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

Fergus J Chadwick[a,c], Yacob Haddou[a,c], Tasnuva Chowdhury[a], David Pascall[d], Shayan Chowdhury[e], Jessica Clark[a,c], Joanna Andrecka[f], Mikolaj Kundergorski[b,c], Craig Wilkie[b,c], Eric Brum[f], Tahmina Shirin[g], A S M Alamgir[g], Mahbubur Rahman[g], Ahmed Nawsher Alam[g], Farzana Khan[g], Janine Illian[b,c], Ben Swallow[b,c], Davina L Hill[a,c], Dirk Husmeier[b], Jason Matthiopoulos[a,c], Katie Hampson[a,c], Ayesha Sania[h]

[a] *Institute of Biodiversity, Animal Health and Comparative Medicine, University of Glasgow*
[b] *School of Mathematics and Statistics, University of Glasgow*
[c] *COVID-19 in LMICs Research Group, University of Glasgow*
[d] *MRC Biostatistics Unit, University of Cambridge*
[e] *a2i Programme, ICT Ministry/UNDP Bangladesh*
[f] *UN FAO in support of the UN Interagency Support Team, Bangladesh*
[g] *Institute of Epidemiology Disease Control and Research, Ministry of Health, Bangladesh*
[h] *Division of Developmental Neuroscience, Department of Psychiatry, Columbia University*

1. Abstract

Background

The majority of the world's population live in low- and middle-income countries 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.

*Corresponding Author

Email addresses: f.chadwick.1@research.gla.ac.uk (Fergus J Chadwick), yacob.haddou@glasgow.ac.uk (Yacob Haddou), tasnuvachowdhury2004@gmail.com (Tasnuva Chowdhury), david.pascall@mrc-bsu.cam.ac.uk (David Pascall), shayan.chowdhury@a2i.gov.bd (Shayan Chowdhury), Jessica.Clark@glasgow.ac.uk (Jessica Clark), aandrecka@gmail.com (Joanna Andrecka), mikolaj.kundegorski@gmail.com (Mikolaj Kundergorski), craig.wilkie@glasgow.ac.uk (Craig Wilkie), eric.brum@fao.org (Eric Brum), tahmina.shirin14@gmail.com (Tahmina Shirin), aalamgir@gmail.com (A S M Alamgir), dr_mahbub@yahoo.com (Mahbubur Rahman), anawsher@yahoo.com (Ahmed Nawsher Alam), farzanakhan_25@yahoo.com (Farzana Khan), janine.illian@glasgow.ac.uk (Janine Illian), ben.swallow@glasgow.ac.uk (Ben Swallow), davina.hill@glasgow.ac.uk (Davina L Hill), dirk.husmeier@glasgow.ac.uk (Dirk Husmeier), jason.matthiopoulos@glasgow.ac.uk (Jason Matthiopoulos), katie.hampson@glasgow.ac.uk (Katie Hampson), ays328@mail.harvard.edu (Ayesha Sania)

A sample (n = 511) 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, sensitivity and specificity.

Findings

Combined data models yielded improved performance over syndromic- and RAT-only models across all three metrics, with sensitivity of W (CI . . .) {relative to {X and Y}, respectively}, specificity of A (CI . . .) {relative to {B and C}, respectively} and log-loss of D (CI . . .) {relative to {E and F}, respectively}.

Interpretation

We demonstrate that integrating these imperfect data sources greatly improves the detection of COVID-19. Low-cost and accessible surveillance methods make pandemic control in low- and middle- income countries a possibility.

Funding

The Bill and Melinda Gates Foundation and the Wellcome Trust.

2. Introduction

Identification and isolation of COVID-19 cases remains the key component of the pandemic response across the globe. From informing individual care to targeting population-level interventions, the faster and more accurately we can identify who is infected, the more effectively we can provide clinical care and reduce transmission of infection. RT-PCR has rapidly become the default, gold-standard test for COVID-19 in applied settings, with high sensitivity and specificity for COVID-19 and relatively ease of sample collection. Most of the world's population, however, live in low- and middle-income countries where the laboratory facilities needed to carry out RT-PCR tests are often scarce and hard to reach for many people. Case identification worldwide, therefore, must be made accessible using inexpensive methods that can be carried out locally.

