SUPPLEMENTARY MATERIALS: EXTENDING THE SUSCEPTIBLE-EXPOSED-INFECTED-REMOVED (SEIR) MODEL TO HANDLE THE *fa*LSE *n*EGATIVE RATE AND *sy*MPTOM-BASED ADMINISTRATION OF COVID-19 DIAGNOSTIC TESTS: *SEIR-fansy*

Ritwik Bhaduri [*]
Indian Statistical Institute
Kolkata, India.
ritwik.bhaduri@gmail.com

Ritoban Kundu [†]
Indian Statistical Institute
Kolkata, India.
ritoban.kundu@gmail.com

Soumik Purkayastha
Department of Biostatistics
University of Michigan, Ann Arbor, USA.
soumikp@umich.edu

Michael Kleinsasser
Department of Biostatistics
University of Michigan, Ann Arbor, USA.
mkleinsa@umich.edu

Lauren J. Beesley
Department of Biostatistics
University of Michigan, Ann Arbor, USA.
lbeesley@umich.edu

Bhramar Mukherjee [‡]
Department of Biostatistics and Epidemiology
University of Michigan, Ann Arbor, USA.
bhramar@umich.edu

Jyotishka Datta
Department of Statistics
Virginia Tech, Blacksburg, VA, USA.
jyotishka@vt.edu

February 5, 2022

S.1 Additional details for the SEIR-*fansy* model

S.1.1 Basic reproduction number

The basic reproduction number (or reproductive ratio) is defined as the number of infections that are expected to occur on average in a homogeneous population as a result of infection by a *single infectious individual* when the entire population is susceptible at the start of the pandemic. We derive an analytical formula for the reproduction number using the Next Generation Matrix Method shown in [13]. In the Next Generation Matrix Method, we start by calculating the next generation matrix, and the spectral radius of the next generation matrix gives us the basic reproduction number. The resulting expression for R_0 is as follows:

$$R_0 = \frac{\beta_t \cdot S_0}{\mu D_E + 1} \left(\frac{\alpha_u (1 - r_t)}{\frac{1}{\beta_1 D_r} + \delta_1 \mu_c + \mu} + \frac{\alpha_p r_t (1 - f)}{\frac{1}{D_r} + \mu_c + \mu} + \frac{r_t f}{\frac{\beta_2}{D_r} + \frac{\mu_c}{\delta_2} + \mu} \right) \tag{S.1}$$

where we define

$$S_0 = \begin{cases} \lambda/\mu & \text{if } \mu \neq 0 \\ 1 & \text{if } \mu = 0. \end{cases}$$

[*]Co-First Author
[†]Co-First Author
[‡]Corresponding author

S.1.1.1 Calculation of R_0 for Misclassification Model

We calculate the basic reproduction number R_0 using the **The Next Generation Matrix Method** as described by van den Driessche [13]. Suppose the whole population is divided into n compartments in which there are $m < n$ infected compartments. Let $x_i, i = 1, 2, .., m$ be the number of infected individuals in the i^{th} infected compartment at time t. Now, the epidemic model is:

$$\frac{\partial x_i}{\partial t} = F_i(x) - V_i(x)$$

Here, $V_i(x) = [V_i^-(x) - V_i^+(x)]$, where $V_i^+(x)$ represents the rate of transfer of individuals into compartment i from all other components containing individuals infected with the disease (here E, U, P and F) and where $V_i^-(x)$ represents the rate of transfer of individuals out of compartment i. Here, $F_i(x)$ represents the rate of appearance of new infections in compartment i. Let x_0 denote the disease free equilibrium. Now $\mathcal{F}$ and $\mathcal{V}$ are $m \times m$ matrices such that :

$$\mathcal{F}_{ij} = \frac{\partial F_i}{\partial x_j}(x_0) \qquad\qquad \mathcal{V}_{ij} = \frac{\partial V_i}{\partial x_j}(x_0)$$

Now, $\mathcal{F}\mathcal{V}^{-1}$ is called the **Next Generation Matrix**. The basic reproduction number R_0 is calculated by the spectral radius or the largest eigenvalue of $\mathcal{F}\mathcal{V}^{-1}$. For our case,

$$\mathcal{F} = \begin{pmatrix} \beta S(\alpha_p P + \alpha_u U + F) \\ 0 \\ 0 \\ 0 \end{pmatrix}, \qquad \mathcal{V} = \begin{pmatrix} \dfrac{E}{D_E} + \mu E \\ U\left(\dfrac{1}{\beta_1 D_r} + \delta_1\mu_c + \mu\right) - \left(\dfrac{1-r}{D_E}\right)E \\ P\left(\dfrac{1}{D_r} + \mu_c + \mu\right) - \left(\dfrac{r(1-f)}{D_E}\right)E \\ F\left(\dfrac{\beta_2}{D_r} + \dfrac{\mu_c}{\delta_2} + \mu\right) - \left(\dfrac{rf}{D_E}\right)E \end{pmatrix}$$

Now, we calculate the jacobian of $\mathcal{F}$ and $\mathcal{V}$ at the Disease Free Equilibrium (DFE).

$$\dot{\mathcal{F}} = \frac{\partial \mathcal{F}}{\partial X} = \begin{bmatrix} 0 & \beta\alpha_u S_0 & \beta\alpha_p S_0 & \beta S_0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

$$\dot{\mathcal{V}} = \frac{\partial \mathcal{V}}{\partial X} = \begin{bmatrix} \left(\dfrac{1}{D_E} + \mu\right) & 0 & 0 & 0 \\ -\left(\dfrac{1-r}{D_E}\right) & \left(\dfrac{1}{\beta_1 D_r} + \delta_1\mu_c + \mu\right) & 0 & 0 \\ -\left(\dfrac{r(1-f)}{D_E}\right) & 0 & \left(\dfrac{1}{D_r} + \mu_c + \mu\right) & 0 \\ -\left(\dfrac{rf}{D_E}\right) & 0 & 0 & \left(\dfrac{\beta_2}{D_r} + \dfrac{\mu_c}{\delta_2} + \mu\right) \end{bmatrix}$$

Now, we need to find the inverse of $\dot{\mathcal{V}}$. Since it is a lower triangular matrix, it is easy to find the inverse.

$$\dot{\mathcal{V}}^{-1} = \begin{bmatrix} \dfrac{1}{\left(\dfrac{1}{D_E} + \mu\right)} & 0 & 0 & 0 \\ \left(\dfrac{1-r}{D_E}\right)\dfrac{1}{\dfrac{1}{\beta_1 D_r} + \delta_1\mu_c + \mu} \cdot \dfrac{1}{\left(\dfrac{1}{D_E} + \mu\right)} & \dfrac{1}{\dfrac{1}{\beta_1 D_r} + \delta_1\mu_c + \mu} & 0 & 0 \\ \dfrac{r(1-f)}{D_E} \cdot \dfrac{1}{\dfrac{1}{D_r} + \mu_c + \mu} \cdot \dfrac{1}{\left(\dfrac{1}{D_E} + \mu\right)} & 0 & \dfrac{1}{\dfrac{1}{D_r} + \mu_c + \mu} & 0 \\ \left(\dfrac{rf}{D_E}\right)\dfrac{1}{\dfrac{\beta_1}{D_r} + \dfrac{\mu_c}{\delta_2} + \mu} \cdot \dfrac{1}{\left(\dfrac{1}{D_E} + \mu\right)} & 0 & 0 & \dfrac{1}{\dfrac{\beta_1}{D_r} + \dfrac{\mu_c}{\delta_2} + \mu} \end{bmatrix}$$

Now, we multiply $\dot{\mathcal{F}}$ and $\dot{\mathcal{V}}^{-1}$. The spectral radius of $\dot{\mathcal{F}}\dot{\mathcal{V}}^{-1}$ gives the basic reproduction number. Note that the matrix $\dot{\mathcal{F}}\dot{\mathcal{V}}^{-1}$ has only one non-zero row, which is the first one. All other rows of $\dot{\mathcal{F}}\dot{\mathcal{V}}^{-1}$ are 0. Hence, the spectral radius is given by $\left(\dot{\mathcal{F}}\dot{\mathcal{V}}^{-1}\right)_{11}$ (i.e., the $(1,1)^{th}$ element of $\dot{\mathcal{F}}\dot{\mathcal{V}}^{-1}$). So,

$$R_0 = \left(\dot{\mathcal{F}}\dot{\mathcal{V}}^{-1}\right)_{11} = \frac{\beta \cdot S_0}{\mu D_E + 1}\left(\frac{\alpha_u(1-r)}{\frac{1}{\beta_1 D_r} + \delta_1\mu_c + \mu} + \frac{\alpha_p r(1-f)}{\frac{1}{D_r} + \mu_c + \mu} + \frac{rf}{\frac{\beta_2}{D_r} + \frac{\mu_c}{\delta_2} + \mu}\right)$$

S.1.2 Special Cases

We develop an intuitive understanding of the above expression by studying some special cases of the SEIR-*fansy* model below:

S.1.2.1 Special case I: SIR model

In the SIR model, there are only 3 compartments: S (Susceptible), I (Infectious), and R (Removed). The death and recovered compartments are merged into one compartment called R (Removed). In the SIR model, there is only one infectious compartment I (which, in our model, is the P compartment), so we assume $r = 1$, $f = 0$ and $\alpha_p = 1$. As in the derivation of R_0 for the SIR model [13], we assume the birth (λ) and death (μ) rates to be zero. With these constraints and assumptions in place, our model reduces to SIR model, and we can simplify the expression in Equation S.1 as follows

$$R_0 = \frac{\beta_t S_0}{D_r^{-1} + \mu_c} = \frac{\beta_t S_0}{\nu} = \frac{\beta_t}{\nu},$$

where $S_0 = 1$, $\mu = 0$ is assumed and $\nu = D_r^{-1} + \mu_c$ specifies the removal rate. As such, we recover the well-known form of R_0 for the SIR model.

S.1.2.2 Special case II: SEIR model

Another special case is the popular SEIR model, where we have 4 compartments: S (Susceptible), E (Exposed), I (Infectious) and R (Removed). To obtain the expression of R_0 for SEIR model, we make all the assumptions as for SIR model except that D_E takes non-zero value $D_E = \frac{1}{k}$. We can also assume non-zero natural birth and death rates λ and μ.

Under these assumptions, the expression of R_0 in (S.1) becomes

$$R_0 = \frac{\beta_t S_0}{\mu D_E + 1} \cdot \frac{\alpha_p}{\nu + \mu} = \frac{k\beta_t S_0}{\mu + k} \cdot \frac{1}{\nu + \mu} = \frac{k\beta_t \lambda}{\mu(\mu + k)(\nu + \mu)}$$

This is the expression of R_0 for the SEIR model as derived in [13] as a special case.

S.1.3 Non-Instantaneous Testing

In the previous models we have assumed instantaneous testing. Without this assumptions (i.e, for $D_T \neq 0$), the differential equations would be as follows:

$$\begin{aligned}
\frac{\partial S}{\partial t} &= -\beta\frac{S(t)}{N}\Big(\alpha_P P(t) + \alpha_U U(t) + F(t) + T(t)\Big) + \lambda - \mu S(t) \\
\frac{\partial E}{\partial t} &= \beta\frac{S(t)}{N}\Big(\alpha_P P(t) + \alpha_U U(t) + F(t) + T(t)\Big) - \frac{E(t)}{D_E} - \mu E(t) \\
\frac{\partial U}{\partial t} &= \frac{(1-r)E(t)}{D_E} - \frac{U(t)}{\beta_1 D_r} - \delta_1\mu_c U(t) - \mu U(t) \\
\frac{\partial T}{\partial t} &= \frac{r\,E(t)}{D_E} - \frac{T(t)}{D_T} - \mu\,T(t) \\
\frac{\partial P}{\partial t} &= \frac{(1-f)\,T(t)}{D_T} - \frac{P(t)}{D_r} - \mu_c\,P(t) - \mu\,P(t)
\end{aligned}$$

$$
\begin{aligned}
\frac{\partial F}{\partial t} &= \frac{f\ T(t)}{D_T} - \frac{\beta_2\ F(t)}{D_r} - \frac{\mu_c\ F(t)}{\delta_2} - \mu\ F(t) \\
\frac{\partial RU}{\partial t} &= \frac{U(t)}{\beta_1 D_r} + \frac{\beta_2\ F(t)}{D_r} - \mu\ RU(t) \\
\frac{\partial RR}{\partial t} &= \frac{P(t)}{D_r} - \mu\ RR(t) \\
\frac{\partial DU}{\partial t} &= \delta_1 \mu_c U(t) + \frac{\mu_c\ F(t)}{\delta_2} \\
\frac{\partial DR}{\partial t} &= \mu_c P(t)
\end{aligned}
$$

