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[c], Shayan Chowdhury[e], Jessica Clark[a,b], Joanna Andrecka[f], Mikolaj Kundergorski[d,b], Craig Wilkie[d,b], Eric Brum[f], Tahmina Shirin[g], A S M Alamgir[g], Mahbubur Rahman[g], Ahmed Nawsher Alam[g], Farzana Khan[g], Janine Illian[d,b], Ben Swallow[d,b], Davina L Hill[a,b], Dirk Husmeier[d], 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] *MRC Biostatistics Unit, University of Cambridge*
[d] *School of Mathematics and Statistics, University of Glasgow*
[e] *a2i, United Nations Development Program, ICT Ministry, 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 (252 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

Community support teams in Dhaka, Bangladesh identified potential COVID-19 patients in Dhaka using syndromic surveillance. A sample (n = 1172) of

*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)

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. «««< HEAD In the "Rising Cases" scenario, which most closely represents the current situation in many LMICs, the combined data model false negative rate is 26 (IQR: 24-29) percentage points lower that of the RAT only model.

Interpretation

Small, scalable improvements in the accuracy of mass-deployed but imperfect diagnostic methods 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, the Wellcome Trust and the Engineering and Physical Sciences Research Council.

2. Introduction (1079 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 [1]) due to its high sensitivity and specificity for COVID-19 [3]. 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 [5], and 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 [7].

An increasingly popular alternative to RT-PCR is rapid antigen testing (RAT) [8]. Like RT-PCR, these tests have high specificity for COVID-19 while being less expensive, easier to implement, and faster but with lower sensitivity [9]. For RT-PCR testing, patients must travel to a designated site or have officials visit their home in enhanced personal protective equipment. 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 [10] leads to more false negative diagnoses.

Another diagnostic that has been used since the start of the pandemic is symptom-thresholding [11]. 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 [12]. 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 [14].

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 [15]. 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 or less common [16]. 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 [16].

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 [17]. 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 [18]. 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 [17]. 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 [5].

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.

3. Methods (950 words)

3.1. 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.

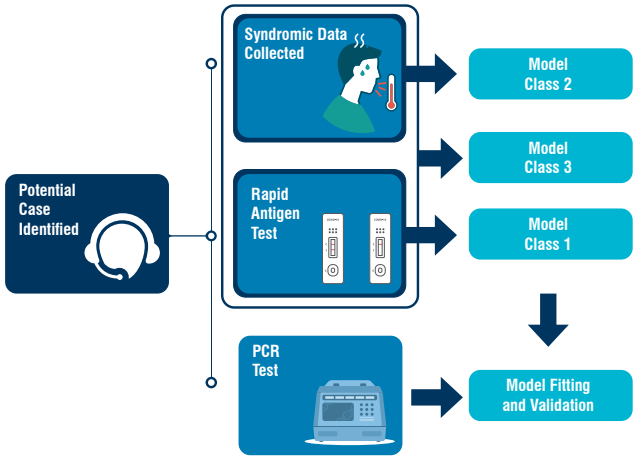

Figure 1: Schematic description of identification of likely COVID-19 patients by community support teams (CSTs), swab collection and model definitions. The teams collected syndromic data (age, gender and presence/absence of 14 predetermined symptoms), and two sets of naso-pharyngeal swabs (one each for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

3.2. Modelling

3.2.1. 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 1).

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 [19]. 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 [20].

3.2.2. 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 2). 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).

3.2.3. 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 [21]. 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.

3.2.4. 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 [22]) and under three epidemiological scenarios (using error terms described in Table 1). 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

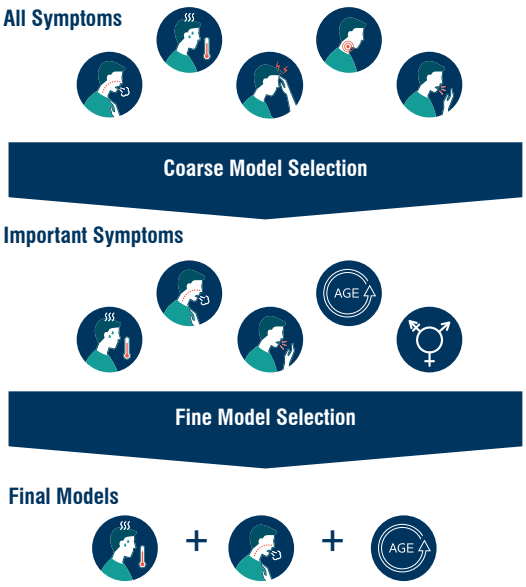

Figure 2: Schematic for rounds of model selection in the multivariate probit component of Model Classes 2 and 3. With 14 symptoms (only 5 shown here for demonstration purposes) and two covariates there are over 131000 possible model combinations. To make exploring these possible models computationally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

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 (IEDCR), Bangladesh, for illustrative purposes.

