S2 Statistical Methodology for *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*

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

0.1. Modelling

0.1.1. 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

*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)

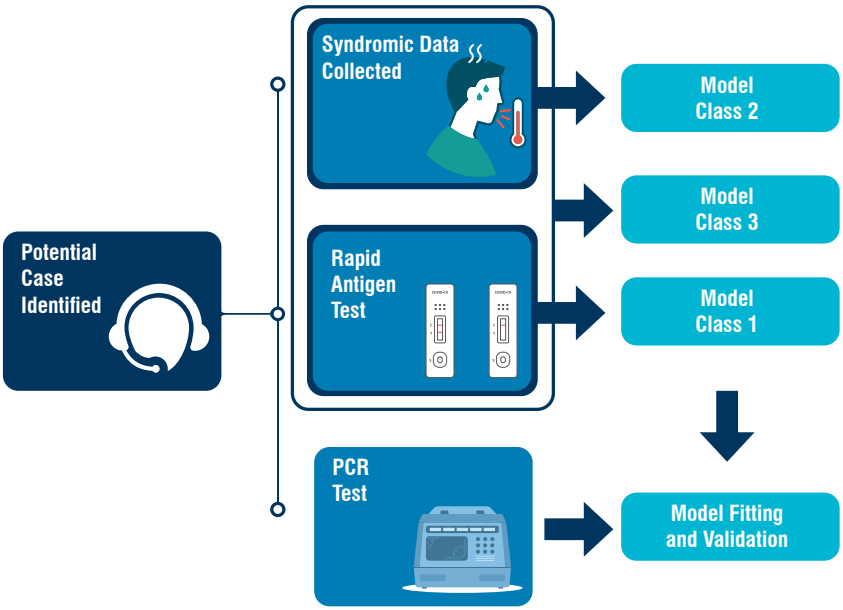

Figure 1: Schematic description of identification of likely COVID-19 patients by community support teams (CSTs), swab collection and model definitions. The teams collected syndromic data (age, gender and presence/absence of 14 predetermined symptoms), and two sets of naso-pharyngeal swabs (one each for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

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. We can write this formally for the ith individual as:

$$\begin{aligned} \text{PCR}_i &= \text{RAT}_i \\ \text{PCR} &\in \{0, 1\} \\ \text{RAT} &\in \{0, 1\} \end{aligned} \tag{1}$$

Model Class 2 uses only the syndromic data. For this model, we used a Bayesian multivariate probit model [1]. The multivariate probit structures the outcomes of the PCR test and symptoms presence/absence as a D-dimensional vector of binary outcomes ($\boldsymbol{y}_i = (y_{i1}, y_{i2}, \ldots, y_{id})$, $y_{ij} \in \{0, 1\}$). These outcomes are determined by an indicator function which takes a D-dimensional vector of *continuous latent* variables ($\boldsymbol{z}_i = (z_{i1}, z_{i2}, \ldots, z_{id})$, $z_{ij} \in \mathbb{R}$). These latent continuous variables then covary as realisations of a D-dimensional multivariate normal, with the mean of the error structure informed by a linear predictor, $\sum_{j=1}^{J} x_{ij}\beta_{jd} + \epsilon_{id}$, and a covariance between dimensions, Σ. The linear predictor allows us to condition the outcomes on risk factor variables (here, age and gender). The covariance structure allows us to account for the correlated nature of the symptoms with each other and the outcome. This multivariate approach (multiple response variables) is also a very efficient way of encoding higher order interactions. The contribution of symptoms might be heavily context dependent, for example, having a fever, a cough or a loss of taste might be uninformative symptoms independently, but in combination they might be highly diagnostic. These interaction terms use a large number of parameters and can be hard to fit to data. Using a multivariate structure allows us to exploit more efficient posterior sampling algorithms, and in higher dimensional settings like this uses fewer parameters.

As Albert and Chib identify in their paper [1], the covariance matrix formulation of the model is not identifiable, with the variance, $diag(\Sigma)$ and means of the latent variables, $\boldsymbol{z}_i$ trading off against each other. For this reason, we use a correlation matrix, Ω, formulation with the variance set to 1. A correlation based framework also makes communication with clinicians and other practitioners smoother as correlations are more familiar.

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

Model Class 3 combines the two data sources. We utilise the specificity of RAT by treating RAT-positive patients as PCR-positive patients. The RAT-negative patients are modelled using the sensitive syndromic approach using Model Class 2 to capture PCR-positive patients that are missed by the RAT. This approach leverages the potential different syndromic profiles of PCR-positve patients who are RAT-positive and -negative, allowing the model to adapt solely to the latter. Structurally, the model combines Equations (1) and (2), with RAT-positive patients being modelled using Equation (1), and RAT-negative patients with Equation (2).

By using a Bayesian formulation, we generate full posteriors for our parameter estimates, allowing natural quantification of uncertainty. Bayesian methods also facilitate the use of more informative priors. While we used minimally informative priors here (standard normals in the probit scale for betas and an LKJ correlation prior with minimal shrinkage, $\eta = 1$ [2]), more informative priors that incorporate spatio-temporal effects, for instance, would be natural extensions. The models were fitted to the data using Bayesian inference techniques based on Hamiltonian Monte Carlo in the Stan programming language [3]. The models all converged with zero divergent transitions and large effective sample sizes.