S.1.4 Misclassification model - complete distributional assumptions

In the main paper, we have given the distribution of observed nodes given the other nodes and parameters. Here, we describe the distribution of the latent nodes also. After getting the estimates of the parameters using MCMC, we want to obtain model-based forecasts. In order to predict the future counts, we use the following multinomial random sampling strategy:

$$
\begin{aligned}
\zeta_{S\to E}, \zeta_{S\to O}, \zeta_{S\to S} &\sim \text{Multinomial}\ (S(t-1), p_{S\to E}, \mu, 1 - p_{S\to E} - \mu) \\
\zeta_{E\to U}, \zeta_{E\to P}, \zeta_{E\to F}, \zeta_{E\to O}, \zeta_{E\to E} &\sim \text{Multinomial}\ (E(t-1), \frac{(1-r)}{D_E}, \frac{r(1-f)}{D_E}, \frac{rf}{D_E}, \mu, \\
&\qquad 1 - p_{E\to U} - p_{E\to P} - p_{E\to F} - \mu) \\
\zeta_{U\to RU}, \zeta_{U\to DU}, \zeta_{U\to O}, \zeta_{U\to U} &\sim \text{Multinomial}\ (U(t-1), \beta_1^{-1} D_r^{-1}, \delta_1 \mu_c, \mu, 1 - \beta_1^{-1} D_r^{-1} - \delta_1 \mu_c - \mu) \\
\zeta_{P\to RR}, \zeta_{P\to DR}, \zeta_{P\to O}, \zeta_{P\to P} &\sim \text{Multinomial}\ (P(t-1), D_r^{-1}, \mu_c, \mu, 1 - D_r^{-1} - \mu_c - \mu) \\
\zeta_{F\to RU}, \zeta_{F\to DU}, \zeta_{F\to O}, \zeta_{F\to F} &\sim \text{Multinomial}\ (F(t-1), \beta_2 D_r^{-1}, \delta_2^{-1} \mu_c, \mu, 1 - \beta_2 D_r^{-1} - \delta_2^{-1} \mu_c - \mu) \\
\zeta_{RU\to O}, \zeta_{RU\to RU} &\sim \text{Multinomial}\ (RU(t-1), \mu, 1 - \mu) \\
\zeta_{RR\to O}, \zeta_{RR\to RR} &\sim \text{Multinomial}\ (RR(t-1), \mu, 1 - \mu)
\end{aligned}
$$

where $\zeta_{X\to Y}$ denotes the number of individuals moving from compartment X to compartment Y at time t. $\zeta_{X\to 0}$ denotes the number of individuals in compartment X that die at time t. The counts in each compartment at time t are given by,

$$
\begin{aligned}
S(t) &= \zeta_{S\to S} \\
E(t) &= \zeta_{E\to E} + \zeta_{S\to E} \\
U(t) &= \zeta_{U\to U} + UE \to U \\
P(t) &= \zeta_{P\to P} + \zeta_{E\to P} \\
F(t) &= \zeta_{F\to F} + \zeta_{E\to F} \\
RU(t) &= \zeta_{RU\to RU} + \zeta_{U\to RU} + \zeta_{F\to RU} \\
RR(t) &= \zeta_{RR\to RR} + \zeta_{P\to RR} \\
DU(t) &= \zeta_{DU\to DU} + \zeta_{U\to DU} + \zeta_{F\to DU} \\
DR(t) &= \zeta_{DR\to DR} + \zeta_{P\to DR}
\end{aligned}
$$

Given the parameters and the counts at time $(t-1)$, we obtain the predicted counts for time t. Using this approach, we obtain the posterior means of the future predicted counts at each of the 9 compartments using the MCMC estimated parameters. For the purpose of future prediction beyond the training period, we use the parameter estimates from the last time period. The estimation process is described in details below.

S.2 Extensions of the SEIR-fansy model

S.2.1 Extension 1. Time varying Case-Fatality Rate (mCFR)

Empirical analysis shows that the death rates and in turn the case-fatality rates are changing during the course of this pandemic between and within countries. The usual case fatality rate (CFR) is defined as:

$$\textbf{Case fatality rate (CFR)} = \frac{\text{Reported Cumulative Deaths}}{\text{Reported Cumulative Cases}}$$

The modified CFR or mCFR includes only the removed cases (deaths+recoveries) in the denominator as the outcomes are known only for this subset of individuals.

$$\textbf{Modified case fatality rate (mCFR)} = \frac{\text{Reported Cumulative deaths}}{\text{Reported Cumulative deaths} + \text{Reported Cumulative Recoveries}}$$

Figure S.1 shows that while countries like Belgium, USA, Italy, and Spain have very high mCFR, India and Russia

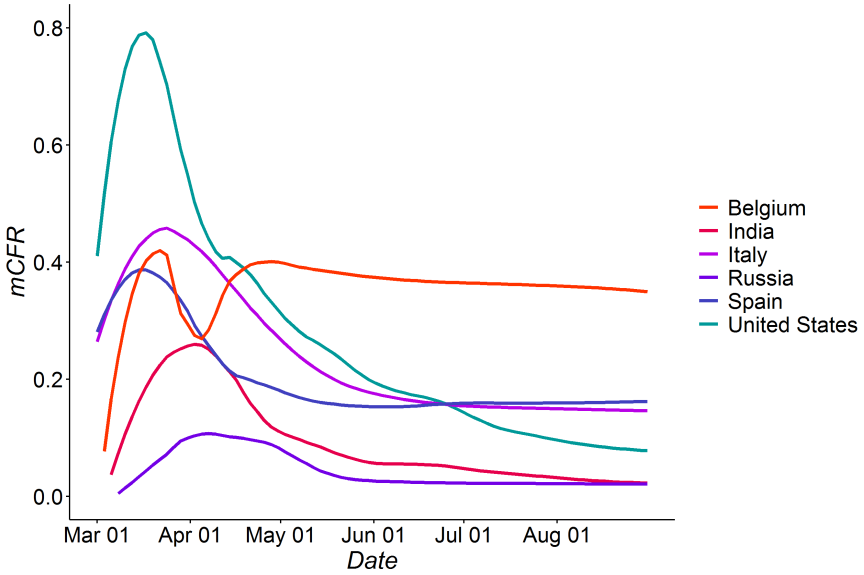

Figure S.1: Variation of mCFR with time: mCFR varies widely not only across countries but also within country for different time points. Each of the countries considered here exhibit an initial phase of high variation in mCFR followed by a decreasing trend which gradually stabilizes to some constant.

have comparatively much lower mCFR. We also note that initially most countries experienced a high mCFR, and mCFR gradually settled to a lower value as the case counts and recoveries rose. Hence, we hypothesize that modeling mCFR as a time varying quantity will improve the prediction of active cases and deaths.

Thus, we introduce a third time varying parameter called the mCFR along with β_t and r_t in the previous multinomial likelihood (Extensions after §2.3.4 of main manuscript). With this change, the differential equations in §2.2 of the main paper will remain the same. We use mCFR as opposed to CFR because in our model we use it to determine what is the probability that an infected person from node P moves to the death node (DR). The remaining will go to the recovered node RR. The new recovery rate will be $\frac{(1-mCFR)}{D_r}$, while the new death rate will be mCFR$\times\mu_c$.

S.2.2 Extension 2. Testing of infectious people based on symptoms

The problem with the base model is that we have implicitly assumed that the probability of a person being tested is equal for all infected individuals. However, that is certainly not the case in reality. The probability of being tested for a truly infected person largely depends on symptoms. On average, a person with severe symptoms will have the highest probability of being tested followed by the mild symptomatics and the asymptomatics. We extend the model (See Figure S.2) to account for symptom-dependent testing. We split the **Exposed(E)** compartment into three nodes: **Severe Symptomatic (Se)**, **Mild Symptomatic (Mi)** and **Asymptomatic (As)**.

We assume that individuals with severe symptoms will be tested with probability 1, while the mild and asymptomatic individuals will be tested with probability $t_{1,t}$ and $t_{2.t}$ at time t, respectively. We will also assume the probabilities of an infected person having severe, mild, or no symptoms are p_1, p_2, p_3, respectively. Under this setting the differential

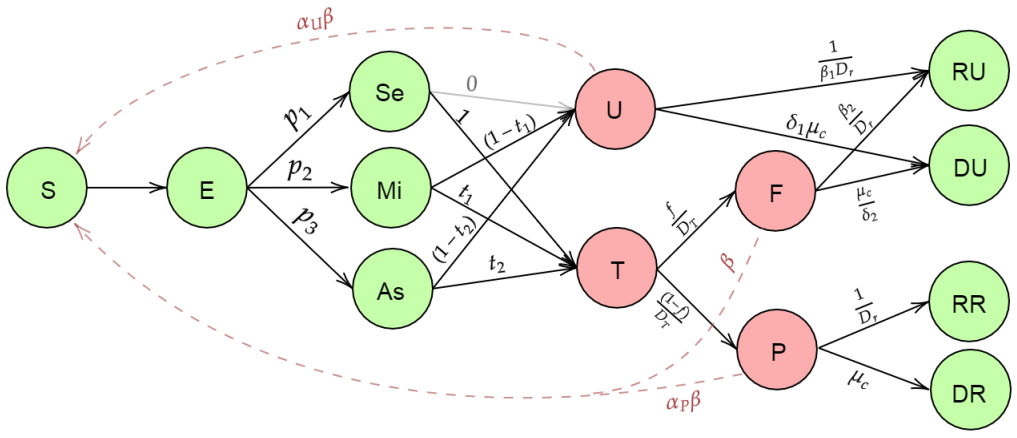

Figure S.2: Misclassification Model with Symptoms: In addition to the original compartments in the basic misclassification model, this model includes 3 other compartments based on symptoms: Se (severe symptomatic), Mi (mild symptomatic) and As (Asymptomatic). Each of these compartments are assumed to have different rates of being tested with testing rate increasing with severity of symptoms.

equations remain the same except the ones corresponding to the nodes P, U and F. The new set of differential equations corresponding to these three nodes are:

$$\begin{aligned}
\frac{\partial U}{\partial t} &= \frac{(p_2(1-t_{1,t}) + p_3(1-t_{2,t}))E(t)}{D_E} - \frac{U(t)}{\beta_1 D_r} - \delta_1\mu_c U(t) - \mu U(t) \\
\frac{\partial P}{\partial t} &= \frac{(p_1 + p_2 t_{1,t} + p_3 t_{2,t})(1-f)\,E(t)}{D_E} - \frac{P(t)}{D_r} - \mu_c\,P(t) - \mu\,P(t) \\
\frac{\partial F}{\partial t} &= \frac{f\,(p_1 + p_2 t_{1,t} + p_3 t_{2,t})\,E(t)}{D_E} - \frac{\beta_2\,F(t)}{D_r} - \frac{\mu_c\,F(t)}{\delta_2} - \mu\,F(t)
\end{aligned}$$

Due to identifiability issues, all the parameters described above cannot be estimated simultaneously. Therefore, We assume values for p_1, p_2 and p_3 are fixed based on existing data. We also assume that $t_{1,t} = kt_{2,t}$, where k is greater than 1 and known. This assumption implies that the probability of receiving a test for a person with mild symptoms is more than for a person with no symptoms. We run a sensitivity analysis for different values of k.

This model is nearly equivalent to the multinomial-2-parameter model described in Section 2.4.4. The only additional information that we are obtaining here is the allocation of tests conditional on symptoms. We essentially have expressed the probability of getting tested or the ascertainment rate r_t as the sum of three different probabilities by using the theorem of total probability. Namely,

$$r_t = (p_1 + p_2 t_{1,t} + p_3 t_{2,t}) \tag{S.2}$$
$$r_t = (p_1 + k p_2 t_{2,t} + p_3 t_{2,t}) \tag{S.3}$$

Our main parameters of interest are now reparameterized as β_t and $t_{2,t}$ instead of β_t and r_t. Since this model is a simple reparameterization of our original model we do not discuss the estimation again.

We shall refer to Extensions 1 and 2 as **Multinomial-3-parameter** and **Multinomial Symptoms** models, respectively.

S.2.3 Extension 3. Selection model: Who is getting tested?

So far we have been concerned with only the testing of truly infected individuals. However, respiratory or flu-like symptoms may manifest in an infected individual or an uninfected individual. The cause of symptoms (both mild and severe) in susceptible individuals may be due to other respiratory diseases such as influenza and the common cold.

It may be reasonable to assume that each individual, regardless of their underlying true disease status, has a probability of being tested that depends on the symptoms they have (or do not have). This probability could also depend on other covariates such as occupation or comorbidities. For simplicity, we will consider the case where testing is determined by symptoms only. Our goal is to extend the Multinomial-2-parameter model in §2.3.2 of main manuscript to directly incorporate the mechanism by which people in the population are tested.

To this end, we will assume individuals with severe symptoms are always tested provided sufficient tests are available. After all the individuals with severe symptoms are tested, the remaining tests are divided among those with mild symptoms and asymptomatics according to some given allocation rule that is independent of their true disease status given observed symptoms. We also assume that the number of tests to be performed in a given day does not depend on the true infection counts and is an external input. One advantage of using the number of tests as an input to the model is that we can study how the number of available tests influences the population infection rate in the long term. Figure S.3 provides a visualization of this expanded model.

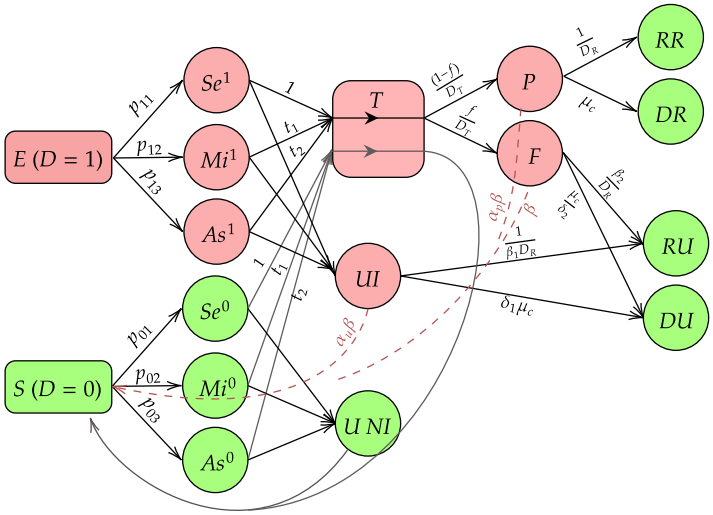

Figure S.3: Selection Model: This model assumes probability of being tested depends only on symptoms and not on the underlying disease status (D) which is unknown. As an example, probability of being tested is same for As^1 and As^0 which denote infected and uninfected asymptomatic individuals respectively. Following similar notations as in figure S.2, Se, Mi and As denote severe symptomatic, mild symptomatic and Asymptomatic individuals while the numbers in the superscript denote the underlying disease status with $D = 1$ and $D = 0$ denoting infected and uninfected individuals respectively.