4. Results (550 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 $3 \cdot 24$ (cross-entropy

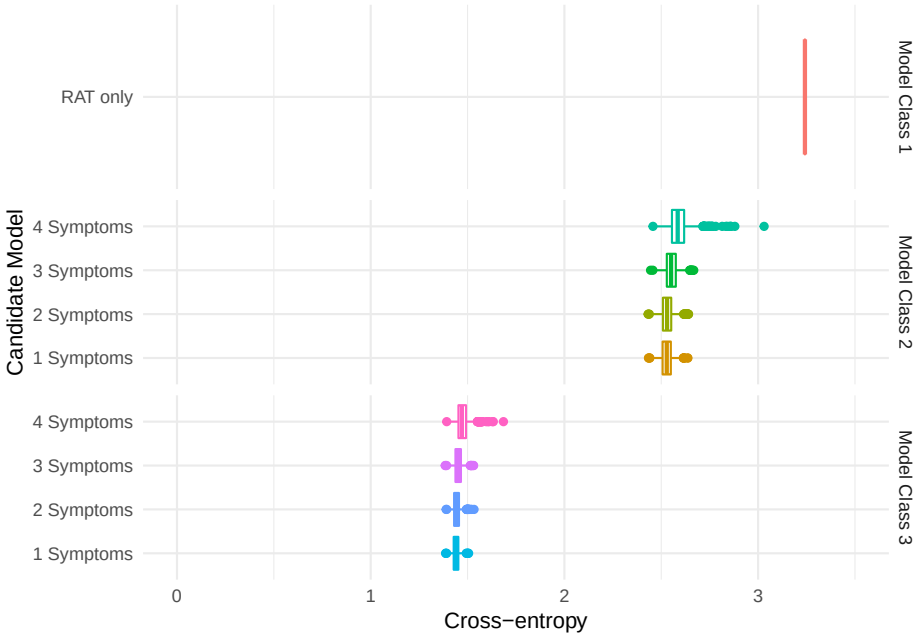

Figure 3: Predictive performance of candidate models. Interquartile ranges for the posterior cross-entropy of the best candidate models at each level of model complexity tested under temporal cross-validation. cross-entropy is a measure of distance from the truth, so values closer to zero indicate better 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).

values further from zero correspond to worse predictive performance). The median cross-entropy values were between $2 \cdot 53$ and $2 \cdot 59$ for models in Class 2. Models in Class 3 performed best with cross-entropy values between $1 \cdot 44$ and $1 \cdot 47$ (see Figure 3).

Generic model classification performance is shown by their ROC curves (Figure 4).

Scenario specific classification performance is shown in Figure 5. Across all three scenarios (defined in Table 1), 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. The median error was $0 \cdot 47$ for models in Class 1 and Class 3 and between $0 \cdot 87$ and $0 \cdot 9$ for models in Class 2 (Figure 5). 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. The median errors were between $0 \cdot 75$ and $0 \cdot 76$ for models in Class 2, and $0 \cdot 44$ and $0 \cdot 49$ for models in Class 3 (Figure 5). In Scenario 3 ("Low-Level Cases"),

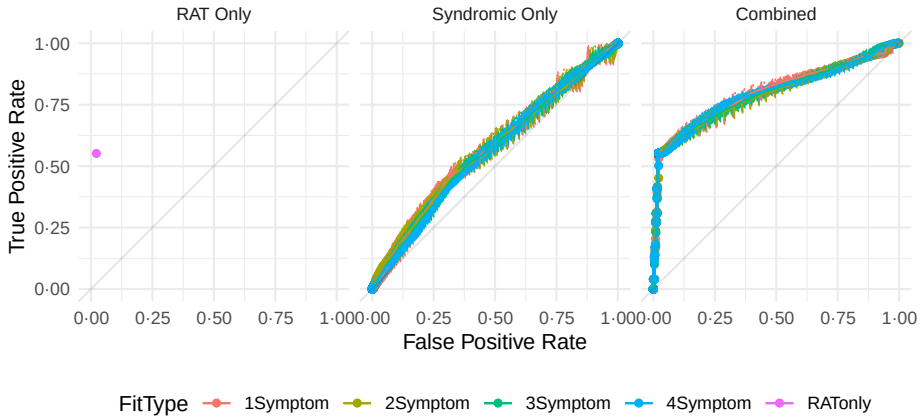

Figure 4: Receiver operating characteristics for rapid antigen testing (RAT) only approach (Model Class 1) and posterior median and interquartile range ROC for Class 2 (syndromic data only) and 3 (syndromic and RAT data) models. These curves demonstrate the performance of the model for any hypothetical 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).

