

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

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

^a*Institute of Biodiversity, Animal Health and Comparative Medicine, University of Glasgow*

^b*COVID-19 in LMICs Research Group, University of Glasgow*

^c*School of Mathematics and Statistics, 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*

here() starts at /Users/fergusjchadwick/Dropbox/Git/COVID19_SyndromicRATDiagnosis

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

*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.kundergorski@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)

Bangladesh’s Institute of Epidemiology Disease Control And Research (IEDCR) identified potential COVID-19 patients in Dhaka using syndromic surveillance. 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 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 due to its high sensitivity and specificity for COVID-19 [2]. 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 [4]. COVID-19 diagnosis worldwide, therefore, must be made accessible using inexpensive methods that can be carried out locally [6].

An increasingly popular alternative to RT-PCR is rapid antigen testing (RAT) [7]. Like RT-PCR, these tests have high specificity for COVID-19 while being less expensive, easier to implement, and faster to produce results [8]. 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) or have highly visible PPE-clad officials visit their home. Then, invasive nasopharyngeal swabs must be taken and there is a delay in receiving the result (between one day and a week in Bangladesh). In contract, RAT can be conducted on nasal or saliva samples, completed in the home 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, several concerns have been raised about the sensitivity of RAT [9] leading to more false negative diagnoses.

Another alternative to RT-PCR, one that has been used since the start of the pandemic, is identifying cases through symptom-thresholding [10]. In this

71 approach, a patient presenting with a fever and one or more viral pneumonia
 72 symptoms is treated as a COVID-19 positive patient. The main advantage of
 73 this approach is the ease of implementation. For example, in Bangladesh, a lower
 74 middle-income country, much of the initial support and reporting of infections
 75 locally is provided by community support teams (CSTs) composed of local
 76 volunteers with basic training. The CSTs can easily collect symptomatic data in
 77 the community and provide care where the criteria are met. Unfortunately, these
 78 symptom-thresholds were developed early in the outbreak, and thus necessarily
 79 drawn from clinical intuition, rather than data. Consequently, the relationship
 80 between the criteria and the true COVID-19 status is often weak, with low
 81 specificity leading to a very large number of false positive diagnoses.

82 A natural extension to these symptom-threshold approaches is syndromic
 83 modelling. Here, a patient presenting with a fever and one or more viral
 84 pneumonia symptoms is treated as a potential COVID-19 patient. However,
 85 rather than using a set of pre-determined criteria, a range of symptomatic and
 86 risk factor data are collected and then a sub-sample of patients are tested using
 87 RT-PCR for COVID-19 [11]. These data are used to fit a model that allows
 88 more accurate prediction of how likely a patient is to have COVID-19 through
 89 the identification of COVID-19 syndromes [13]. It is worth highlighting at this
 90 point that in resource-limited settings, there is very limited provision for testing
 91 of asymptomatic cases, despite their important role in disease transmission [14].
 92 Even while focusing solely on symptomatic patients, syndromic modelling is a
 93 complex and nuanced task. The strength of relationships between symptoms
 94 and diseases is not stable through time or across sampling strategies since the
 95 relative importance of each symptom for disease diagnosis, in part, depends on
 96 the prevalence of other diseases causing similar symptoms in the community
 97 [15]. For example, if another disease for which loss of smell is a symptom
 98 becomes common, that symptom becomes a worse predictor for COVID-19.
 99 Similarly, if everyone who presents has a cough and thus is included in the
 100 sample, then coughing will likely have a very low correlation with COVID-19
 101 (even if the two are strongly related in the general population). While these
 102 issues can be overcome by properly considering the population sampled and
 103 using appropriately sophisticated statistical methods, the many types of common
 104 respiratory disease generally means that even then these models tend to have
 105 relatively high false positive rates (low specificity) for COVID-19 [15], although
 106 much lower than the symptom-threshold approach.