There are eight new compartments in this model. Se^1, Mi^1 and As^1 consist of individuals who have developed severe, mild and no symptoms due to COVID-19 respectively. On the other hand, Se^0, Mi^0 and As^0 are comprised of individuals who have developed the similar degrees of symptoms but do not have COVID-19 infection. Their symptoms might be attributed to other diseases which exhibit similar symptoms as COVID-19 like influenza etc. Finally, UI (and UNI) compartment is formed of individuals who are untested with (without) an active COVID-19 infection.

The differential equations corresponding to this model have been provided in section S.2.3.1. Though conceptually appealing, this model has identifiability issues, and estimation requires substantial additional information. In particular, we need to know the mechanism by which people are tested, including the corresponding probabilities of testing. Additionally, we need to know the true symptom distributions for the exposed and susceptible people. It will be often quite hard to obtain this information as part of regularly-released data sources by countries across the world. Thus, implementation of this model may be wrinkled with too many subjective choices. However, we still think this is a

valuable formulation as it helps us to understand, analytically and intuitively, how selection bias can influence our estimates of interest.

Basic Reproduction number: We can calculate R_0 for this expanded model framework (derived using the Next Generation Matrix Method) as follows:

$$R_0 = \frac{S_0\beta_t}{\left(\mu + \frac{1}{D_E}\right)}\left(\frac{\alpha_u\left(\left(\frac{p_{12}+p_{13}}{D_E}\right) - \frac{T_0-(1-\mu)Se^0}{D_E(1-\mu)}\left(\frac{t_1p_{12}}{Mi^0}+\frac{t_2p_{13}}{As^0}\right)\right)}{\left(\mu+\delta_1\mu_c+\frac{1}{\beta_1 D_r}\right)} + \frac{\alpha_p\left(\frac{(1-f)p_{11}}{D_E} + \frac{(1-f)(T_0-(1-\mu)Se^0)}{D_E(1-\mu)}\left(\frac{t_1p_{12}}{Mi^0}+\frac{t_2p_{13}}{As^0}\right)\right)}{\left(\mu+\mu_c+\frac{1}{D_r}\right)} + \frac{\left(\frac{f\,p_{11}}{D_E} + \frac{f\,(T_0-(1-\mu)Se^0)}{D_E(1-\mu)}\left(\frac{t_1p_{12}}{Mi^0}+\frac{t_2p_{13}}{As^0}\right)\right)}{\left(\mu+\mu_c+\frac{1}{D_r}\right)}\right). \tag{S.4}$$

S.2.3.1 Differential Equations for the Selection Model

The differential equations for the selection model are as follows:

$$\begin{aligned}
\frac{\partial S}{\partial t} &= \frac{\partial Se^0}{\partial t} + \frac{\partial Mi^0}{\partial t} + \frac{\partial As^0}{\partial t}\\
\frac{\partial Se^0}{\partial t} &= \frac{-\beta Se^0(\alpha_p P(t) + \alpha_{ui}UI(t) + F(t))}{N} - \mu Se^0(t)\\
\frac{\partial Mi^0}{\partial t} &= \frac{-\beta Mi^0(\alpha_p P(t) + \alpha_{ui}UI(t) + F(t))}{N} - \mu Mi^0(t)\\
\frac{\partial As^0}{\partial t} &= \frac{-\beta As^0(\alpha_p P(t) + \alpha_{ui}UI(t) + F(t))}{N} - \mu As^0(t)\\
\frac{\partial E}{\partial t} &= \frac{-\beta S(\alpha_p P(t) + \alpha_{ui}UI(t) + F(t))}{N} - \frac{p_{11}E(t)}{D_e} - \frac{p_{12}E(t)}{D_e} - \frac{p_{13}E(t)}{D_e}\\
Se^1(t) &= \frac{p_{11}E(t)}{D_e} \qquad Mi^1(t) = \frac{p_{12}E(t)}{D_e} \qquad As^1(t) = \frac{p_{13}E(t)}{D_e}\\
\frac{\partial UI}{\partial t} &= Mi^1(t) - t_1(T - Se^1(t) - (Se^0(t) + \frac{\partial Se^0}{\partial t}))\left(\frac{Mi^1(t)}{Mi^1(t) + (Mi^0(t) + \frac{\partial Mi^0}{\partial t})}\right)\\
&\quad + As^1(t) - t_2(T - Se^1(t) - (Se^0(t) + \frac{\partial Se^0}{\partial t}))\left(\frac{As^1(t)}{As^1(t) + (As^0(t) + \frac{\partial As^0}{\partial t})}\right)\\
&\quad - \mu UI(t) - \delta_1\mu_c UI(t) - \frac{UI(t)}{\beta_1 D_r}\\
\frac{\partial P}{\partial t} &= (1-f)(t_1(T - Se^1(t) - (Se^0(t) + \frac{\partial Se^0}{\partial t}))\left(\frac{Mi^1(t)}{Mi^1(t) + (Mi^0(t) + \frac{\partial Mi^0}{\partial t})}\right)\\
&\quad + (1-f)(t_2(T - Se^1(t) - (Se^0(t) + \frac{\partial Se^0}{\partial t}))\left(\frac{As^1(t)}{As^1(t) + (As^0(t) + \frac{\partial As^0}{\partial t})}\right)\\
&\quad + (1-f)Se^1(t) - \mu P(t) - \mu_c P(t) - \frac{P(t)}{D_r}
\end{aligned} \tag{S.5}$$

$$\frac{\partial F}{\partial t} = f(t_1(T - Se^1(t) - (Se^0(t) + \frac{\partial Se^0}{\partial t})) \left(\frac{Mi^1(t)}{Mi^1(t) + (Mi^0(t) + \frac{\partial Mi^0}{\partial t})} \right)$$
$$+ f(t_2(T - Se^1(t) - (Se^0(t) + \frac{\partial Se^0}{\partial t})) \left(\frac{As^1(t)}{As^1(t) + (As^0(t) + \frac{\partial As^0}{\partial t})} \right)$$
$$+ fSe^1(t) - \mu F(t) - \frac{\mu_c F(t)}{\delta_2} - \frac{\beta_2 F(t)}{D_r}$$
$$\frac{\partial RU}{\partial t} = \frac{UI(t)}{\beta_1 D_r} + \frac{\beta_2\, F(t)}{D_r} - \mu\, x_{R_U} \qquad \frac{\partial RR}{\partial t} = \frac{P(t)}{D_r} - \mu\, R_R$$
$$\frac{\partial DU}{\partial t} = \delta_1 \mu_c UI(t) + \frac{\mu_c\, F(t)}{\delta_2} \qquad \frac{\partial DR}{\partial t} = \mu_c P(t)$$

S.2.3.2 Selection Model : Complete Distributional Assumptions

To generate data using the test model, we perform the following steps.

$$\zeta_{S\to E}, \zeta_{S\to O}, \zeta_{S\to S} \sim Multinomial(S(t-1), p_{S\to E}, \mu, 1 - p_{S\to E} - \mu)$$

Now, we assume the probability of an individual being severely symptomatic, mildly symptomatic or asymptomatic given he/she is susceptible is given by the probability vector $\mathbf{p_0} = (p_{01}, p_{02}, p_{03})$. The probability for an infected individual is given by $\mathbf{p_1} = (p_{11}, p_{12}, p_{13})$. To obtain the number of individuals in the groups $Se^0, Mi^0, and As^0$, we assume that the outgoing individuals from the susceptible group follow the distribution given by $\mathbf{p_0}$.

$$Se^0_{new}(t),\ Mi^0_{new}(t),\ As^0_{new}(t) \sim Multinomial(\zeta_{S\to S}, \mathbf{p_0})$$

Now, from our assumption that the individuals in E follow the distribution given by $\mathbf{p_1}$, we can write,

$$\zeta_{E\to Se^1}, \zeta_{E\to Mi^1}, \zeta_{E\to As^1}, \zeta_{E\to O}, \zeta_{E\to E} \sim Multinomial\left(E(t-1), \left(\frac{\mathbf{p_1}}{D_E}, \mu, 1 - \frac{1}{D_E} - \mu\right)\right)$$

Recall, we assume that all individuals with severe symptoms are tested provided adequate tests are available. This implies

$$Se^0_{tested} = Se^0(t) \qquad Se^1_{tested} = Se^1(t) \qquad Se_{tested} = Se^0_{tested} + Se^1_{tested}$$

In the case when number of test $T(t)$ is less than that of severe individuals, we assume that the number of tested Se^1 and Se^0 individuals is proportional to their respective counts.

$$Se^0_{tested}, Se^1_{tested} \sim Multinomial\left(Se(t), \left(\frac{Se^0(t)}{Se^0(t) + Se^1(t)}, \frac{Se^1(t)}{Se^0(t) + Se^1(t)}\right)\right)$$

If the total number of remaining tests is greater than or equal to the number of mild and asymptomatic individuals, then all of them are tested i.e :

$$Mi^0_{tested} = Mi^0(t),\ Mi^1_{tested} = Mi^1(t),\ As^0_{tested} = As^0(t),\ As^1_{tested} = As^1(t)$$

If number of tests are not adequate for all the mild symptomatic and asymptomatic people to be tested, then the remaining tests (after testing the severe symptomatic people) are distributed among the mildly symptomatic and asymptomatic individuals in the ratio $t_1 : t_2$.

$$Mi_{tested}, As_{tested} \sim Binomial\left(T - Se_{tested}, (t_1, t_2)\right)$$

As we did in the case of severely symptomatic, we allocate the tests among infected and uninfected mildly symptomatic (and also asymptomatic) individuals randomly.

$$Mi^0_{tested}, Mi^1_{tested} \sim Binomial\left(Mi_{tested}, \left(\frac{Mi^0(t)}{Mi^0(t) + Mi^1(t)}, \frac{Mi^1(t)}{Mi^0(t) + Mi^1(t)}\right)\right)$$
$$As^0_{tested}, As^1_{tested} \sim Binomial\left(As_{tested}, \left(\frac{As^0(t)}{As^0(t) + As^1(t)}, \frac{As^1(t)}{As^0(t) + As^1(t)}\right)\right)$$

$$\begin{aligned}
\zeta_{UI\to RU}, \zeta_{UI\to DU}, \zeta_{UI\to O}, \zeta_{UI\to UI} &\sim Multinomial(UI(t-1), \beta_1^{-1} D_r^{-1}, \delta_1 \mu_c, \mu, \\
&\qquad 1 - \beta_1^{-1} D_r^{-1} - \delta_1 \mu_c - \mu) \\
\zeta_{P\to RR}, \zeta_{P\to DR}, \zeta_{P\to O}, \zeta_{P\to P} &\sim Multinomial(P(t-1), D_r^{-1}, \mu_c, \mu, 1 - D_r^{-1} - \mu_c - \mu) \\
\zeta_{F\to RU}, \zeta_{F\to DU}, \zeta_{F\to O}, \zeta_{F\to F} &\sim Multinomial(F(t-1), \beta_2 D_r^{-1}, \delta_2^{-1} \mu_c, \mu, \\
&\qquad 1 - \beta_2 D_r^{-1} - \delta_2^{-1} \mu_c - \mu) \\
\zeta_{RU\to O}, \zeta_{RU\to RU} &\sim Multinomial(RU(t-1), \mu, 1-\mu) \\
\zeta_{RR\to O}, \zeta_{RR\to RR} &\sim Multinomial(RR(t-1), \mu, 1-\mu)
\end{aligned}$$

We also assume the false negative probability = f. The numbers of new individuals to P and F states are given by :

$$P_{new}, F_{new} \sim Multinomial\Big(Se^1_{tested} + Mi^1_{tested} + As^1_{tested}, (1-f, f)\Big)$$

Finally we write the number of people in each state at time t as follows :

$$\begin{aligned}
UI(t) &= \zeta_{UI\to UI} + Se^1_{untested} + Mi^1_{untested} + As^1_{untested} \\
P(t) &= \zeta_{P\to P} + P_{new} \\
F(t) &= \zeta_{F\to F} + F_{new} \\
RU(t) &= \zeta_{RU\to RU} + \zeta_{UI\to RU} + \zeta_{F\to RU} \\
RR(t) &= \zeta_{RR\to RR} + \zeta_{P\to RR} \\
DU(t) &= \zeta_{DU\to DU} + \zeta_{UI\to DU} + \zeta_{F\to DU} \\
DR(t) &= \zeta_{DR\to DR} + \zeta_{P\to DR}
\end{aligned}$$

S.2.3.3 Special case of Selection Model : Uniform testing

To understand the effect of selection bias on R_0, we consider a special case of the Selection model where we assume uniform testing. Here, uniform testing means tests are offered independently of symptoms. The model is represented diagrammatically in Figure (S.4).

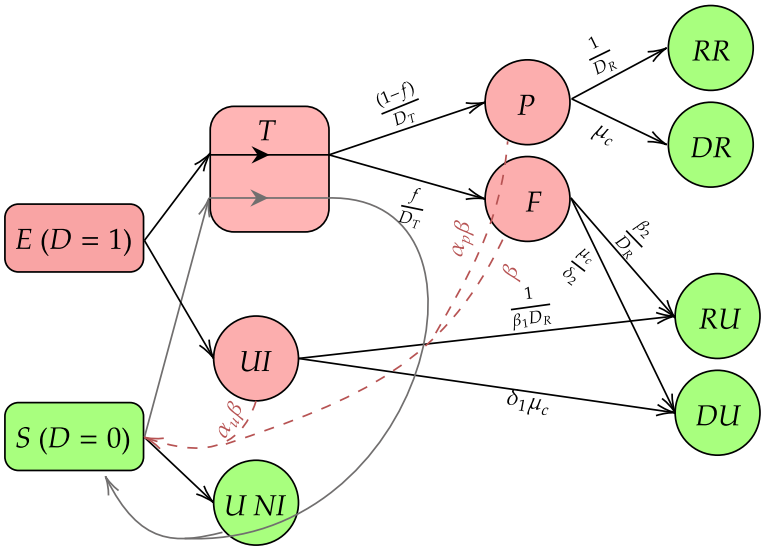

Figure S.4: Special case of Selection Model : Uniform testing. This is a special case of the selection model where we assume probability of being tested is same for Se, Mi or As compartments (irrespective of true disease status). This simplifies the selection model and in fact, under this assumption, the selection model is equivalent to our misclassification model.