An increasingly popular alternative to RT-PCR are rapid antigen tests (RATs). Like RT-PCR, these tests have high specificity for COVID-19 and are widely used as they are less expensive, easier to use, and faster. RATs also require less commitment and discomfort for patients. For RT-PCR testing, patients must travel to a designated site (such as a hospital or testing booth) and invasive nasopharyngeal swabs must be taken. RATs can be conducted on saliva samples and completed in the home. RATs can also be done 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, several concerns have been raised about the sensitivity of RAT.

One of the alternatives to RT-PCR that has been used since the start of the pandemic is symptom-thresholding. In this approach, a patient presenting with a fever and one or more viral pneumonia symptoms is treated as a COVID-19 positive patient. The main advantage of this approach is the ease of implementation. For example, in Bangladesh, a lower middle income country, initial support and reporting of infections locally is provided by community support

teams (CSTs) composed of volunteers from the community with basic training. The CSTs can easily collect symptomatic data in the community and provide care where the criteria are met. Unfortunately, these symptom-thresholds were developed early in the outbreak, necessarily drawn from clinical intuition, rather than data. Consequently, the relationship between the criteria and the target diagnosis is often weak, with low specificity.

A natural extension to these symptom-threshold approaches is syndromic modelling. Here, rather than using a set of criteria, a range of symptomatic and risk factor data are collected and then a sub-sample of patients are tested using RT-PCR for COVID-19. These data are used to generate a predictive model that allows more accurate estimation of how likely a given patient is to have COVID-19. In such resource-limited settings, there is very limited provision for testing of asymptomatic cases, despite their important role in disease transmission. Syndromic modelling is a complex, nuanced task. The strength of relationships between symptoms and diseases is not stable through time or across sampling strategies since the relative importance of each symptom for disease diagnosis, in part, depends on the prevalence of other diseases causing similar symptoms in the community.

For example, if another disease for which loss of smell is a symptom becomes common, that symptom becomes a worse predictor for COVID-19. Similarly, if everyone who has a cough is considered a likely COVID-19 patient and thus is included in the training data, the cough will likely have a very low correlation with COVID-19 (even if the two are strongly related in the general population). While these issues can be overcome by properly considering the population sampled and using appropriately sophisticated statistical methods, the many types of common respiratory disease generally means that these models tend to have relatively high false positive rates for COVID-19 (although much lower than the symptom-threshold approach).

Neither syndromic modelling nor RATs have the high sensitivity and specificity of RT-PCR, but higher error rates may be tolerable depending on their scale and impact. Low specificity will mean a large number of false positive classifications, where the patient is told they have COVID-19 but they actually do not. This might lead to patients unnecessarily self-isolating and receiving support which can be expensive to the individuals and local public health bodies, as well as reducing available resources for those who need them. Similarly, low sensitivity will result in more false negative classifications, where the patient is told they do not have COVID-19 but they actually do, which can lead to a health-risk for the individual and to the disease spreading further. The costs of these misclassifications will depend on local context. When the prevalence of the disease is low, false positives may create local skepticism about the value of testing, or when there are strong population-level mitigations already in place (such as a nationwide lockdown), then false positives might be more costly than false negatives. If the disease is abundant or increasing rapidly then false negatives are likely to be more costly. In most situations, a balance will need to be struck.

The two dominant testing methods available in resource limited settings,

therefore, are both 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 how to exploit complementarity between these data types to ameliorate their respective weaknesses. 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; and to demonstrate that the combined data achieve equal to much lower error rates than the next best method. 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.