107 Poor sensitivity and specificity are problematic in diagnostics but higher
 108 error rates may be tolerable depending on their scale and impact given the local
 109 situation. Low specificity means a large number of false positive classifications,
 110 where the patient is told they have COVID-19 but they actually do not. This
 111 might lead to patients unnecessarily self-isolating and receiving support which
 112 can be expensive to the individuals and local public health bodies, as well
 113 as reducing available resources for those who need them [16]. Similarly, low
 114 sensitivity means more false negative classifications, where the patient is told
 115 they do not have COVID-19 but they actually do, which can lead to a health-risk
 116 for the individual and to the disease spreading further [17]. The costs of these

117 misclassifications will depend on local context. When the prevalence of the
118 disease is low, false positives may create local skepticism about the value of
119 testing, or when there are strong population-level mitigations already in place
120 (such as a nationwide lockdown), then false positives might be more costly than
121 false negatives [16]. If the disease is abundant or increasing rapidly then false
122 negatives are likely to be more costly. In most situations, a balance will need to
123 be struck [4].

124 The two dominant “alternative” testing methods available in resource limited
125 settings, therefore, are both flawed. Relying solely on symptomatic diagnosis
126 will likely overestimate the number of individuals with COVID-19 due to its lack
127 of specificity. Conversely, RATs will give a false impression of control due to the
128 number of positive cases that will be missed. In this paper, we demonstrate how
129 to combine these data types to exploit their complementarity and ameliorate
130 their respective weaknesses. We aim to compare the performance of these two
131 testing methods and the combined approach both in terms of general prediction
132 and as diagnostics under three epidemiological scenarios; and to demonstrate
133 that the combined data achieve equal to much lower error rates than the next
134 best method. We then discuss the role of statistically integrating data from
135 multiple imperfect testing methods in resource limited settings to improve the
136 diagnosis of diseases, particularly COVID-19.

137 3. Methods

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

142 Patients were selected for further testing conditional on the presence of a
143 fever ($>38^{\circ}\text{C}$) at the point of testing and one or more of 14 additional symptoms
144 associated with COVID-19 (breathing problems, coughing, diarrhoea, fever
145 (ongoing), a headache, loss of taste, loss of smell, muscle pain, red eyes, a runny
146 nose, a sore throat, tiredness, vomiting or a wet cough). The patient’s age
147 and gender were also recorded, but these data were not included in the patient
148 selection criteria.

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

156 We examined the ability of the two imperfect identification methods, the
157 syndromic profile and RAT result, to predict the patient’s COVID-19 status
158 when used separately and together. The different data combinations define three
159 model classes (Figure @ref(fig:data-flowchart)).

Find more
ways to cite
Figure 1 in
this text.

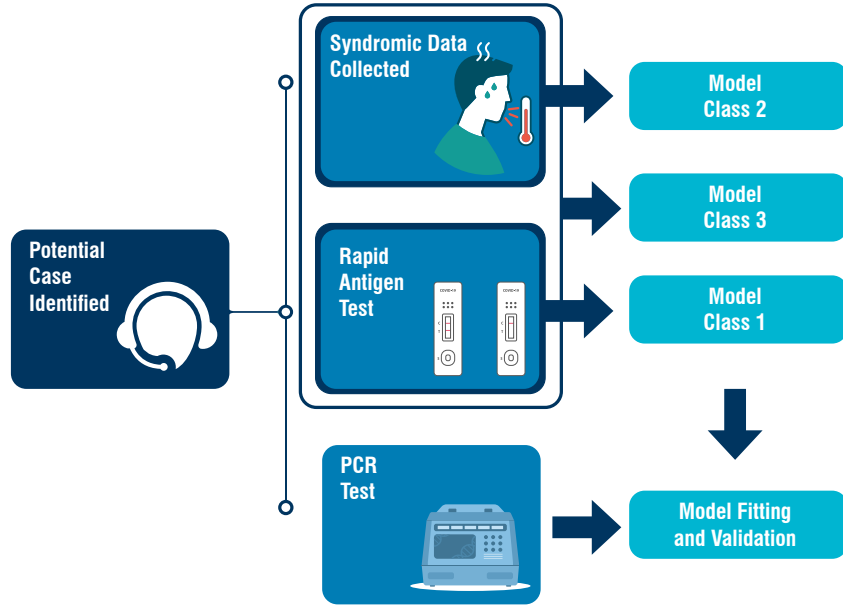


Figure 1: Schematic description of identification of likely COVID-19 patients by CSTs, swab collection and model definitions. The teams collected syndromic data (age, gender and presence/absence of 14 predetermined symptoms), and two sets of naso-pharyngeal swabs (one each for Rapid Antigen Testing and RT-PCR). We then used 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.

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 [18]. 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 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 [19].