The transmission dynamics of this model are very similar to the Selection model. We provide the differential equations describing the dynamics of the nodes S, E, UI, P and F. The rest of the nodes (RU, RR, DU and DR) have differential

equations exactly same as in Selection Model.

$$\begin{aligned}
\frac{\partial S}{\partial t} &= -\beta\frac{S(t)}{N}\Big(\alpha_P P(t) + \alpha_U U(t) + F(t)\Big) + \lambda - \mu S(t)\\
\frac{\partial E}{\partial t} &= \beta\frac{S(t)}{N}\Big(\alpha_P P(t) + \alpha_U U(t) + F(t)\Big) - \frac{E(t)}{D_E} - \mu E(t)\\
\frac{\partial UI}{\partial t} &= \frac{E(t)}{D_E} - T\frac{E(t)}{S(t)+E(t)} - \frac{UI(t)}{\beta_1 D_r} - \delta_1\mu_c UI(t) - \mu UI(t)\\
\frac{\partial P}{\partial t} &= (1-f)T\frac{E(t)}{S(t)+E(t)} - \frac{P(t)}{D_r} - \mu_c\, P(t) - \mu\, P(t)\\
\frac{\partial F}{\partial t} &= f\cdot T\frac{E(t)}{S(t)+E(t)} - \frac{\beta_2\, F(t)}{D_r} - \frac{\mu_c\, F(t)}{\delta_2} - \mu\, F(t)
\end{aligned}$$

Now that we have all the differential equations governing the dynamics of this model, we calculate the basic reproduction number using Next Generation Matrix method [13]. Using calculations similar to what we did for the Misclassification model, we arrive at the following expression of R_0 for Uniform testing model.

$$R_0 = \frac{\beta}{\mu + \frac{1}{D_E}}\left[\frac{\alpha_u\left(\frac{1}{D_E} - T\right)}{\frac{1}{\beta_1 D_r} - \delta_1\mu_c - \mu} + \frac{\alpha_p\,(1-f)\,T}{\frac{1}{D_r} - \mu_c - \mu} + \frac{fT}{\frac{\beta_2}{D_r} - \frac{\mu_c}{\delta_2} - \mu}\right] \tag{S.6}$$

S.2.3.4 Decoupling selection and misclassification:

To capture the effect of selection bias and misclassification on R_0, we consider an example setting where we first compute the value of R_0 (see equation S.7) with $f = 0$ (no misclassification, selection), $f = 0.3$ (misclassification+selection) using the same set of parameters. To isolate the effect of selection, we consider a model where selection is random (for further details refer to section S.2.3.3 of supplementary materials) and evaluate R_0 when $f = 0$ (no misclassification or selection) and $f = 0.3$ (only misclassification, no selection).

Data Generation: We consider a hypothetical population of 1 million people. We set $\beta = 0.25$, $r_t = 0.1$, $p_0 = (10^{-6}, 10^{-5}, 1 - 10^{-6} - 10^{-5})$ $p_1 = (0.02, 0.18, 0.8)$, $t_{1,t} = 0.7$ and $t_{2,t} = 0.3$. We consider three different values of the number of tests per thousand population $(0.1, 0.5, 1.0, 2.0)$ consistent with values observed in different countries across the world. Remaining parameters are same as in §S.4.

Results: Table S.1 presents values of R_0 for the four configurations. From table (S.1), we conclude that under random

Number of tests per thousand	Model			
	No Selection Bias		Selection Bias	
	f = 0	f = 0.3	f = 0	f = 0.3
0.1	1.54	1.53	1.64	2.09
0.5	1.54	1.54	2.02	4.08
1.0	1.54	1.54	2.50	6.56
2.0	1.54	1.55	3.45	11.54

Table S.1: Effect of misclassification and selection bias on R_0: The table lists values of basic reproduction number R_0 calculated under different values of false negative rate f and under presence or absence of selection bias. As is evident from the table, the value of R_0 is not sensitive to the value of f when selection bias is absent. However, under the presence of selection bias, the value of R_0 is sensitive to different values of sensitivity of the test. This phenomenon is further supported by the simulation studies in sections 4.1 and 4.2 in the main paper.

selection, R_0 is not sensitive to the total number of tests and false negatives with values remaining around 1.54, the true value. The values in the 3$^{\text{rd}}$ column are inflated, which indicates a substantial effect of selection bias on R_0, especially when the number of tests are large, even when the tests are perfect. The fourth column underpins the key issue that the R_0 can be very far from the true value with both selection bias and misclassification, especially with large number of tests being distributed in a non-probabilistic way.

S.3 Real Data Analysis for India

S.3.1 Initial values and parameter setting:

Using observed counts on April 1, 2020 (for wave 1) and February 1, 2021 (for wave 2), we fix the values P_0, RR_0 and DR_0 for the three compartments for which data are reported. The counts in the unobserved compartments are set proportionately to the observed ones, with $E_0 = 3(U_0 + P_0 + F_0)$. The false negative rate is fixed at $f = 0.15$ [9], and the initial value for the ascertainment rate r_t is set to 0.15. We assume all the parameters in our model except r_t and β_t remain constant through the entire course of the disease. We set the latency period $D_E = 5.2$ days and assume that the latency period is equal to the incubation period [10]. We assume $D_r = 17.8$ days following the report by WHO and set this as the average time till death for deceased COVID patients. We set $\mu_c = \text{mCFR}/17.8$ where mCFR is as defined in section S.2.1. The natural birth and death rates are assumed to be equal. According to the worldbank report, the average life span of Indians is approximately 69.4 years. So we take $\lambda = \mu = 1/(69.416 \cdot 365)$..The various scaling factors are fixed: $\alpha_p = 0.5,\ \alpha_u = 0.5,\ \beta_1 = 0.6,\ \beta_2 = 0.7,\ \delta_1 = 0.3,\ \delta_2 = 0.7$. With these values and the sub-intervals described in Table S.4, we estimate R_0 for each sub-interval. Additionally, we predict COVID-counts in the test period for wave 2 using the estimated parameter values based on the training set.

For India, we have fitted data from April 1, 2020 to January 31, 2021 for wave 1 and February 1 to August 31, 2021 . So for our prediction, we need the counts of the different compartments on the initial date, that is on April 1, 2020 and February 1, 2021 for waves 1 and 2 respectively. The tables S.2 and S.3 present the counts of the compartments for India on the start dates of waves 1 and 2 respectively.

Variable	Value	Justification
S(0)	1340940853	N-(E(0) + U(0) + P(0) + F(0) + RU(0) + RR(0) + DU(0) + DR(0)) (N = 1341 million)
E(0)	43221	Thrice the number of current infected
U(0)	12246	$\frac{1-r}{r}$ (P(0) + F(0))
P(0)	1837	Reported current infected on 1st April, 2020
F(0)	324	$\frac{f}{1-f}$ P(0)
RU(0)	987	$\left(\frac{1-r}{r} + \frac{f}{1-f}\right)$ RR(0)
RR(0)	169	Reported recovered on 1st April, 2020
DU(0)	310	$\left(\frac{1-r}{r} + \frac{f}{1-f}\right)$ DR(0)
DR(0)	53	Reported deceased on 1st April, 2020

Table S.2: Initial Values of the Different Compartments in Wave 1. We have taken April 1, 2020 to January 31, 2021 as the first wave training period and in the first wave, we have not taken any testing period. The date on which these initial counts are reported is April 1,2020.

S.3.2 Performance of different models for India

To asses the performance of different models described earlier, we check the estimates of R_0 as well as the accuracy of estimates of number of cases and deaths.

First, we look at the estimates of R_0. From table S.6, we observe that the estimates of R_0 from different models are qualitatively similar with numerical differences.

Now, we look at the prediction accuracies of the different models in estimating case and death counts. Noting that the number of cases and deaths varies widely across different time points (e.g., the number of reported cumulative cases go from below 5000 on April to 31 million in August, 2021), we use a scale-independent metric, **mean squared relative**

Variable	Value	Justification
S(0)	1263249237	N-(E(0) + U(0) + P(0) + F(0) + RU(0) + RR(0) + DU(0) + DR(0)) (N = 1341 million)
E(0)	3903900	Thrice the number of current infected
U(0)	1106105	$\frac{1-r}{r}$ (P(0) + F(0))
P(0)	165916	Reported current infected on 1st Feb,2021
F(0)	29279	$\frac{f}{1-f}$ P(0)
RU(0)	61044909	$\left(\frac{1-r}{r} + \frac{f}{1-f}\right)$ RR(0)
RR(0)	10447283	Reported recovered on 1st Feb, 2021
DU(0)	899440	$\left(\frac{1-r}{r} + \frac{f}{1-f}\right)$ DR(0)
DR(0)	153931	Reported deceased on 1st Feb, 2021

Table S.3: Initial Values of the Different Compartments in Wave 2. In the wave 2, we have taken Feb 1,2021 to June 30,2021 as the training period, while, July 1,2021 to August 31,2021 was taken to be the test period. The date on which these iniital counts are reported is Feb 1,2021.

prediction error (MRPE) or relative mean square error or RMSE [3], defined as follows:

$$MRPE = \frac{1}{n}\sum_{i=1}^{n}\left(1 - \frac{\hat{v}_i}{v_i}\right)^2$$

for observed data $v = (v_1, v_2, ..., v_n)$ and predicted vector $\hat{v} = (\hat{v}_1, \hat{v}_2, ..., \hat{v}_n)$. We present MRPE (multiplied by 10) for each of the models evaluated both on the training and test set. Table S.5 presents the values of MRPE for reported cumulative cases, deaths and active cases for all the five models. The column-wise minima are indicated in bold letters.

The Poisson and binomial models perform very similarly. This is expected, since the binomial likelihood approaches the Poisson likelihood for large case-counts. We note that the multinomial models are doing better in test data in predicting reported cumulative cases. Overall, the multinomial-2-parameter model does well when predicting reported cumulative cases, deaths and active cases. Specifically, the multinomial-2-parameter model predicts reported active cases better than the other models considered.

Model	Reported Cumulative Cases		Reported Deaths		Reported Active Cases	
	Train	Test	Train	Test	Train	Test
Poisson	**0.07**	0.10	0.95	1.24	**0.36**	0.75
Binomial	0.07	0.10	0.96	**1.22**	0.36	0.74
Multinomial-2-parameter	0.21	0.09	1.25	1.27	0.57	**0.71**
Multinomial-3-parameter	0.13	**0.02**	**0.40**	1.52	0.45	1.10
Multinomial symptoms	0.12	0.02	1.19	1.87	0.44	1.23

Table S.5: MRPE (multiplied by factor of ten) for reported cumulative cases, deaths and active cases for different models on train and test dataset. The lowest values in each column are written in **bold** characters. As can be seen from the table, Poisson and Binomial models perform very similarly while multinomial models outperform Poisson and Binomial models in test data when predicting case counts.

Table S.5 presents the values of MRPE for reported cumulative cases, deaths and active cases for all the five models. The column-wise minima are indicated in bold letters. The Poisson and binomial models perform very similarly. This is expected, since the binomial likelihood approaches the Poisson likelihood for large case-counts. We note that the

Phase	From	To	Wave
Lockdown 1	1st April	14th April	1
Lockdown 2	15th April	3rd May	1
Lockdown 3	4th May	17th May	1
Lockdown 4	18th May	31st May	1
Unlock 1.0	1st June	30th June	1
Unlock 2.0	1st July	31st July	1
Unlock 3.0	1st August	31st August	1
Unlock 4.0	1st September	30th September	1
Unlock 5.0	1st October	31st October	1
Unlock 6.0	1st November	30th November	1
Unlock 7.0	1st December	31st December	1
Unlock 8.0	1st January	31st January	1
Unlock 9.0	1st February	28th February	2
Unlock 10.0	1st March	31st March	2
Unlock 11.0	1st April	30th April	2
Unlock 12.0	1st May	31st May	2
Unlock 13.0	1st June	30th June	2
Unlock 14.0	1st July	31st July	2
Unlock 15.0	1st August	31st August	2

Table S.4: Phases of public health interventions in India from 1st April, 2020 to 1st August, 2021 which is the time period considered for data analysis in this paper. The last column indicates if a particular time period has been considered as wave 1 or 2. The period from 1st April, 2020 to 31st January, 2021 is considered as wave 1 while 1st February, 2021 to 31st August, 2021 is considered as wave 2.

multinomial models are doing better in test data in predicting reported cumulative cases. Overall, the Multinomial-2-parameter model does well when predicting reported cumulative cases, deaths and active cases. Specifically, the Multinomial-2-parameter model predicts reported active cases better than the other models considered. Figure S.5 provides the daily trajectories for predicted active cases and deaths from April 1, 2020 to August 31, 2020 for the different models under consideration.

