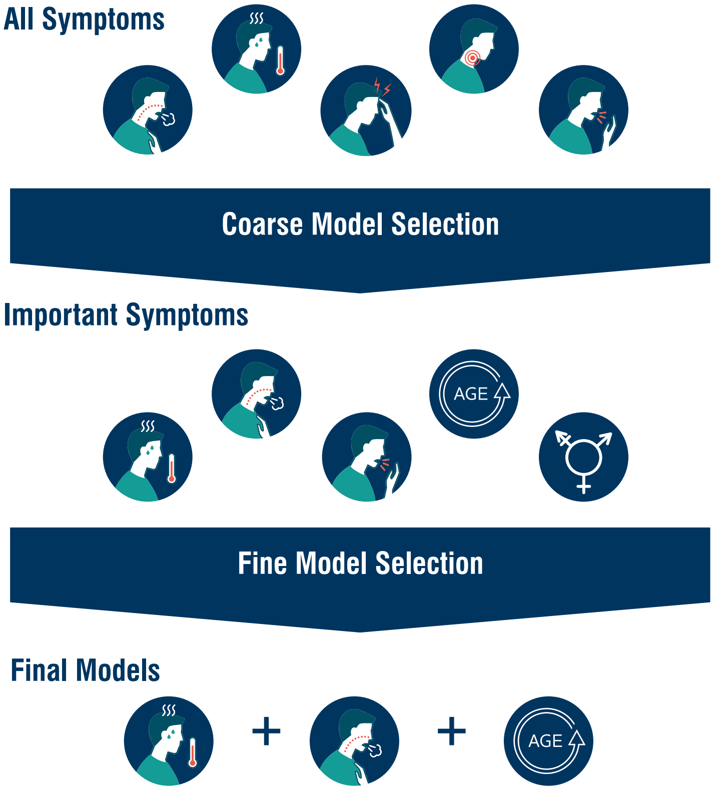


Figure 11: Schematic for rounds of model selection in the multivariate probit component of Model Classes 2 and 3. With 14 symptoms (only 5 shown here for demonstration purposes) and two covariates there are over 131000 possible model combinations. To make exploring these possible models computationally feasible and to reduce the risk of overfitting, we carried out two rounds of model selection. First, the data are divided into temporal cross-validation sets. The multivariate probit connects symptoms to the RT-PCR result through a correlation matrix. In the coarse model selection, the most complex feasible model (all symptoms and covariates) is fit to the training data. The estimated correlations between each symptom and the RT-PCR result are compared for each cross-validation set. The symptoms that have non-zero correlations in a systematic direction (i.e. all positively or all negatively correlated with RT-PCR result) are retained. The process is then repeated on each retained set of symptoms until the four symptoms in each model class with the strongest correlation to RT-PCR result. We then conduct a more exhaustive model selection on all the possible permutations of the four symptoms and two covariates. In this round, each model is fit to training data and used to predict for the test set, and the quality of those predictions is measured using cross-entropy scoring. The cross-entropy score is then used to select the best predictive model for each level of model complexity. Only these final models are then used for classification. This reduces the set of models tested as classifiers from >131 000 to just four per model class.

probability the patient might be COVID-19 positive or negative). To generate a classification, $\hat{Y}$, a probability threshold, $\hat{p}$, must be chosen over which patients are classified as COVID-19 positive:

$$\hat{Y} = \begin{cases} 1, & \text{if } p(y) \geq \hat{p} \\ 0 & \text{otherwise} \end{cases} (\#eq: ClassificationThreshold) \tag{5}$$

Receiver operating characteristics (ROCs) are a way to measure the performance of a set of classifications in terms of true and false positives and negatives (TP, FP, TN and FN) and the rates of each of these classification types (e.g. $TPR = \frac{TP}{TP+FN}$, and $FPR = \frac{FP}{FP+TN}$). The error rates are calculated with respect to a particular threshold, $\hat{p}$, or across the range of possible $\hat{p}$s to generate a ROC curve [@hoo2017roc]. In our epidemiological scenarios (outlined below) we use our ROC curve calculations to identify single thresholds which yield a required error rate.

We strongly emphasise that generic performance here is only used to show the flexibility of the model classes; the best model for a local situation can only be determined if the relative cost of false positives and false negatives is known. Here, we choose three representative scenarios. Each scenario has a requirement and error rate (defined in Table @ref(tab:scenarios-tab2)).

In Scenario 1, we do not consider epidemiological context but simply minimise false negative and false positive rates equally. We do this by maximising the two correct classification rates both individually and in total, as measured by the harmonic mean (as opposed to the arithmetic mean which would only maximise the rates in total).

$$\text{Find } \hat{p} \text{ which } \max\left(\frac{2 \cdot \text{TPR} \cdot \text{TNR}}{\text{TPR} + \text{TNR}}\right), \epsilon = \text{FPR} + \text{FNR}$$

I'm not sure if any of these equations are correct or sufficient or clear

Scenario 2 corresponds to the current situation in Bangladesh at time of writing (July 2021), with COVID-19 cases beginning to rapidly increase again. Under these circumstances, false negatives are extremely costly relative to false positives due to the exponential growth of the disease.

$$\text{Find } \hat{p} \text{ which } \max\left(FNR - 0.2 \leq 0\right), \epsilon = \text{FPR}$$

In Scenario 3, the pandemic is not declining but maintaining a steady rate of cases. In this situation, policy-makers may be keen to keep false positive diagnoses low to prevent lockdown fatigue and to keep the workforce active.

$$\text{Find } \hat{p} \text{ which } \max\left(FPR - 0.2 \leq 0\right), \epsilon = \text{FNR}$$

The requirements in Scenario 2 and 3 were developed in discussion with the Institute of Epidemiology, Disease Control and Research (IEDR), Bangladesh, for illustrative purposes.

Table 2: For each epidemiological scenario there is a requirement and a performance criterion. The requirement refers to a base level of performance the model must achieve; in general this will be a maximum acceptable error rate of some kind. These requirements were determined in discussion with members of the Institute of Epidemiology, Disease Control and Research, Ministry of Health, Bangladesh (IEDCR). The requirement determines a probability threshold for each model which most closely meets that requirement. The performance criterion is then used to determine which model performs the 'best' given that the requirement has been met.

Scenario Name	Requirement	Performance Criterion (Error)
1 Agnostic	Maximise correct classification rates	Sum of error rates
2 Rising Cases	Max. 20% false negative rate	False negative rate
3 Low-Level Cases	Max. 20% false positive rate	False positive rate