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. Reducing the number of possible models to a small number of the most predictive models was necessary 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 (>131000 for 14 symptoms and two covariates) which might, by chance, 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 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 [20]. 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

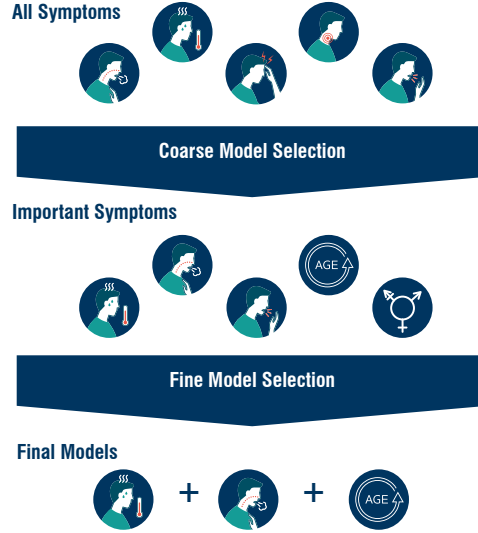


Figure 2: Schematic for rounds of model selection in the multivariate probit component of Model Classes 2 and 3. With 14 symptoms (only 5 shown here for demonstration purposes) and two covariates there are over 131000 possible model combinations. To make exploring these possible models computationally 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

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 by aiming to maximise the overall 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).

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 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 and other low- and middle- income countries. 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, samples need to be taken by trained technologists and results take a day to a week to come back, both of the imperfect diagnostics can be delivered in the community with almost instant