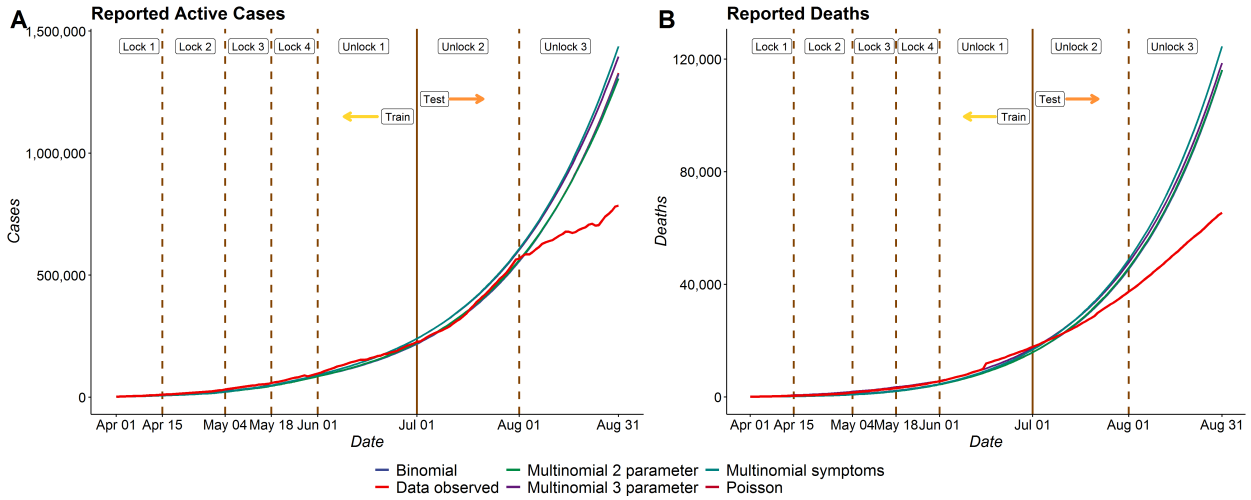

Figure S.5: Reported Active Cases in India - Comparison between different models: We consider 5 different models here, namely, Binomial, Poisson, Multinomial-2-parameter, Multinomial-3-parameter and Multinomial symptoms models. We observe that the prediction of all the models are quite similar and all of them are quite accurate for the training period. However, in the test period, all of the models suffer due to the sudden decrease in cases during Unlock 3, which none of them could predict.

=Model	Basic Reproduction Number				
	1-14 Apr	15 Apr-3 May	4-17 May	18-31 May	1-30 Jun
Poisson	3.36 [3.29, 3.43]	2.24 [2.21, 2.27]	1.69 [1.66, 1.72]	1.76 [1.73, 1.79]	1.76 [1.62, 1.64]
Binomial	3.42 [3.35, 3.5]	2.25 [2.22, 2.28]	1.69 [1.66, 1.72]	1.76 [1.73, 1.79]	1.63 [1.62, 1.64]
Multinomial-2-parameter	3.74 [3.62, 3.85]	2.36 [2.31, 2.4]	1.75 [1.72, 1.78]	1.70 [1.67, 1.72]	1.61 [1.61, 1.62]
Multinomial-3-parameter	3.22 [3.15, 3.29]	2.25 [2.21, 2.29]	1.75 [1.72, 1.78]	1.75 [1.72, 1.78]	1.73 [1.72, 1.74]
Multinomial symptoms	3.30 [3.23, 3.37]	2.18 [2.14, 2.21]	1.64 [1.61, 1.66]	1.59 [1.57, 1.61]	1.52 [1.51, 1.53]

Table S.6: Estimates of R_0 for 5 different time periods in India by different models. All the models give quite similar estimates. For each of the 5 models, there is an overall decreasing trend in the estimates of R_0 as we move forward in time periods. This shows the effectiveness of the lockdown and public health measures implemented in India during the beginning of wave 1.

S.3.3 Confidence intervals for different compartments for India

We have done our estimation using the MCMC Metropolis Method and predicted the counts for the different compartments by using the posterior means conditional on the estimated parameters. So the large number of iterations of MCMC provide a 95% credible interval for the parameters as well as for the predictions of the compartments. So the following figure shows the credible regions for the Reported Active, False negative active and Untested Active cases.

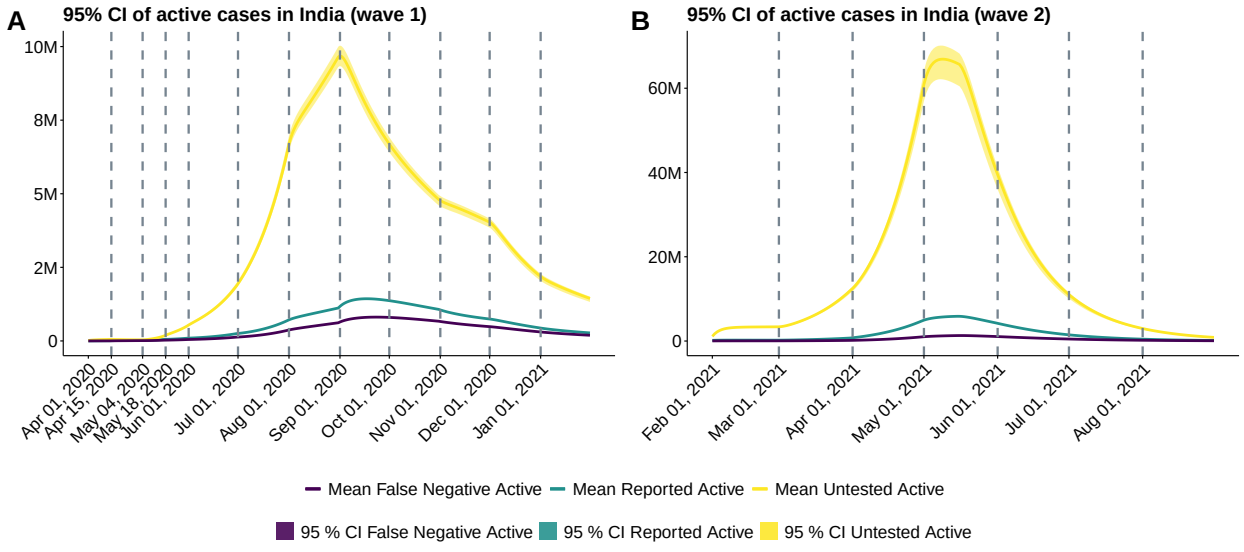

Figure S.6: 95% Credible Intervals of estimates of Current Active Cases in India for (A) wave 1 and (B) wave 2. We see that the credible intervals of reported active cases is much narrower than that of untested active cases and false negative cases which is expected as the former is observed, which reduces the variance of the estimates while the other 2, being unobserved, have high variance.

Subfigure (A) of figure (S.6), shows the 95% CI's of the estimates of Current Active cases in India from 1st April, 2020 to 31st January, 2021 and subfigure (B) shows the same from 1st February, 2021 to 31st August, 2021.

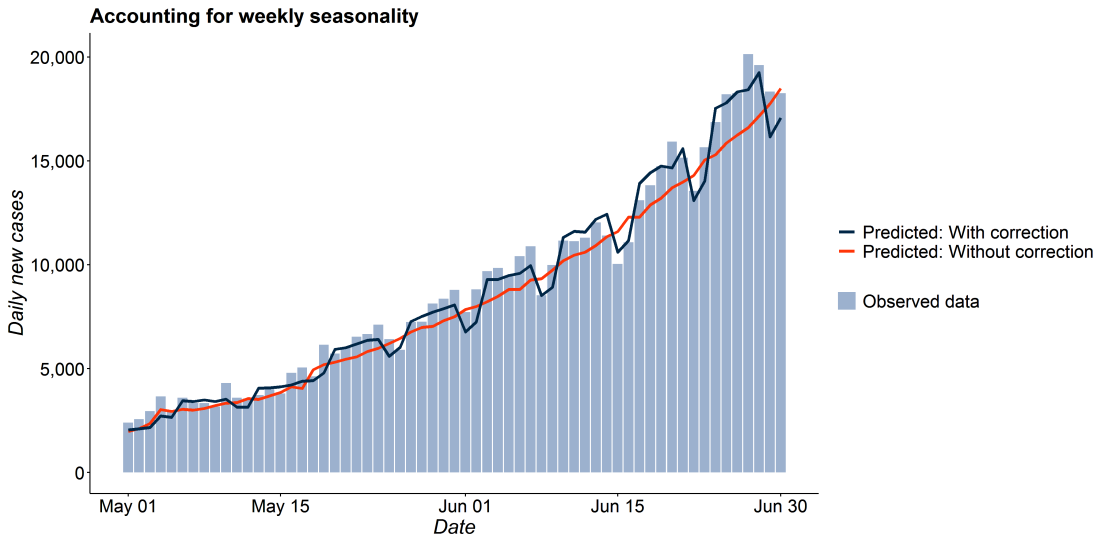

Figure S.7: Daily Fluctuations in May and June. The red line denotes our usual predictions (using base Multinomial-2-parameter model without day-of-the-week correction) and the blue line denotes predictions with the correction. One can clearly see how the correction makes the prediction much more accurate.

S.3.4 Accounting for weekly seasonality

From figure S.7, we can easily observe a seasonal trend in the observed number of daily confirmed cases. More specifically a weekly trend is visible, where the number of cases for Monday and Tuesday is a bit lower than the other five days of the week. So, in this case, it might not be reasonable to assume that the outgoing rate from the Exposed Node to the Infectious Nodes is same for all the days in the week. So we have assumed that the rate for these two days (Monday and Tuesday) is lower than that of other days of the week. This is achieved by multiplying the rate for Monday and Tuesday by a constant ω_1 (to be estimated by MCMC as a third parameter), and for the remaining 5 days, it is multiplied by a constant ω_2. But we have assumed that the mean outgoing rate from the Exposed Node to the Infectious Nodes is $\frac{1}{D_E}$ over a week. So we can readily get the relationship between ω_1 and ω_2 which is given by $\omega_2 = \frac{7-2\omega_1}{5}$. Now using the new adjustments, there will be changes in the following 4 differential equations and the remaining D.E. will remain exactly same.

$$
\begin{aligned}
\frac{\partial E}{\partial t} &= \frac{\beta_t S(t)}{N}\Big(\alpha_P P(t) + \alpha_U U(t) + F(t)\Big) - \frac{E(t) * \omega(t)}{D_E} - \mu E(t) \\
\frac{\partial U}{\partial t} &= \frac{(1-r_t)E(t) * \omega(t)}{D_E} - \frac{U(t)}{\beta_1 D_r} - \delta_1 \mu_c U(t) - \mu U(t) \\
\frac{\partial P}{\partial t} &= \frac{(1-f) r_t\, E(t) * \omega(t)}{D_E} - \frac{P(t)}{D_r} - \mu_c\, P(t) - \mu\, P(t) \\
\frac{\partial F}{\partial t} &= \frac{f\, r_t\, E(t) * \omega(t)}{D_E} - \frac{\beta_2\, F(t)}{D_r} - \frac{\mu_c\, F(t)}{\delta_2} - \mu\, F(t)
\end{aligned}
$$

where $\omega(t) = \omega_1$ if the day is Monday or Tuesday and $\omega(t) = \omega_2$ if the day is otherwise. So basically we need to estimate an additional parameter ω_1 other than β_t and r_t. Now we have assumed that ω_1 is not time-varying, it is constant over time to reduce the complexity of the model. Now we can easily see from the Figure S.7 that using this correction, the model has been able to capture the true trend of the daily confirmed cases more accurately than that of without correction. The prediction has been become more accurate with this adjustment. This is also evident from the MRPE of the daily confirmed cases in both cases. The MRPE with correction is 0.009, while the MRPE without correction is 0.012. The estimate for ω_1 came out to be 0.82 and therefore ω_2 is equal to 1.07. Actually the ratio of ω_1 and ω_2 denotes the actual scaling factor by which the daily positive counts on Monday and Tuesday are dipping down compared to the remaining 5 days and this ratio came out to be 0.76. So, with this correction, we have not only been able to improve the prediction accuracy but also estimate that there is an approximately 76% lower reporting of daily new cases on Monday and Tuesday than other days of the week. It is easy to see that this method is also applicable for

other countries like USA where reported cases tend to be lower on weekends than on weekdays. However, one must assume beforehand which days have comparatively lower rates of reporting. If such information is unavailable, one may try to incorporate more than 2 parameters all of which can vary freely or resort to other time series techniques for isolating the seasonal (weekly) component.

S.4 Results for Delhi and Mumbai

There is tremendous heterogeneity in the virus curves across time in India. We focus on two of the worst-hit places in India - Delhi and Mumbai.

The city of Mumbai has nearly 0.7 million reported cumulative cases by 1st August, 2021 while Delhi had over 1.4 million cases by the same date.

We estimate the basic reproduction number for Mumbai and Delhi for both wave 1 and wave 2. The periods for wave 1 and wave 2 remain same as before. Wave 1 is defined from April 1, 2020 to January 31, 2021 and wave 2 is defined as February 1, 2021 to June 30, 2021. We also provide a 1 month prediction of reported active cases for July, 2021 for both Delhi and Mumbai.