3. Methods

Participants included in this study were identified for COVID-19 testing after self-reporting symptoms to the Bangladesh government's national hotlines for COVID-19 support. Recruitment took place across Dhaka (the capital city of Bangladesh) between 2nd April 2021 and 5th May 2021.

Patients were selected for further testing conditional on the presence of a fever (>38°C) and one or more of 13 additional symptoms associated with COVID-19 (breathing problems, coughing, diarrhoea, a headache, loss of taste, loss of smell, muscle pain, red eyes, a runny nose, a sore throat, tiredness, vomiting or a wet cough). The patient's age and gender were also recorded, but these data were not included in the patient selection criteria.

Nasopharyngeal swabs and syndromic data were collected from the patient by medical technologists. One swab each was used for Rapid Antigen Testing (RAT) and RT-PCR (gold-standard for COVID-19 status). The syndromic profile comprises the patient's symptomatic information, age and gender). The full questionnaire and testing protocols are provided in Appendix XX. Participants provided written informed consent to sample collection and for their test results to be analyzed in the study.

We examined the ability of the two imperfect identification methods, the syndromic profile and RAT result, to predict the patient's COVID-19 status when used separately and together. The different data combinations define three model classes.

Model Class 1 uses only the RAT result and is the simplest of the three. It simply equates a positive RAT result with the patient being PCR positive, and a negative RAT result with PCR negativity. Model Class 2 uses only the syndromic data and Model Class 3 combines the RAT result with the syndromic data.

For Model Class 2, we used a Bayesian multivariate probit model. The multivariate probit structure allows the model to account for the correlations between, and binary nature of, the symptoms (e.g. loss of taste is often correlated with loss of smell). By using a Bayesian formulation, we are able to better quantify

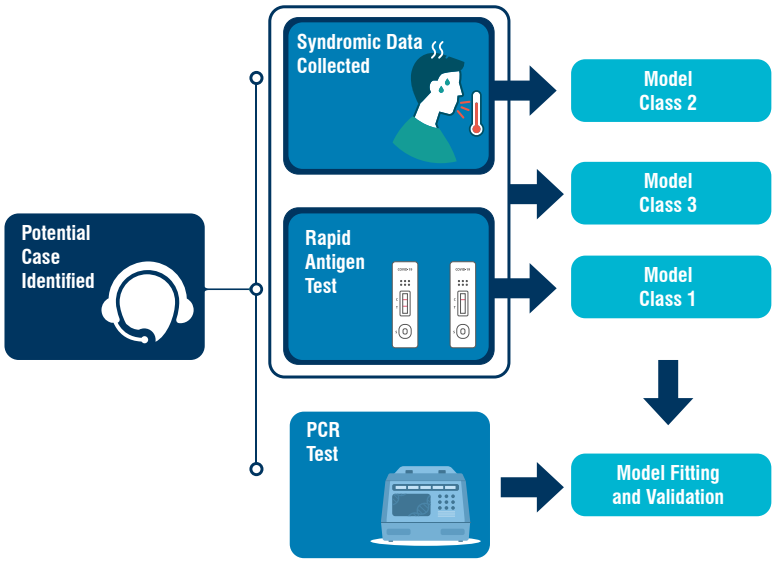

Figure 1: Schematic description of identification of likely COVID-19 patients by CSTs. The teams collected syndromic data (age, gender and presence/absence of 14 predetermined symptoms), and two sets of naso-pharyngeal swabs (one each for Rapid Antigen Testing and RT-PCR). We then used these two imperfect diagnostics (RAT and syndromic data) 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 PCR test result is used to train and test each model using temporal cross validation.

the uncertainty in the parameter estimates. Structurally, the multivariate probit model allows the symptoms and COVID-19 status to be treated as correlated binary outcomes with an intrinsic rate (the intercept for each variable) and the patient's age and gender, while propagating and quantifying uncertainty.

In Model Class 3, we model RAT positive patients as PCR positive and use the syndromic approach outlined for Model Class 2 for the RAT negative patients. The models were fitted to the data using Hamiltonian Monte Carlo in the Stan programming language.