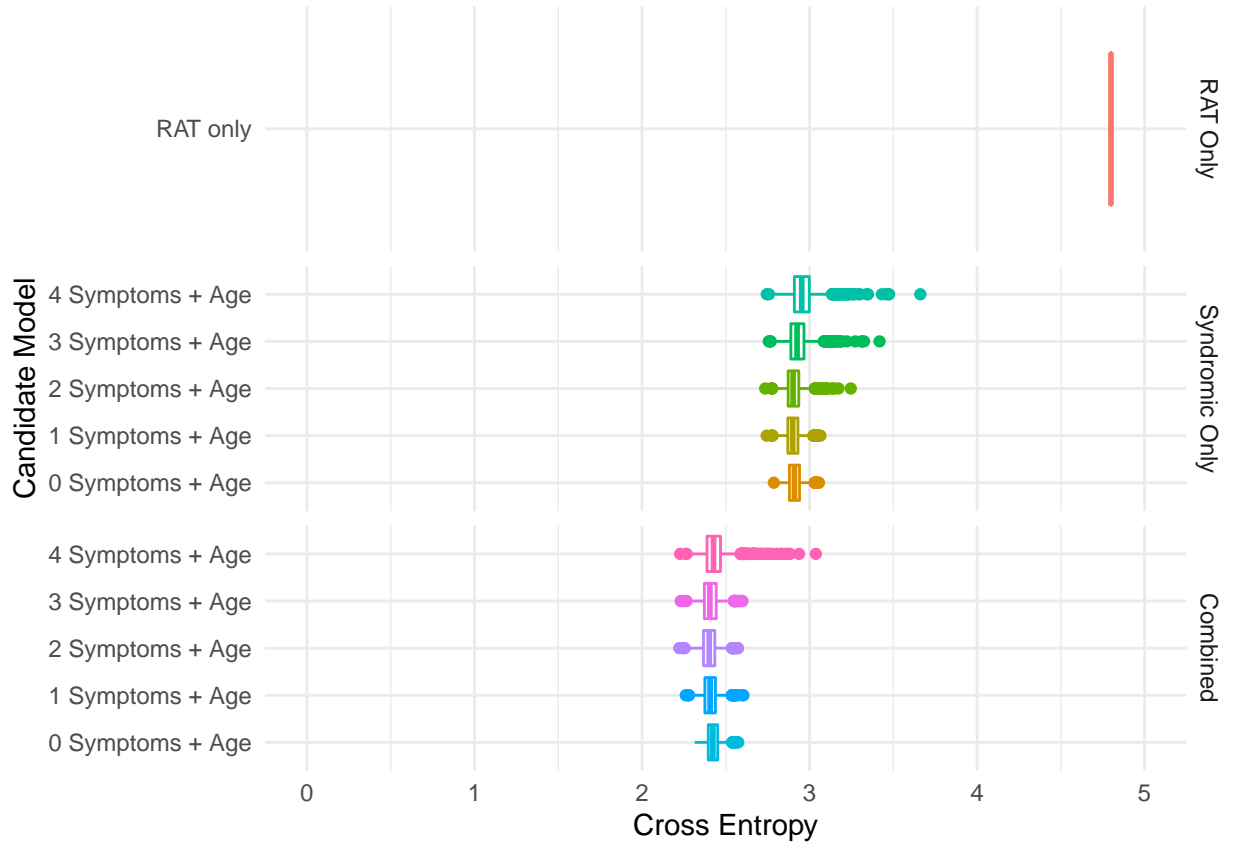


Figure 3: Interquartile ranges for the posterior cross entropy of the best candidate models at each level of model complexity tested under temporal cross-validation. Cross entropy is a measure of distance from the truth, so values closer to zero indicate better models. The intermediate complexity models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

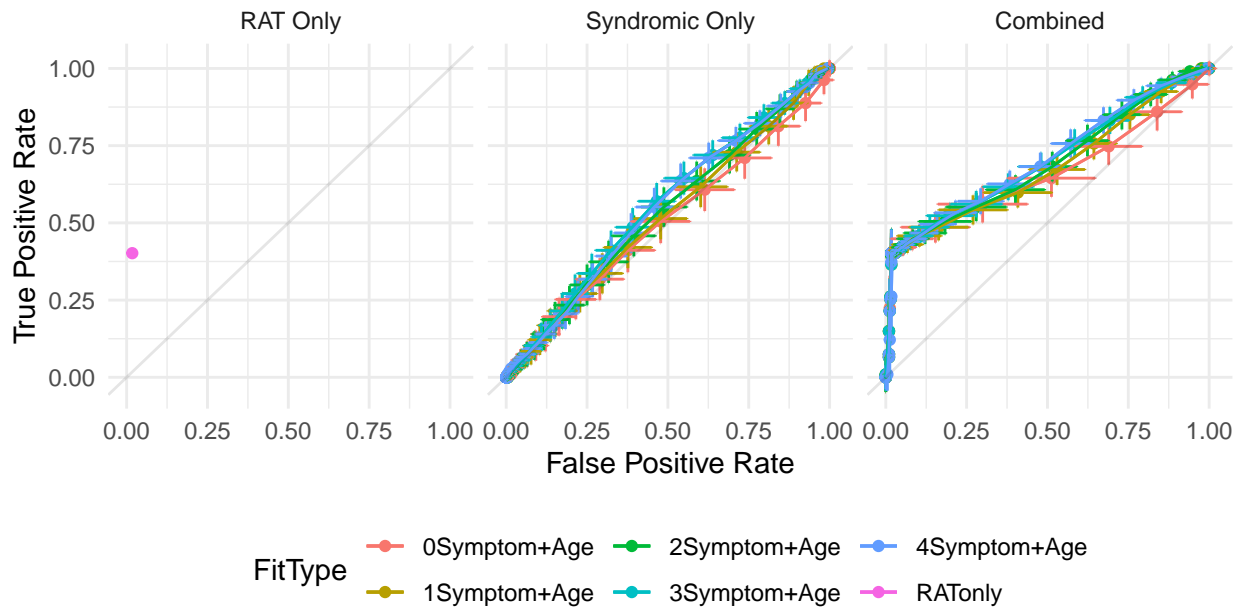


Figure 4: Receiver operating characteristics for RAT only approach and posterior mean (\pm posterior standard deviation) receiver operating characteristics for Class 2 and 3 models. These curves demonstrate the performance of the model for any given scenario as defined by false and true positive rates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