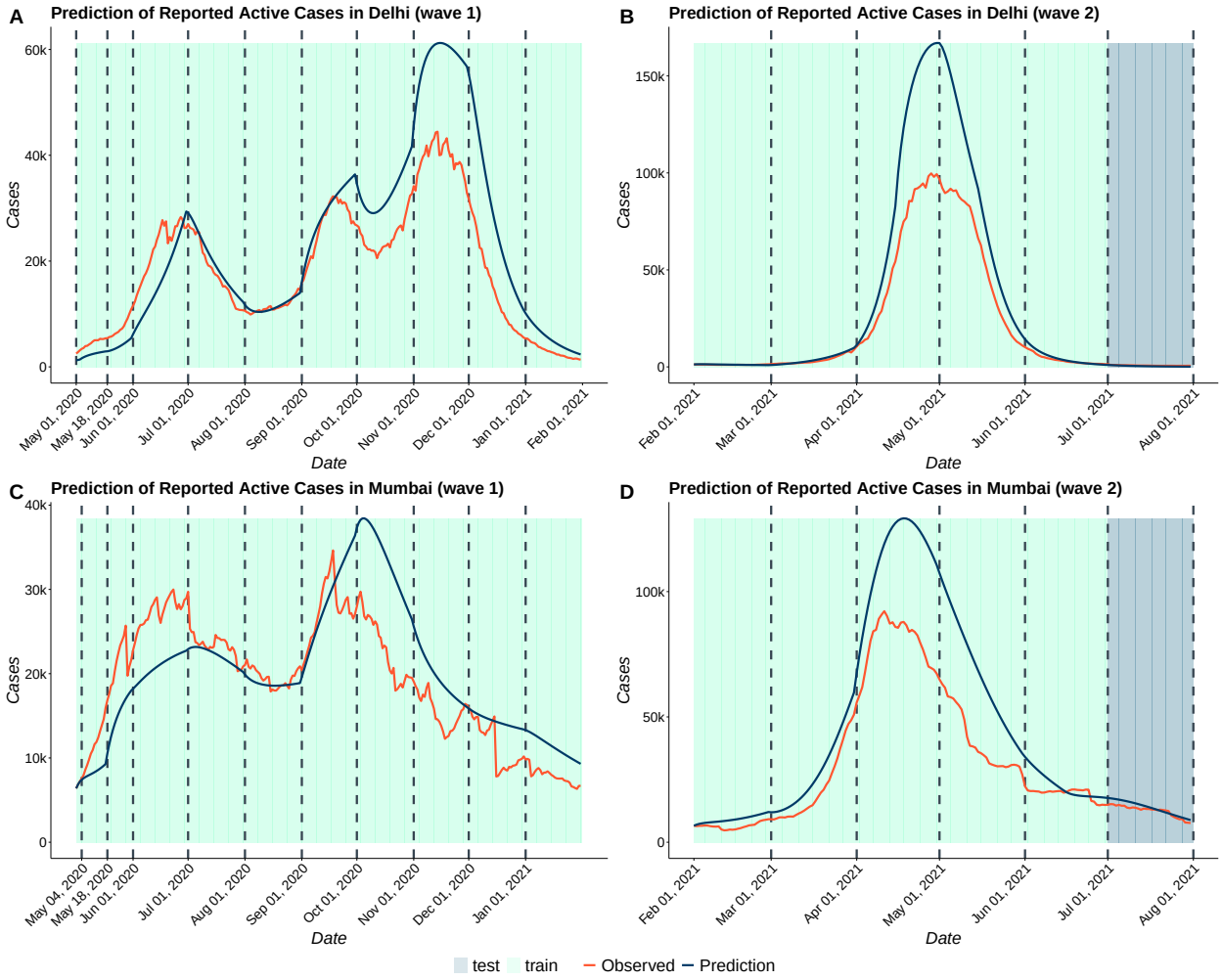

Figure S.8: Predicted Reported Active Cases for (A) Delhi (wave 1), (B) Delhi (wave 2), (C) Mumbai (wave 1) and (D) Mumbai (wave 2) using the Multinomial-2-parameter model. We have taken April 1, 2020 to January 31, 2021 as the first wave training period and in the first wave, we have not taken any testing period. In the wave 2, we have taken Feb 1,2021 to June 30,2021 as the training period, while, July 1,2021 to August 31,2021 was taken to be the test period.

Figure S.8 shows the predicted reported active cases of Delhi and Mumbai. Figure S.8 shows that our model fits the training data for both the places reasonably well. We also see that in the test data, our model the number of reported cases predicted by our model matches the true observed cases quite accurately.

Table S.8 presents a comparison of the under-reporting factors in Delhi and Mumbai for cases and deaths. There is tremendous heterogeneity between states, with case underreporting factors of approximately 28.3 and 13.6 in the first wave, while the underreporting factors in Wave 2 are 19 and 42 and death underreporting factors of 6.29, and 3.86 in Delhi, and Mumbai respectively in the first wave, while in the second wave they came out to be 4.4 and 5.6.

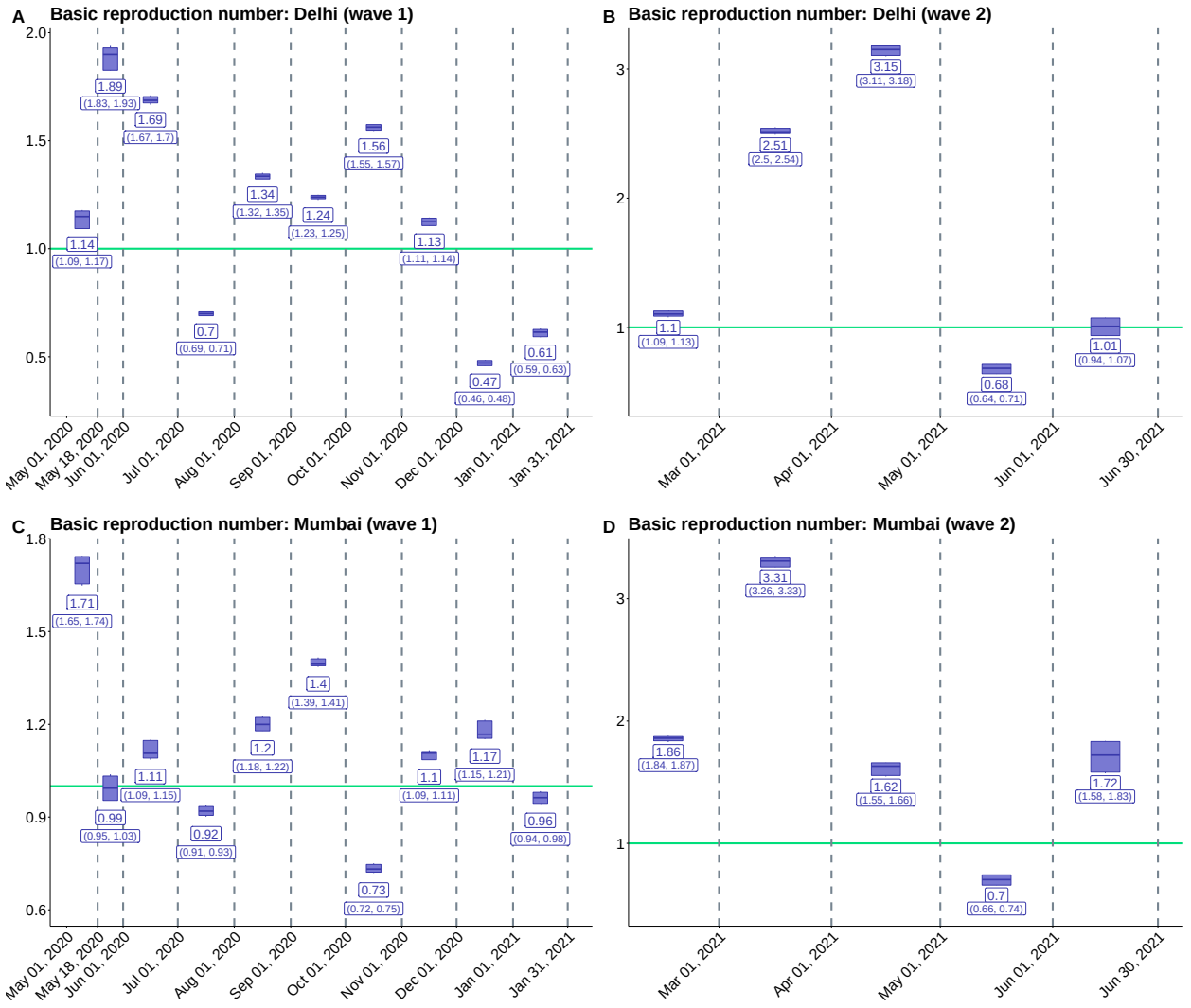

Figure S.9: Estimated basic reproduction number for (A) Delhi (wave 1), (B) Delhi (wave 2), (C) Mumbai (wave 1) and (D) Mumbai (wave 2) using the Multinomial-2-parameter model. The reproduction numbers are estimated for the training periods only in each of the 2 waves: April 1,2020 to Jan 31, 2021 for the first wave and Feb 1,2021 to June 30,2021 for the second wave.

Figure (S.9) shows the estimates of basic reproduction number for Delhi and Mumbai. We note that the peak mean estimates of R_0 for wave 1 is lower than the same in wave 2 for both the places. In fact, the peak R_0 values for wave 1 are 1.89 and 1.71 while in wave 2, they are 3.15 and 3.31 for Delhi and Mumbai respectively. This can be explained by the fact that during the second wave, both these places, as well as the rest of India, saw a much higher rate of daily new cases than that in the first wave.

S.5 Simulations

We will here discuss about the different simulations and scenarios and their methods of data generation.

S.5.1 Effect of Misclassification

Data Generation: Our generative model is the multinomial-2-parameter model with $f = 0.3$. The other parameters are fixed as in the data analysis for India: $N = 1.341$ billion, $\lambda = \mu = (69.416 \times 365)^{-1}$, $\alpha_p = 0.5,\ \alpha_u = 0.5,\ \beta_1 = 0.6,\ \beta_2 = 0.7,\ \delta_1 = 0.3,\ \delta_2 = 0.7$, $\mathrm{D_e} = 5.2$, $\mathrm{D_R} = 14$, $\mu_c = (1 - \text{mCFR})/14$ where mCFR = 0.054. We generate the data for a period of 101 days, divided into five time periods: days $1-10$, $11-31$, $32-50$, $51-64$, $65-101$. The values of $\boldsymbol{\beta}_t$ across the five periods are set at 0.8, 0.65, 0.4, 0.3, 0.3 and the corresponding values of $\boldsymbol{r}_t$ are set at 0.1, 0.2, 0.15, 0.15, 0.2. The chosen true values closely mimic the estimates for India from March 15 to June 23 and the periods mimic the phases of lockdown in India. We choose the multinomial-2-parameter model for estimation. We fit the model using the same parameters as in the model used for generating the data except $\boldsymbol{\beta}_t$, $\boldsymbol{r}_t$ and f. We consider $f \in \{0, 0.15, 0.3\}$ for prediction in the 3 scenarios. The values of $\boldsymbol{\beta}_t$ and $\boldsymbol{r}_t$ are then estimated in each of the 3 scenarios for the 5 time periods. The entire process is repeated 1000 times. In the main paper we have shown the effect of misclassification on number of total active cases. We concluded that the effect of misclassification on total active cases was substantial, but it was negligible on reported active cases. Here, we provide the mean estimates of R_0 obtained by the 3 different models with 3 different false negative rates $f = 0,\ 0.15$ and 0.3.

	Basic Reproduction Number					**MRE**		
	$\mathbf{R_{01}}$	$\mathbf{R_{02}}$	$\mathbf{R_{03}}$	$\mathbf{R_{04}}$	$\mathbf{R_{05}}$	**Lower C.I**	**Mean**	**Upper C.I**
Actual	3.99	3.65	2.12	1.59	1.69	-	-	-
Predicted Using $f = 0$	3.64	3.51	1.97	1.48	1.65	0.0036	0.0041	0.0045
Predicted Using $f = 0.15$	3.52	3.64	2.01	1.51	1.69	0.0035	0.004	0.0044
Predicted Using $f = 0.3$	3.83	3.73	2.04	1.53	1.71	0.0009	0.0012	0.0015

Table S.7: Effect of misclassification on Basic Reproduction Number. The first row denotes the true values used to simulate the data (which are based on estimates for India during the beginning of wave 1) while the later rows show estimates obtained using Multinomial-2-parameter model with 3 different values of fm which are 0, 0.15 and 0.3 respectively. We observe that the estimates of R_0 do not vary substantially with different values of f which shows the robustness of our estimate of R_0 against misclassification.

It is quite evident from the table S.7 that the R_0 is quite robust with the change of the value of false negative rate (f). Under all the false negative rates, the estimation of R_0 is quite accurate which is evident from the MRE provided in the table S.7.

S.5.2 Effect of selection

We generate data using the model described in extension 3 with most of the parameters taking the same values as the above previous simulation except for $\alpha_u = 0.7,\ \mu_c = 0.047 \cdot \frac{1}{14},\ \mathrm{D_e} = 5.2,\ \mathrm{D_r} = 14/0.953$. For Selection Model we have some additional parameters. They are set as $p_0 = (10^{-6}, 10^{-5}, 1 - 10^{-6} - 10^{-5})$ and $p_1 = (0.02, 0.18, 0.8)$. As before, the data are generated for a period of 101 days with 5 periods $1-10$, $11-31$, $32-50$, $51-64$ and $65-101$. $\boldsymbol{\beta}_t = (0.6, 0.4, 0.3, 0.25, 0.2)'$ for the 5 periods. Predictions are based on the Multinomial-2-parameter model, where the probability of being tested is assumed to be independent of symptoms with $f = 0.3$ (the simulation truth). Ignoring the misclassification will lead to even larger biases but we chose to decouple the effect of the two. As we have observed in section 4.2 of main paper, selection bias has a substantial impact on the estimates of R_0 and case counts.

S.5.3 Effect of number of tests

In this section, we study the effect of increasing testing on the course of the pandemic. We expect that with increasing number of tests we have a better chance of identifying the infectious people, which might result in a faster end to the pandemic. To test this hypothesis, we use our Selection Model(Extension 3) to explore the population infection rates as a function of the number of available tests. **Generation Model:** We use the same Selection model for generating these data as in the previous section. To generate the data, we use five different scenarios where values of all the parameters except the number of tests is the same as in the previous simulation. The data are generated for a period of 1000 days with 5 periods $1-10$, $11-31$, $32-50$, $51-64$ and $65-1200$. The values of β_t for the 5 periods were

$(0.6, 0.4, 0.3, 0.2, 0.05)$ for all the models. The only difference between the models is the number of tests. For the first scenario, we generated the number of tests such that the number of test increases exponentially from 20,000 on the first day to 1 million on the 1200^{th} day. We used the following equation for generating such a sequence :

$$\mathcal{T}(t) = \text{start} \times e^{(t-1)\log(\lambda)} \; \forall t \in \{1, 2, ..., 1200\} \text{ where } \lambda = \frac{\log\left(\frac{\text{start}}{\text{end}}\right)}{1200 - 1}$$

Let us denote the sequence of no of tests generated by $\mathcal{T}$. Now, for the 5 scenarios, we generated data using number of tests equal to 1, 2, 3, 4 and 5 times $\mathcal{T}$ respectively. We repeat the process 1000 times and take analyze the mean predictions of total active cases.