We conducted backwards model selection (starting with the most complex model feasible, with all 14 symptoms and both covariates) to identify a subset of models with the highest predictive power under temporal cross validation. We identify a subset of the most predictive models to reduce computational demand and reduce the risk of overfitting models to the test scenarios. The large number of symptoms means that there is a high number of potential model configurations which might (>131000 for 14 symptoms and two covariates), by chance, perform well on the test sets (even under the challenging conditions of temporal cross validation). By first using general predictive power to narrow down the number of candidate models and then testing those models under more specific scenarios, we are more likely to choose models which generalise well to new data.

We scored the models' predictive power using cross entropy. Cross entropy measures the accuracy of probabilistic predictions for models that predict binary outcomes using probabilities. A cross entropy value close to zero corresponds to high levels of accuracy, with larger values indicating lower accuracy.
As the score only uses the predicted probability and true values, it is possible to directly compare the predictions of any model for the same test set. More details on the model structure and selection process, including code, are available in Appendix XX.

We then compared models as classifiers using their false positive and false negative rates in three epidemiological scenarios. 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.

ROC curves were drawn to show classifier performance across a range of unspecified scenarios, and error rates under three specific epidemiological scenarios were compared. ROC curves show the true and false positive rates that each model can achieve. Comparing specific scenarios allows classifier performance to be demonstrated in relevant scenarios. Whether measuring classifier performance in specific scenarios or more generally, decisions need to be made about the relative cost and acceptable levels of the two types of misclassification (false positives and negatives). We strongly emphasise that local context should be the guide in applying these methods.

In Scenario 1, we do not consider epidemiological context but simply weight false negative and false positive rates equally and aim to maximise the overall

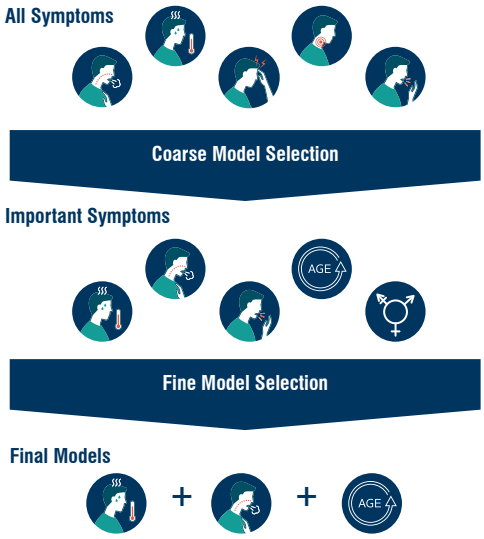

Figure 2: Schematic for rounds of model selection in the multivariate probit component of Model Classes 2 and 3. With 14 symptoms (only 5 shown here for demonstration purposes) and two covariates there are over 131000 possible model combinations. To make exploring these possible models computationally feasible and to reduce the risk of overfitting we carried out two rounds of model selection. 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 fine 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 >131000 to just four per model class.

Table 1: For each 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. The requirement determines a 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	20% max false positive rate	False negative rate
3 Steady, Low-Level Cases	20% max false negative rate	False positive rate

correct classification rate. Scenario 2 corresponds to the current situation in Bangladesh at time of writing (June 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.

4. Results

A total of 511 subjects had data available for the current analyses. The mean age of women participants (56% of the sample) was 38.5 (SD = 15.4), and for men (44% of the sample) was 41.8 (SD = 17.2). Participants were self-selecting and drawn from across Dhaka, the capital of Bangladesh.

Model selection for Model Class 2 and 3 each retained age as an explanatory variable and showed a marked decline in predictive power at more than 4 symptoms. 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 wet cough, runny nose, loss of smell and breathing problems for Model Class 2, and fever, wet cough, tiredness and diarrhoea for Model Class 3. For both Model Class 2 (syndromic data only) and Model Class 3 (syndromic and RAT data), model selection retained age as a covariate but not gender.