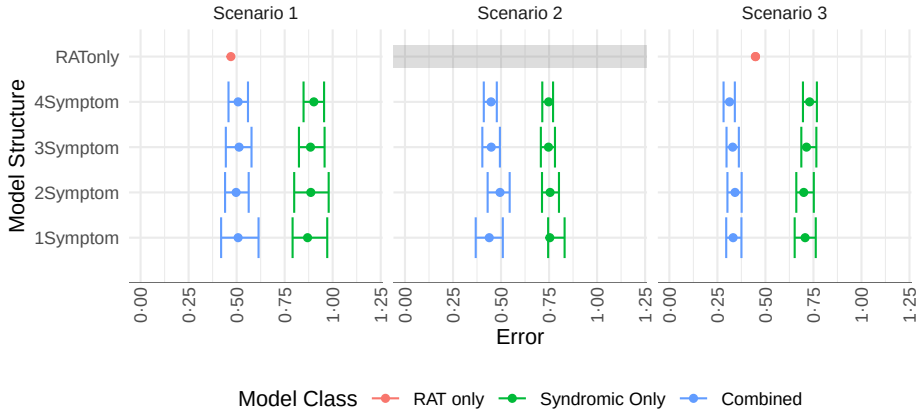

Figure 5: Performance of models under each scenario measured by posterior median and interquartile range for errors defined in Table 1. Low errors correspond to better model performance. There is no error rate defined for the Model Class 1 (RAT only model) in Scenario 2 as the model failed to meet the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

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). The error in Class 1 was $0 \cdot 02$ and the median errors ranged from $0 \cdot 19$ to $0 \cdot 2$ for Class 2, and $0 \cdot 18$ to $0 \cdot 2$ for Class 3 (Figure 5). 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.

5. Discussion (816 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 1 where the pandemic is once again accelerating. In this scenario, the combined data model (Model Class 3) false negative rate is 26 (IQR: 24-29) 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 (IQR: 29- 34) 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 5). 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 3 and 4, respectively). 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 [23].

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 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 adaptable 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 testing at scale. 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.

6. Funding (26 words)

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

Find Fergus PhD funding code

7. 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, Frances Mair, 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)

[1] Dramé M, Teguo MT, Proye E, Hequet F, Hentzien M, Kanagaratnam L, et al. Should RT-PCR be considered a gold standard in the diagnosis of covid-19? Journal of Medical Virology 2020.

[2] 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.

[3] 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.

[4] 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.

[5] Vandenberg O, Martiny D, Rochas O, Belkum A van, Kozlakidis Z. Considerations for diagnostic COVID-19 tests. Nature Reviews Microbiology 2021;19:171–83.

[6] Cash R, Patel V. Has COVID-19 subverted global health? The Lancet 2020;395:1687–8.

[7] 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.

[8] 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.

[9] 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 patients in cameroon: A clinical, prospective, diagnostic accuracy study. The Lancet Infectious Diseases 2021.

[10] 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 Clinical Virology 2020;129:104500.

[11] 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 Research 2020;7:1–23.

[12] 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 Infection Control 2021;49:685–9.

[13] Undurraga EA, Chowell G, Mizumoto K. COVID-19 case fatality risk by age and gender in a high testing setting in latin america: Chile, march–august 2020. Infectious Diseases of Poverty 2021;10:1–1.

[14] Wenham C, Smith J, Morgan R. COVID-19: The gendered impacts of the outbreak. The Lancet 2020;395:846–8.

[15] 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 epidemic management. BMC Public Health 2020;20:1–1.

[16] 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 humanitarian responses. Conflict and Health 2020;14:1–1.

[17] Surkova E, Nikolayevskyy V, Drobniewski F. False-positive COVID-19 results: Hidden problems and costs. The Lancet Respiratory Medicine 2020;8:1167–8.

[18] West CP, Montori VM, Sampathkumar P. COVID-19 testing: The threat of false-negative results. Mayo clinic proceedings, vol. 95, Elsevier; 2020, p. 1127–9.

[19] Albert JH, Chib S. Bayesian analysis of binary and polychotomous response data. Journal of the American Statistical Association 1993;88:669–79.

[20] Carpenter B, Gelman A, Hoffman MD, Lee D, Goodrich B, Betancourt M, et al. Stan: A probabilistic programming language. Journal of Statistical Software 2017;76:1–32.

[21] Gneiting T, Raftery AE. Strictly proper scoring rules, prediction, and estimation. Journal of the American Statistical Association 2007;102:359–78.

[22] Hoo ZH, Candlish J, Teare D. What is an ROC curve? 2017.

[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.