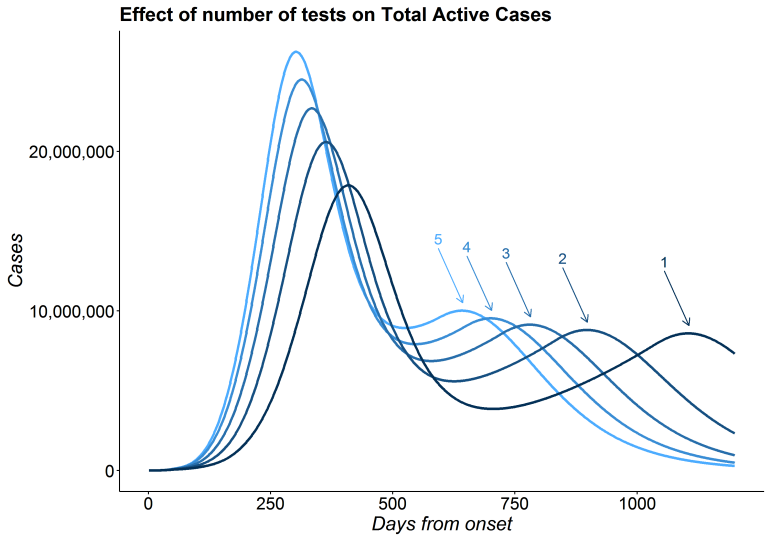

Figure S.10: Effect of test : Plot showing the number of total active cases over time for different number of tests. The numbers above the arrows indicate the multiplication factor of number of tests. We can observe that with higher number of tests, we have earlier peaks and a quicker drop in the number of cases as well.

Results: Figure S.10 shows the mean number of total active cases across 1000 simulations as a function of time since pandemic onset. With a higher number of tests, the pandemic ends faster. We observe that the number of days before the first time the number of total active cases comes below the 1 million mark (after attaining the peak) for the models with number of tests = $3\mathcal{T}$ and $4\mathcal{T}$ is 1.14 and 1.06 times that of model with with number of tests = $5\mathcal{T}$. For the models with number of tests = $\mathcal{T}$ and $2\mathcal{T}$, we note that cases do not come below the 1 million mark withing the 1200 day period.

We also note that all the predictions in the 5 scenarios predict the existence of a 2^{nd} peak. This is a proof of concept illustration with our formulation. We further observe that with higher number of tests, the gap between the first and second peak becomes smaller. Here, for the models with number of tests equal to 2, 3, 4 and 5 times $\mathcal{T}$, we have the gap between the 2 peaks as 0.77, 0.64, 0.56 and 0.49 times that of model with number of tests equal to $\mathcal{T}$. We should note that while the actual values presented in this simulation do not bear any resemblance with those of any country or state, the relative orders and findings convey important insights regarding effect of the number of tests on the number of infections.

S.6 MCMC convergence diagnostics

Since the training period for our simulation study was divided into 1-months bins, resulting in a large number of month-specific parameters, we discuss the convergence diagnostics for the key parameters (β_t, r_t and R_0) corresponding to the last period for the sake of simplicity. Recall that the true values for these three parameters were set at $\beta_t^* = 0.25$, $r_t^* = 0.1$ and and $R_0^* = 1.5$ respectively.

We further assess the convergence and mixing properties for our MH sampler by running two chains each for 200,000 iterations with burn-in of 100,000 iterations, using standard diagnostic tools such as the Gelman-Rubin $\hat{R}$ [6], trace-plot and autocorrelation. To generate diverse initial values, we initialize the first chain using Gaussian MLE estimates and

the second using a random walk perturbation of the initial values of the first chain. The trace-plots and autocorrelation plots are shown in (Fig. S.11, A and B) and the densities of the posterior samples in Fig. S.11)-C. The Gelman–Rubin $\hat{R}$ statistics for the three parameters (β_t, r_t and R_0) came out to be 1.02 (upper CI: 1.1),1.05 (upper CI: 1.23) and 1 (upper CI: 1.01), respectively, indicating convergence. These diagnostics suggest rapid convergence, adequate mixing and low autocorrelation for the sampler. The diagnostic plots have been generated using the `ggmcmc` package [5] in `R`. Note that the results on the main paper and the supplement are based on posterior summaries based on one long chain with 200,000 iterations iterations out of which the first 100,000 iterations are considered to be the burn-in period.

S.7 Sensitivity Analysis

Since we have not estimated the values of quite a few parameters, a sensitivity analysis is necessary. Now, as doing sensitivity analysis of all the parameters and initial values is impractical, we will do sensitivity analysis for the following parameters only.

1. $\mathbf{E_0}$: The initial value of Exposed had been chosen as 3 times the sum of initial values of Untested, Confirmed and False Negative cases. Such a choice might seem arbitrary. Hence, we try 4 different values of E_0 and check how the estimates of R_0 and Current Active cases vary across different values of E_0. We have assumed $E_0 = 1, 2, 3$ and 4 times $(U_0 + P_0 + F_0)$.
2. $\alpha_{\mathbf{U}}$: The value of α_U had been taken as 0.5 in the main analysis. We also assumed α_P = 0.5. So, we effectively assumed that the rate of transmission of disease by untested and tested positive individuals was same. Some things to consider when choosing the value of α_U and α_P were that individuals who were tested positive are quarantined and/or hospitalized reducing their rate of transmitting the disease. And untested cases are predominantly asymptomatic cases whose rate of spreading the virus is much less than symptomatic cases. So, we have $\alpha_U < 1$ and $\alpha_P < 1$. However, we do not know if $\alpha_U > \alpha_P$ or $\alpha_U < \alpha_P$. So we try 4 different values of α_U here which are $\alpha_U = 0.3, 0.5, 0.7$ and 1.
3. $\mathbf{D_E}$: We stated in the beginning of this paper that we have assumed the Incubation period equals the Latency period (= D_E). We have taken $D_E = 5.2$ days following the results by Lauer et al. [10]. However research by other groups suggest different values of incubation period like 6.4 days by Becker et al. [2] etc. So we consider 3 values of D_E for sensitivity analysis. They are D_E = 6.4, 5.2 and 4.1 (lower limit of 95% CI of estimates of incubation period by Lauer et al. [10])
4. $\mathbf{k}$: For Multinomial Symptoms model, one important parameters is k which is the ratio of probability of a mildly symptomatic person being tested to that of an asymptomatic person being tested. Since the probability of testing is higher for a mildly symptomatic person than an asymptomatic person, so $k > 1$. In our main analysis, we assumed $k = 4$. The choice of k was not supported by any data. So, we try 4 different values of k : $k = 3, 4, 5$ and 6 and look at the different estimates.

S.7.1 Effect of initial value of Exposed

We start with the initial value of Exposed individuals. Throughout our analysis we have assumed that the number of exposed individuals on the starting day i.e. 1^{st} April was thrice the number of total expected infected up to that day. So we check how much our estimates vary if we vary the starting value of Exposed (E_0). So, we use 4 starting values for E_0:

1. $E_0 = U_0 + P_0 + F_0$
2. $E_0 = 2(U_0 + P_0 + F_0)$
3. $E_0 = 3(U_0 + P_0 + F_0)$
4. $E_0 = 4(U_0 + P_0 + F_0)$

We can observe from subfigure B of Figure (S.12) that our estimates of R_0 are relatively robust with respect to choice of initial values of exposed. The only substantial variation is observed in the first time period - 1^{st} - 14^{th} April. Now, let us look at how the estimates of number of active cases change with different initial values.

We can observe from subfigure A of Figure (S.12) that all the estimates for total active cases increases with increasing values of E_0. The estimate of total active cases on 30^{th} June for $E_0 = 4(U_0 + P_0 + F_0)$ was more than 2 times that for $E_0 = (U_0 + P_0 + F_0)$. Hence we observe that though the estimates of total active cases vary substantially with different initial number of Exposed people, the estimates of Basic Reproduction Number are much more robust to such variations. Now we look at the effect of α_U on our estimates.

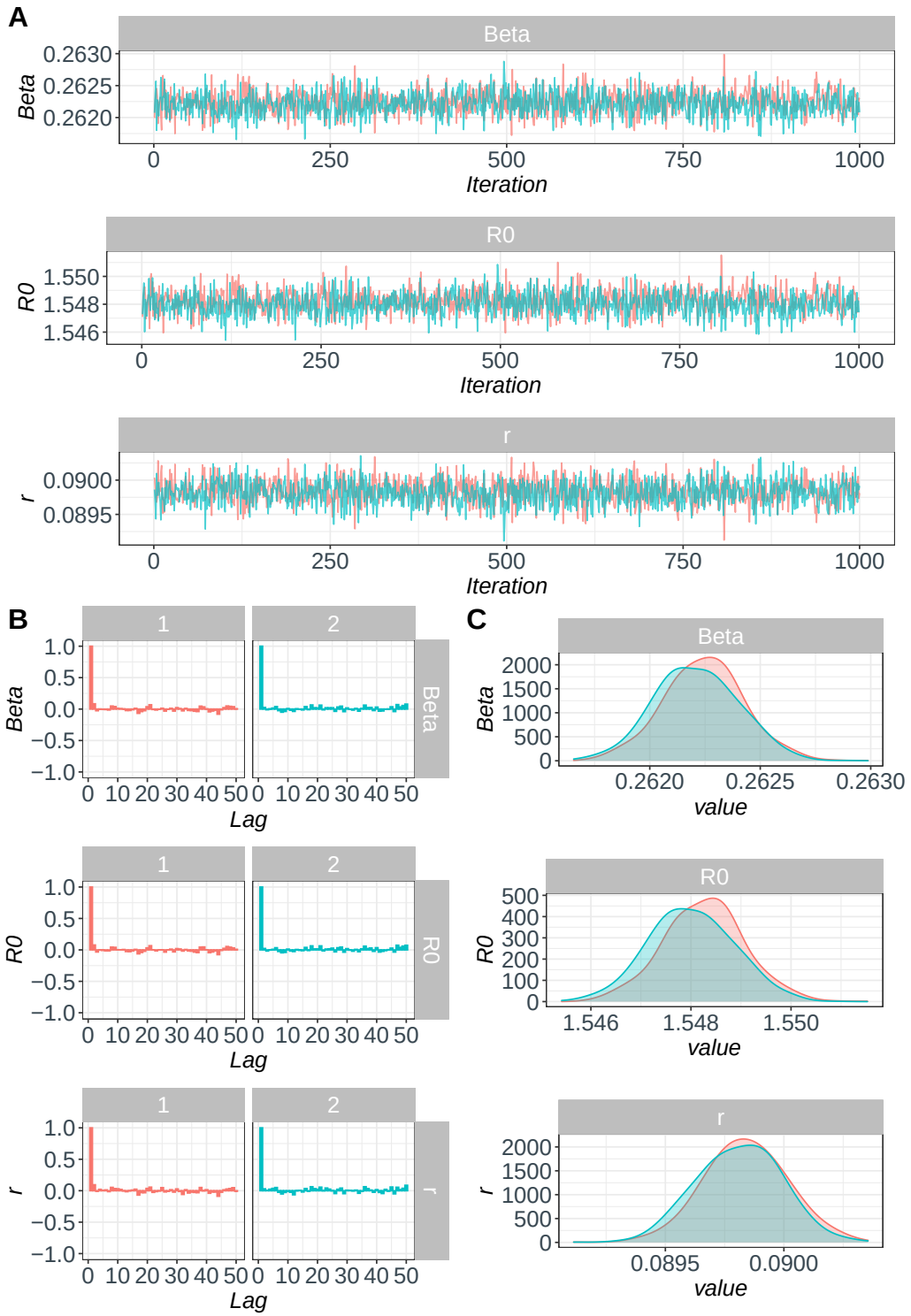

Figure S.11: The Figures under Block A correspond to the trace plots for the three parameters β, r and R_0, while those under Block B correspond to the autocorrelation plots and finally the Block C correspond to the density plots.

S.7.1.1 Effect of α_u

In our main analysis we assumed $\alpha_U = 0.5$. Here, we try 4 different values of α_U, $\alpha = 0.3, 0.5, 0.7$ and 1. First, we

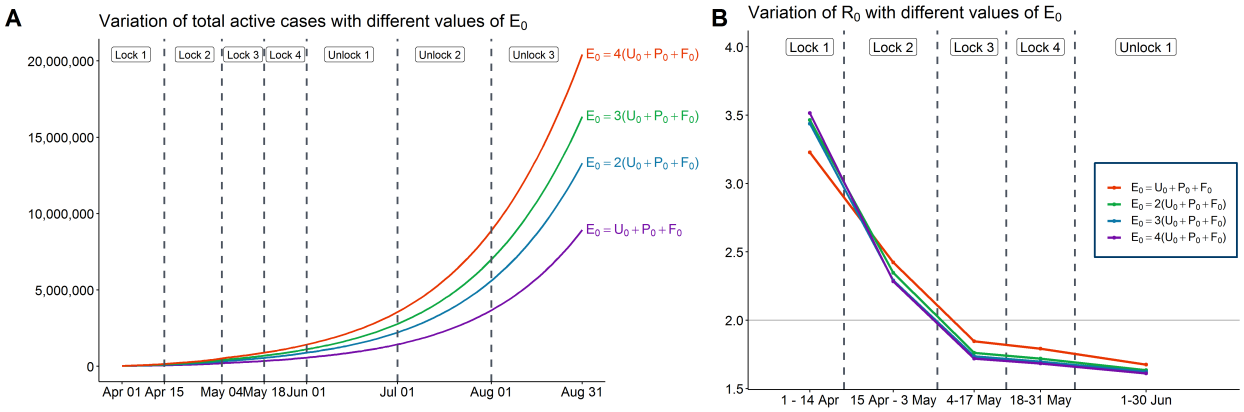

Figure S.12: Variation of estimates of R_0 with different values of E_0 which is the initial number of Exposed individuals. 4 values of E_0 are considered, which are, $E_0 = U_0 + P_0 + F_0$, $E_0 = 2(U_0 + P_0 + F_0)$, $E_0 = 3(U_0 + P_0 + F_0)$ and $E_0 = 4(U_0 + P_0 + F_0)$