In the comparison of model predictive performance, Model Class 1 (RAT only) performed worst with a cross entropy of 4.79 (cross entropy values further from zero correspond to worse predictive performance). The median cross entropy values were between 2.90 and 2.95 for models in Class 2 (syndromic data only). Models in Class 3 (combined data model) performed best with cross entropy values between 2.40 and 2.43 (see Figure 3).

General model classification performance is shown by the full ROC curves for each model (Figure 4).

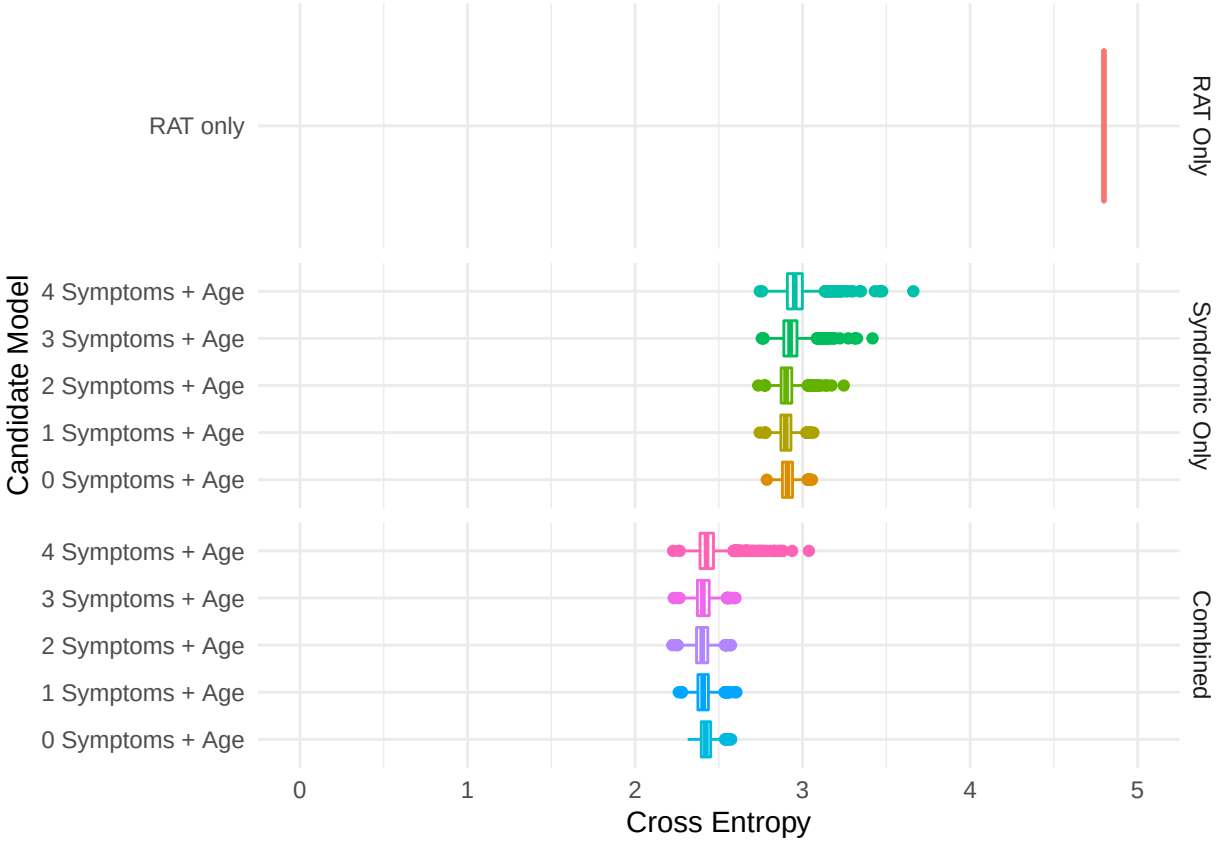

Figure 3: Interquartile ranges for the posterior cross entropy of the best candidate models at each level of model complexity tested under temporal cross-validation. Cross entropy is a measure of distance from the truth, so values closer to zero indicate better models. The intermediate complexity models perform best at prediction, although performance is similar across all the models within each class.