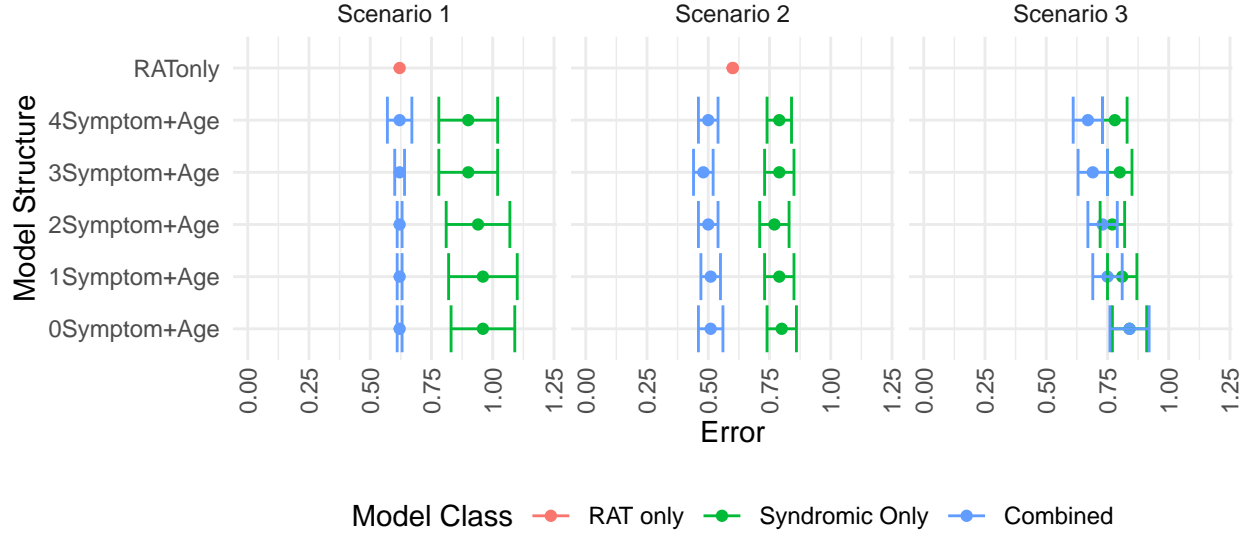


Figure 5: Performance of models under each scenario measured by errors defined in Table 2. Low errors correspond to better model performance. There is no error rate defined for the Model Class 1 (RAT only model) in Scenario 2 as the model failed to meet the requirement for that scenario (making the error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

278 result delivery. This has several advantages. Firstly, increased accessibility by
 279 removing the need to travel for testing, thus reduces bias particularly against
 280 poorer, sicker and older people. Secondly, linked to cost, it is much easier to
 281 scale up testing when specialist training and expensive equipment and biosafety
 282 procedures are not required. Thirdly, it allows for assessment of an individual's
 283 wellbeing in the home context, and thus facilitates tailoring interventions to
 284 where they are most needed.

285 The symptoms retained in the models may hint at the mechanism by which
 286 combining syndromic data and RAT results improves diagnosis. We have de-
 287 liberately not emphasised the final symptoms chosen through model selection
 288 in this paper as we are focusing on prediction and classification for a unique
 289 sub-population: self-referring, symptomatic patients. We do, however, highlight
 290 that the symptoms retained in the final models for the syndromic-only and
 291 combined models are largely different (only one symptom appears in both lists).
 292 We propose that this is due to the fact that RAT is most effective during the
 293 first week of symptom onset, with much worse performance pre- and post- this
 294 period [21]. These first-week symptoms are generally typical of viral pneumonias
 295 [22], and, indeed, when the RAT result is excluded from the model, the most
 296 important symptoms are typical of upper respiratory tract infections. However,
 297 when the RAT result is combined with syndromic data, the most important
 298 symptoms become much more eclectic. It is probable that these symptoms are
 299 either typical of later-stage COVID-19 or are less common presentations of the

300 disease, possibly caused by co-infection or multimorbidities. Further research is
 301 needed to understand the mechanisms by which symptoms predict COVID-19
 302 and by which RAT misses COVID-19. Of particular interest is whether indi-
 303 viduals that are missed by RAT are less infectious, which could be explored
 304 by using Threshold Cycle (Ct) values from the RT-PCR to compare viral load
 305 with respect to prediction by the different methods [23]. We note also that, as
 306 expected, age was retained in model selection. We were, however, surprised that
 307 gender was removed during model selection. Gender is thought to play a major
 308 role in infection risk [25]. As we are looking to predict symptomatic COVID-19
 309 in symptomatic individuals, generalised risk of infection is perhaps less predictive
 310 than expected, potentially due to the balancing of risk and burden [26].