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

The coarse round of model selection (Figure 2) selected candidate symptoms based on whether they had a strong and consistent correlation with PCR as estimated according to Equations (2) and (**??**). The models were fit with both covariates throughout the coarse round and symptoms were compared in nested models. In the fine round of model selection, these candidate symptoms and the

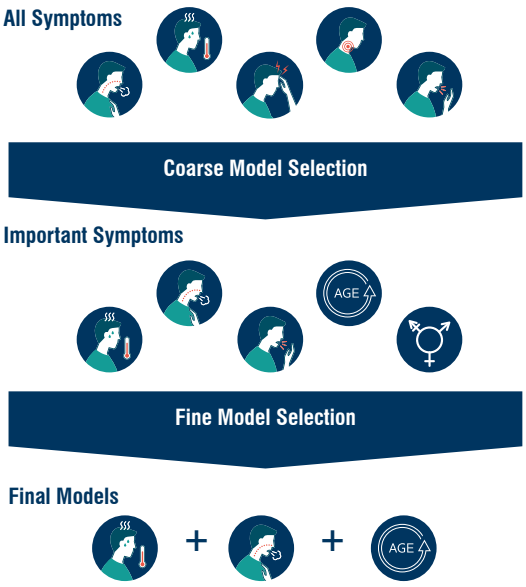

Figure 2: Schematic for rounds of model selection in the multivariate probit component of Model Classes 2 and 3. With 14 symptoms (only 5 shown here for demonstration purposes) and two covariates there are over 131000 possible model combinations. To make exploring these possible models computationally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

covariate combinations (age and gender, age, gender and no covariates) were permuted to more exhaustively explore the model space. Reducing the number of possible models using the two stages of model selection was necessary to reduce computational demand and reduce the risk of overfitting models to the test scenarios. The large number of symptoms corresponds to a high number of potential model configurations (>131 000 for 14 symptoms and two covariates) which might perform well on the test sets (even under the challenging conditions of temporal cross-validation) but lack transferability.

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

0.1.3. Predictive Performance

We scored the models' predictive power using binary cross-entropy (hereafter, cross-entropy). Cross-entropy measures the accuracy of models that generate probabilities of binary outcomes, rather than make binary classifications, similar in concept to a mean square error for normally-distributed data, but adapted for binary data [4]. A cross-entropy value close to zero corresponds to high levels of accuracy, with larger values indicating lower accuracy. More specifically, the metric allows us to compare a binary vector, $\boldsymbol{y} \in [0, 1]$, with a vector of probabilistic predictions ($p(\boldsymbol{y}) \in (0, 1)$) as follows:

$$\boldsymbol{H}_p(q) = \frac{1}{N} \sum_{i=1}^{N} y_i \cdot log(p(y_i)) + (1 - y_i) \cdot log(1 - p(y_i)) \tag{3}$$

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

0.1.4. 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, $\hat{Y}$, a probability threshold, $\hat{p}$, must be chosen over which patients are classified as COVID-19 positive:

$$\hat{Y} = \begin{cases} 1, & \text{if } p(y) \geq \hat{p} \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

Receiver operating characteristics (ROCs) are a way to measure the performance of a set of classifications in terms of true and false positives and negatives (TP, FP, TN and FN) and the rates of each of these classification types (e.g. $TPR = \frac{TP}{TP+FN}$, and $FPR = \frac{FP}{FP+TN}$). The error rates are calculated with respect to a particular threshold, $\hat{p}$, or across the range of possible $\hat{p}$s to generate a ROC curve [5]. In our epidemiological scenarios (outlined below) we use our ROC curve calculations to identify single thresholds which yield a required error rate.

We strongly emphasise that generic performance here is only used to show the flexibility of the model classes; the best model for a local situation can only be determined if the relative cost of false positives and false negatives is known. Here, we choose three representative scenarios. Each scenario has a requirement and error rate (defined in Table 1). We identify the threshold, $\hat{p}$, at which the requirement is most closely exceeded (i.e. if the requirement is an error rate should be a maximum 15%, the threshold that produces an error rate below 15% but as close to 15% as possible will be chosen).

In Scenario 1, we do not consider epidemiological context but simply minimise false negative and false positive rates equally. We do this by maximising the two correct classification rates both individually and in total, as measured by the harmonic mean. The harmonic mean is used widely in the classification literature as it is maximised by achieving large values in all its component parts, rather than the arithmetic mean which can be maximised by having one extremely large component at the expense of other components. In other words, the arithmetic mean could be large because it has a very high TPR but a small TNR, whereas the harmonic mean will maximise both TPR and TNR. While conceptually the harmonic mean is better suited than the arithmetic for this use case, both produce qualitatively the same results for these data.

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.

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

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 exceeds that requirement (i.e. for a 20The 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% false negative rate	False positive rate
3 Low-Level Cases	20% false positive rate	False negative rate

[2] Lewandowski D, Kurowicka D, Joe H. Generating random correlation matrices based on vines and extended onion method. Journal of Multivariate Analysis 2009;100:1989–2001.

[3] 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.

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

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