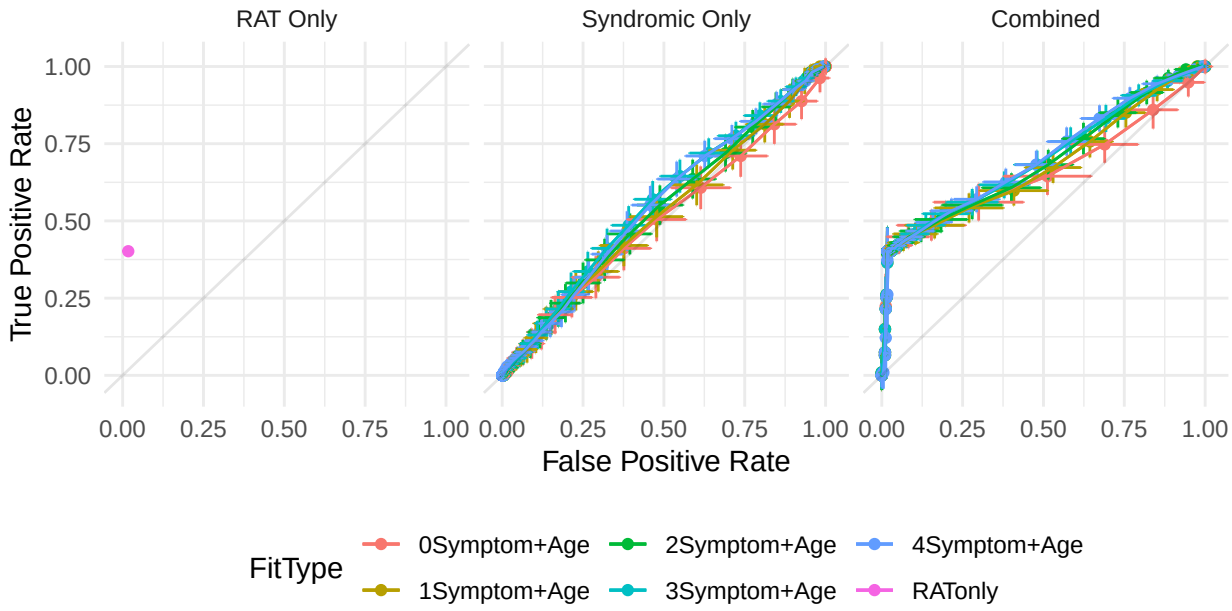

Figure 4: Receiver operating characteristics for RAT only approach and posterior mean (+- posterior standard deviation) receiver operating characteristics for Class 2 and 3 models. These curves demonstrate the performance of the model for any given scenario as defined by false and true positive rates (as opposed to Figure 5 which demonstrates model performance in specific scenarios).

Scenario specific classification performance is shown in Figure 5. In Scenario 1, the median error was 0.62 for models in Class 1 and Class 3 and between 0.90 and 0.96 for models in Class 2 (Figure 5A). In Scenario 2, Model Class 1 was unable to meet the required false negative rate. The median errors were between 0.69 and 0.74 for models in Class 2, and 0.57 and 0.69 for models in Class 3 (Figure 5B). In Scenario 3, the error in Class 1 was 0.60 and the median errors ranged from 0.70 to 0.75 for Class 2, and 0.44 and 0.47 for Class 3 (Figure 5C).