311 We believe that the combined syndromic and rapid testing model represents
 312 the most promising approach to testing for COVID-19 in low- and middle-
 313 income countries at present. By taking a statistical modelling approach to
 314 case identification, we are able to update our diagnostic process in real time,
 315 allowing this method to readily adapt to new variants (or even new diseases) or
 316 new priorities for resource allocation. The modelling frameworks we have used
 317 are also sufficiently flexible to accommodate new data sources. Of particular
 318 interest are extensions to include the “pandemic context” in the model using
 319 space-time data. Furthermore, by using more sophisticated modelling structures
 320 it is possible to tune error rates to better reflect the local relative costs of false
 321 positives and false negatives. Naturally, these strengths have complementary
 322 limitations. Our models require updating in real-time and can only achieve good
 323 performance if the validation data is of high quality. Similarly, targeting error
 324 rates is only sensible if those rates properly reflect local conditions which is
 325 hard to do in practice. These limitations should be seriously considered but the
 326 alternatives for imperfect testing methods are diagnostics that cannot be tailored
 327 to local conditions at all (and, as such may perform worse than a method which
 328 is sub-optimally tailored to local conditions) or diagnostics which make these
 329 decisions implicitly and not explicitly. We believe that in choosing the latter
 330 these decisions are more readily challenged, researched and improved upon. We
 331 also emphasise the need for rigorous experimental design to ensure findings from
 332 the sample population are applicable to the target population and the need for
 333 further research into understanding error rate tradeoffs in applied settings.

334 The methodology we have outlined here is applicable to a wide range of
 335 diseases and settings across low- and middle-income countries. One of the biggest
 336 challenges in diagnosing and tracking many diseases in resource-limited settings
 337 is the low availability of access to gold-standard testing (such as RT-PCR in
 338 the case of COVID-19) and high error rate of alternative testing methods. In
 339 this paper, we have outlined a process for coupling a small number of gold-
 340 standard tests with formal statistical integration of alternative testing methods,
 341 to generate high quality diagnostic models. This process readily maps onto
 342 many other case identification problems, including the diagnosis of several
 343 neglected tropical diseases. For example, malaria (gold standard (GS) is also
 344 RT-PCR, imperfect methods (IM) include antigen tests, syndromic diagnosis
 345 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.

6. Funding

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.

- [1] Corman VM, Landt O, Kaiser M, Molenkamp R, Meijer A, Chu DK, et al. Detection of 2019 novel coronavirus (2019-nCoV) by real-time RT-PCR. *Eurosurveillance* 2020;25:2000045.
- [2] Tahamtan A, Ardebili A. Real-time RT-PCR in COVID-19 detection: Issues affecting the results. *Expert Review of Molecular Diagnostics* 2020;20:453–4.
- [3] Chowdhury R, Luhar S, Khan N, Choudhury SR, Matin I, Franco OH. Long-term strategies to control COVID-19 in low and middle-income countries: An options overview of community-based, non-pharmacological interventions. *European Journal of Epidemiology* 2020;35:743–8.
- [4] Vandenberg O, Martiny D, Rochas O, Belkum A van, Kozlakidis Z. Considerations for diagnostic COVID-19 tests. *Nature Reviews Microbiology* 2021;19:171–83.
- [5] Cash R, Patel V. Has COVID-19 subverted global health? *The Lancet* 2020;395:1687–8.
- [6] Olalekan A, Iwalokun B, Akinloye OM, Popoola O, Samuel TA, Akinloye O. COVID-19 rapid diagnostic test could contain transmission in low- and middle-income countries. *African Journal of Laboratory Medicine* 2020;9:1–8.
- [7] Linares M, Pérez-Tanoira R, Carrero A, Romanyk J, Pérez-García F, Gómez-Herruz P, et al. Panbio antigen rapid test is reliable to diagnose SARS-CoV-2 infection in the first 7 days after the onset of symptoms. *Journal of Clinical Virology* 2020;133:104659.