5. Discussion

We have demonstrated that combining two imperfect diagnostics yields better prediction of COVID-19 status and greater flexibility than each diagnostic individually. These improvements are non-trivial in real-world settings like Bangladesh, where there are currently thousands of new cases being identified every day and the pandemic growth is accelerating meaning every missed case has a compounding effect. In the most relevant scenario, 2 where we try to keep the false negative rate low, the combined data model achieves 40 percentage points lower than the false negative rate of the RAT only model, and a 12 percentage points lower false positive rate than the syndromic only model. These are large

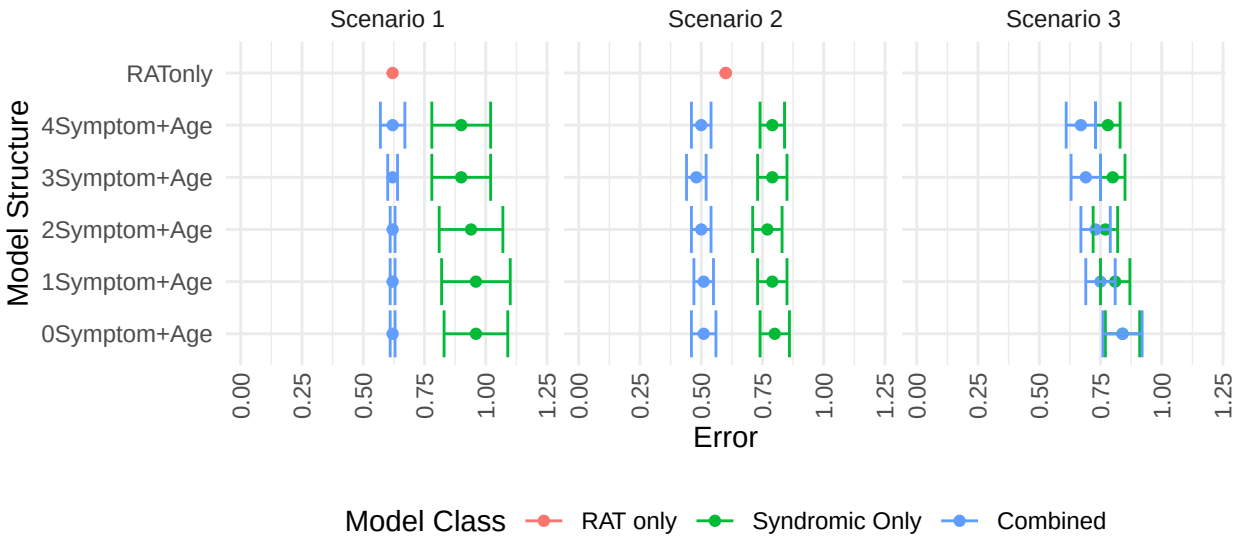

Figure 5: Performance of models under each scenario measured by errors defined in Table 2. Low errors correspond to better model performance. There is no error rate defined for the Model Class 1 (RAT only model) in Scenario 2 as the model failed to meet the requirement for that scenario (making the error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

performance gains for any diagnostic, although under the current situation will have a considerable impact on identifying both COVID-19 positive and negative patients. The pattern is similar in Scenario 3 with the predictive performance metrics showing that the combined model has equal or better performance across most potential scenarios. Furthermore, this boost is achieved with data that are already being collected in Bangladesh. Outwith developing and rerunning the models presented in this paper, these improvements are essentially cost free and eminently scalable.

Syndromic identification and RATs are fast, inexpensive and can be performed at patients' homes by minimally trained personnel. These imperfect detection methods have been developed as inexpensive alternatives to RT-PCR. While even the improvements described above will never allow these methods to compete with RT-PCR in terms of sensitivity and specificity, Rapid Antigen Testing and syndromic diagnosis also hold several further advantages. Unlike RT-PCR where patients have to go to designated testing centres and samples are taken by trained technologists, both of the imperfect diagnostics can be delivered in the community. This has several advantages. Firstly, increased accessibility by removing the need to travel for testing, thus reduces bias particularly against poorer, sicker and older people. Secondly, linked to cost, it is much easier to scale up testing when specialist training and expensive equipment and biosafety procedures are not required. Thirdly, it allows for assessment of an individual's wellbeing in the home context, and thus facilitates tailoring interventions to

where they are most needed.

The symptoms retained in the models may hint at the mechanism by which combining syndromic data and RAT results improves diagnosis. We have deliberately not emphasised the final symptoms chosen through model selection in this paper as we are focusing on prediction and classification for a unique sub-population: self-referring, symptomatic patients. We do, however, highlight that the symptoms retained in the final models for the syndromic-only and combined models are largely different (only one symptom appears in both lists). We propose that this is due to the fact that RAT is most effective during the first week of symptom onset, with much worse performance pre- and post- this period. These first-week symptoms are generally typical of viral pneumonias, and, indeed, when the RAT result is excluded from the model, the most important symptoms are typical of upper respiratory tract infections. However, when the RAT result is combined with syndromic data, the most important symptoms become much more eclectic. It is probable that these symptoms are either typical of later-stage COVID-19 or are less common presentations of the disease, possibly caused by co-infection or multimorbidities. Further research is needed to understand the mechanisms by which symptoms predict COVID-19 and by which RAT misses COVID-19. Of particular interest is whether individuals that are missed by RAT are less infectious, which could be explored by using Threshold Cycle (Ct) values from the RT-PCR to compare viral load with respect to prediction by the different methods. We note also that, as expected, age was retained in model selection. We were, however, surprised that gender was removed during model selection. Gender is thought to play a major role in infection risk. As we are looking to predict symptomatic COVID-19 in symptomatic individuals, generalised risk of infection is perhaps less predictive than expected.

We believe that the combined syndromic and rapid testing model represents the most promising approach to testing for COVID-19 in low- and middle-income countries at present. By taking a statistical modelling approach to case identification, we are able to update our diagnostic process in real time, allowing this method to readily adapt to new variants (or even new diseases) or new priorities for resource allocation. The modelling frameworks we have used are also sufficiently flexible to accommodate new data sources. Of particular interest are extensions to include the "pandemic context" in the model using space-time data. Furthermore, by using more sophisticated modelling structures it is possible to tune error rates to better reflect 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 is of high quality. Similarly, targeting error rates is only sensible if those rates properly reflect local conditions which is hard to do in practice. These limitations should be seriously considered but the alternatives for imperfect testing methods are diagnostics that cannot be tailored to local conditions at all (and, as such may perform worse than a method which is sub-optimally tailored to local conditions) or diagnostics which make these decisions implicitly and not explicitly. We believe that in choosing the latter these decisions are more readily challenged, researched and improved upon. We

also emphasise the need for rigorous experimental design to ensure findings from the sample population are applicable to the target population and the need for further research into understanding error rate tradeoffs in applied settings.

The methodology we have outlined here is applicable to a wide range of diseases and settings across low- and middle-income countries. One of the biggest challenges in diagnosing and tracking many diseases in resource-limited settings is the low availability of access to gold-standard testing (such as RT-PCR in the case of COVID-19) and high error rate of alternative testing methods. In this paper, we have outlined a process for coupling a small number of gold-standard tests with formal statistical integration of alternative testing methods, to generate high quality diagnostic models. This process readily maps onto many other case identification problems, including the diagnosis of several neglected tropical diseases. For example, malaria (gold standard (GS) is also RT-PCR, imperfect methods (IM) include antigen tests, syndromic diagnosis and blood smears), schistosomiasis (GS: RT-PCR or autopsy; IM: Kato Katz egg counts, antibody detection) and rabies (GS: fluorescent antibody test; IM: light microscopy, differential diagnosis).

In conclusion, we believe that the combined syndromic and rapid antigen testing approach represents the most promising approach to large-scale testing in low- and middle- income countries at present. By using the small amount of RT-PCR testing possible and formally integrating multiple imperfect, non-gold-standard methods, we can tune these diagnostics to our local conditions. 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 low- and middle-income countries. As such, these low-cost improvements to existing testing programs have the potential to identify one to two orders of magnitude more cases than either gold-standard or alternative methods alone.