- 376 [8] Boum Y, Fai KN, Nikolay B, Mboringong AB, Bebell LM, Ndifon M,
et al. Performance and operational feasibility of antigen and antibody
rapid diagnostic tests for COVID-19 in symptomatic and asymptomatic
377 patients in cameroon: A clinical, prospective, diagnostic accuracy study.
The Lancet Infectious Diseases 2021.
- 378 [9] Mak GC, Cheng PK, Lau SS, Wong KK, Lau C, Lam ET, et al. Evaluation
of rapid antigen test for detection of SARS-CoV-2 virus. Journal of
379 Clinical Virology 2020;129:104500.
- 380 [10] Jin Y-H, Cai L, Cheng Z-S, Cheng H, Deng T, Fan Y-P, et al. A rapid
advice guideline for the diagnosis and treatment of 2019 novel coronavirus
(2019-nCoV) infected pneumonia (standard version). Military Medical
381 Research 2020;7:1–23.
- 382 [11] Sim JXY, Conceicao EP, Wee LE, Aung MK, Seow SYW, Teo RCY, et
al. Utilizing the electronic health records to create a syndromic staff
surveillance system during the COVID-19 outbreak. American Journal of
383 Infection Control 2021;49:685–9.
- 384 [12] Undurraga EA, Chowell G, Mizumoto K. COVID-19 case fatality risk
by age and gender in a high testing setting in latin america: Chile,
385 march–august 2020. Infectious Diseases of Poverty 2021;10:1–1.
- 386 [13] Wenham C, Smith J, Morgan R. COVID-19: The gendered impacts of
the outbreak. The Lancet 2020;395:846–8.
- 387 [14] Mayorga L, Samartino CG, Flores G, Masuelli S, Sánchez MV, Mayorga
LS, et al. A modelling study highlights the power of detecting and
isolating asymptomatic or very mildly affected individuals for COVID-19
388 epidemic management. BMC Public Health 2020;20:1–1.
- 389 [15] Garry S, Abdelmagid N, Baxter L, Roberts N, Waroux O le P de, Ismail
S, et al. Considerations for planning COVID-19 treatment services in
390 humanitarian responses. Conflict and Health 2020;14:1–1.
- 391 [16] Surkova E, Nikolayevskyy V, Drobniewski F. False-positive COVID-19
results: Hidden problems and costs. The Lancet Respiratory Medicine
392 2020;8:1167–8.
- 393 [17] West CP, Montori VM, Sampathkumar P. COVID-19 testing: The threat
of false-negative results. Mayo clinic proceedings, vol. 95, Elsevier; 2020,
394 p. 1127–9.
- 395 [18] Albert JH, Chib S. Bayesian analysis of binary and polychotomous re-
sponse data. Journal of the American Statistical Association 1993;88:669–
396 79.
- 397 [19] Carpenter B, Gelman A, Hoffman MD, Lee D, Goodrich B, Betancourt M,
et al. Stan: A probabilistic programming language. Journal of Statistical
398 Software 2017;76:1–32.
- 399 [20] Gneiting T, Raftery AE. Strictly proper scoring rules, prediction, and
400 estimation. Journal of the American Statistical Association 2007;102:359–
401 78.

402 [21] Muhi S, Tayler N, Hoang T, Ballard SA, Graham M, Rojek A, et al. Multi-site assessment of rapid, point-of-care antigen testing for the diagnosis of SARS-CoV-2 infection in a low-prevalence setting: A validation and implementation study. *The Lancet Regional Health-Western Pacific* 2021;9:100115.

403

404 [22] Xu T, Chen C, Zhu Z, Cui M, Chen C, Dai H, et al. Clinical features and dynamics of viral load in imported and non-imported patients with COVID-19. *International Journal of Infectious Diseases* 2020;94:68–71.

405

406 [23] Albert E, Torres I, Bueno F, Huntley D, Molla E, Fernández-Fuentes MÁ, et al. Field evaluation of a rapid antigen test (panbio™ COVID-19 ag rapid test device) for COVID-19 diagnosis in primary healthcare centres. *Clinical Microbiology and Infection* 2021;27:472–e7.

407

408 [24] Menni C, Valdes AM, Freidin MB, Sudre CH, Nguyen LH, Drew DA, et al. Real-time tracking of self-reported symptoms to predict potential COVID-19. *Nature Medicine* 2020;26:1037–40.

409

410 [25] Wolff D, Nee S, Hickey NS, Marschollek M. Risk factors for covid-19 severity and fatality: A structured literature review. *Infection* 2020:1–4.

411

412 [26] Joe W, Kumar A, Rajpal S, Mishra U, Subramanian S. Equal risk, unequal burden? Gender differentials in COVID-19 mortality in india. *Journal of Global Health Science* 2020;2.

413