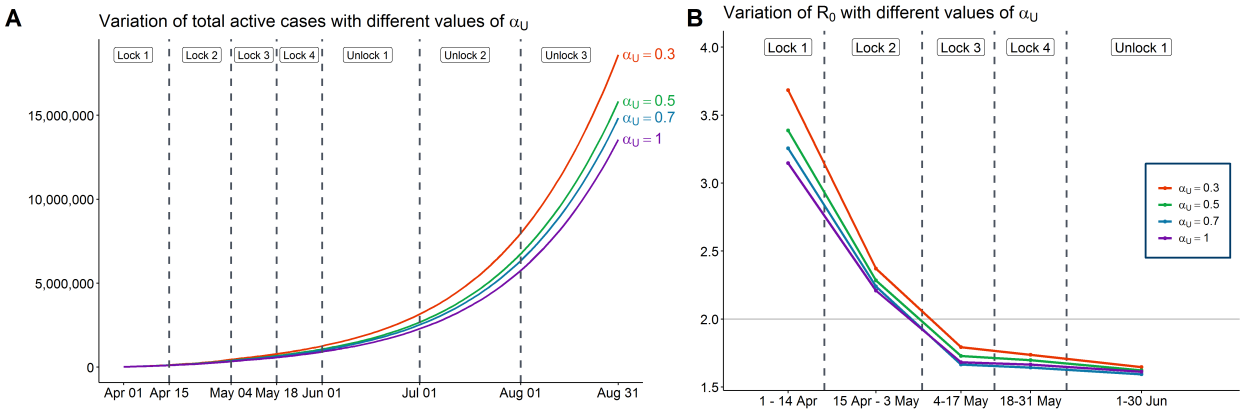

Figure S.13: Variation of estimates of R_0 with different values of α_U which is the scaling factor of the infection rate of untested individuals as compared to false negative individuals. 4 different values of α_U are considered which are $0.3,\ 0.5,\ 0.7$ and. 1 respectively.

look at the estimates of R_0. Similar to the previous section, from subfigure B of (S.13), we observe that the estimates of R_0 are more or less similar for different values of α_U. Once again, the only R_0 that substantially varies with different values of α_U is the first one i.e. R_{01}. Now, we look at the estimates of total active cases.

Subfigure A of (S.13) shows that the estimated value of total active cases decreases with increasing value of α_U. The reason behind this is if the value of α_U is higher, then a smaller number of untested cases will spread the same amount of infection as a larger number of cases would have if the value of α_U had been lower.

So, once again, we observe that while the estimates of the number of active cases are influenced heavily by α_U, the estimates of R_0 remain relatively unaffected by the change.

S.7.2 Effect of D_e

In our model, the time an individual stays in the compartment E is assumed to follow an exponential distribution whose mean is equal to the latency period. Based on estimates from [7] we have taken the value of latency period to be 5.2 days. While the estimates of latency period by different groups of researchers vary substantially, most papers list their mean estimates between 5 and 6 days. For example, a recent paper Xin et al. [14] estimates mean latency period to be 5.5 days. In this context, we would like to point out another implicit assumption in our model. For the sake of simplicity, the exposed individuals E are assumed to move to some compartment among P, U or F after the

latency period. This is equivalent to assuming that the incubation period and latency period are equal. This seems to be a reasonable assumption in the context of COVID-19 as the values of the incubation period and latency period are estimated to be very close. For example, a meta-analysis by [4] finds the mean incubation period to lie in the range 5.2 days (95% CI 4.4 to 5.9) to 6.65 days (95% CI 6.0 to 7.2). These numbers did change with the rise of the Omicron variant [1], but that does not affect our data analysis as it ends in August 2021 before Omicron started its circulation. To check whether the predictions from our model are robust to different latency periods, we have performed sensitivity analyses corresponding to data from waves 1 and 2 in India.

In the first study, we try two values of D_e (4.1 and 6.4) and check how the new estimates stack up against our original predictions using $D_e = 5.2$.

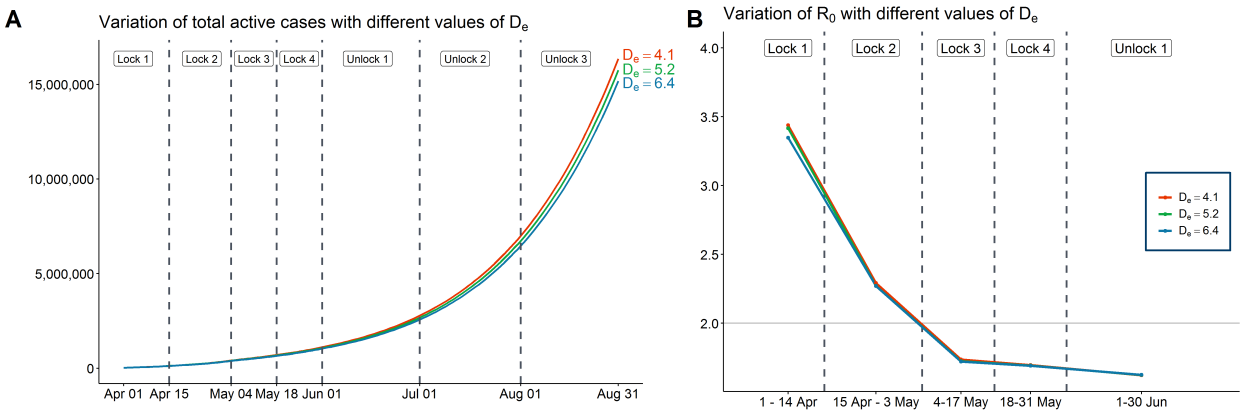

Figure S.14: Variation of estimates of R_0 and total active cases with different values of D_e which is the incubation period. 3 different values of D_e are considered, which are, 4.1, 5.2 and 6.4 respectively.

As we can observe from subfigure B of S.14, the estimates of R_0 are robust with respect to different values of D_e. From subfigure A of Figure (S.14), we note that the predicted number of active cases vary with the different values of D_e. However, unlike the previous cases, we do not observe substantial variation with different variation of D_e. But one thing, we need to observe that in this setup, the value of D_e is just varying from 4.1 to 6.2. So, we performed a second sensitivity analysis on for the second wave (Feb 1, 2021 – Jun 30, 2021), for a broader range of values for D_e, motivated by the predominance of delta variant[12] during the second wave of pandemic in India, with a much smaller latency period (3-4 days) [11]. For the second sensitivity analysis, we have used 5 values of $D_e \in \{2, 3, 4, 5, 6\}$.

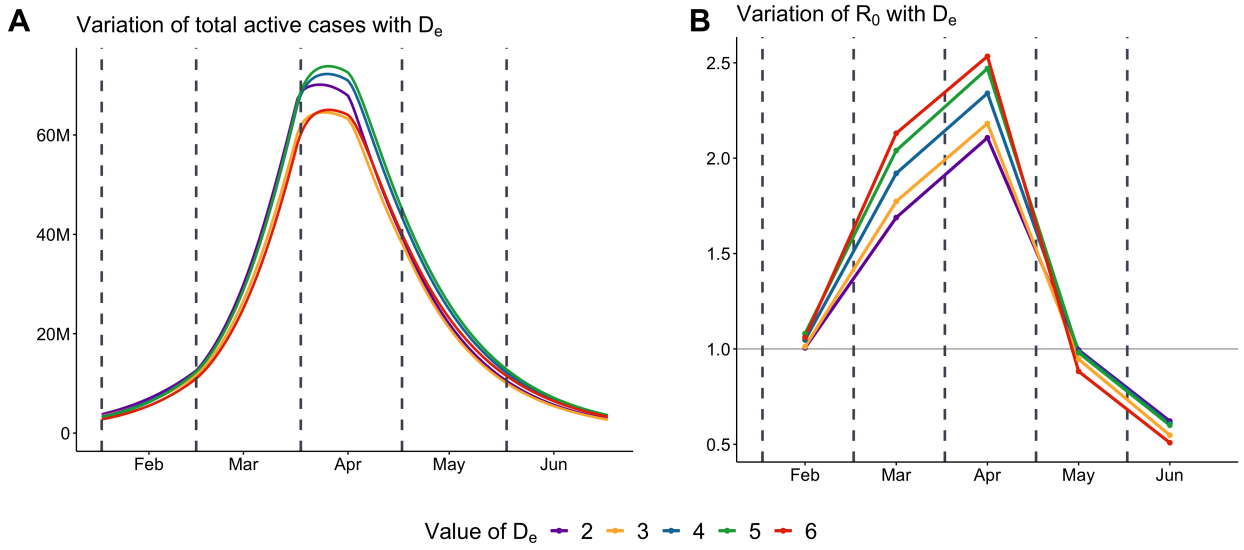

Figure S.15: Variation of estimates of R_0 and total active cases with different values of D_e which is the incubation period. 5 different values of D_e are considered, which are 2, 3, 4, 5 and 6 respectively.

Figure S.15 presents the results of the sensitivity study. We observe that the estimate of R_0 varies moderately with D_e, the maximum R_0 for the model D_e=6 being 2.5 and that for the model D_e=2 being 2 (approximately). In spite of this variation across the different values of D_e, we see that the general trend of the basic reproduction captured by all the models is similar. On the other hand, the peak total active cases for the models varies from 64.5 million to 73.8 million. As with R_0, the general trend appears to be similar even after varying the latency period over a wide range of values. We would like to point out here that while the figure only includes total active cases, the robustness properties extend similarly to estimates of cumulative cases and deaths as well.

To summarize, we observe that the predictions and estimates of our model exhibit some variation with changing values of D_e. As we cited earlier there has been studies showing that the value of D_e for the ancestral variant has been in the 5-6 days range. The alpha and delta variant estimates also fall closer to this range. Since the current paper analyzes data during wave 1 and 2 from India (April 2020 to June 2021) our choice of D_e aligns with the estimates of D_e for the ancestral, alpha and delta variants which have been the principal variants during this period. On the other hand, the latency period of recent variants like Omicron have been estimated to be much shorter at around 3 days [8]. So, with new emerging variants we would recommend choosing reliable and accurate estimates [1] of D_e and other similar parameters from other external studies and incorporate them in our model. Different periods of the pandemic will need different values of D_e based on the dominant variant.

S.7.3 Effect of *k*

In multinomial symptoms model, we defined k as the ratio of the probability of a mildly symptomatic individual getting tested to the same for an asymptomatic individual. We chose the value $k = 4$ for our main analysis. We had argued why the value of k should be greater than 1 but could not provide any justification for choosing that particular value. So, we try 4 different values of k : k = 3, 4, 5 and 6. We will start with the estimates of R_0.

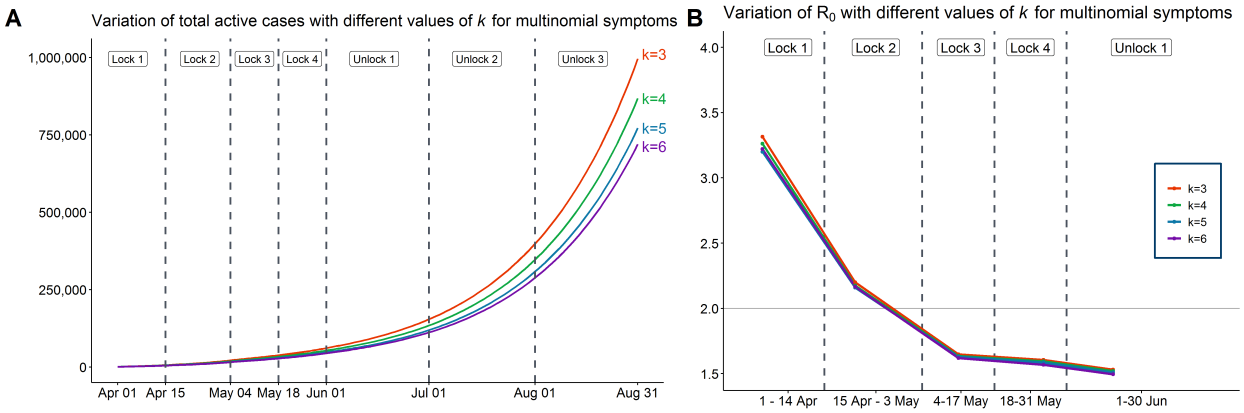

Figure S.16: Variation of estimates of R_0 with different values of k for multinomial symptoms model (which denotes the ratio of probability of being tested for mild symptomatic to that of asymptomatic individuals). k is assumed to be greater than 1. Here we have considered 4 different values of k which are 3, 4, 5 and 6 respectively.

Figure subfigure B (S.16) shows that similar to previous cases, the estimates of R_0 do not vary much with different values of k. We now look at the estimates of total active cases. From subfigure A of Figure (S.16), we note that the estimates of total active cases vary with different values of k and with higher values of k we have lower predictions of number of total active cases.

To summarize, we observe that the estimates of the Basic Reproduction number are not substantially influenced by these parameters with an exception of the first Reproduction number. We also note that the estimated number of active cases varies with different values of parameters in most of the cases. It is clearly visible that the sensitivity of the total active case predictions vary across parameters. While it does not vary much with D_E, there is substantial variation with different values of E_0.

Nation/State			1st February, 2021	1st July, 2021
Delhi	Cases	Predicted Reported (Millions)	0.62 [0.6, 0.64]	1.44 [1.43, 1.45]
		Predicted Total (Millions)	17.6 [17.06, 18.1]	33.3 [32.7, 33.9]
		Observed (Millions)	0.64	1.43
		Under-Reporting Factor	28.3 [28.2, 28.4]	19.2 [18.9, 19.5]
	Deaths	Predicted Reported (Thousands)	11.5 [11.16, 11.9]	25.36 [25.06, 25.7]
		Predicted Total (Thousands)	72.4 [69.6, 75.1]	132 [127, 137]
		Observed (Thousands)	10.8	25
		Under-Reporting Factor	6.29 [6.24, 6.33]	4.3 [4.1, 4.5]
Mumbai	Cases	Predicted Reported (Millions)	0.322 [0.318, 0.325]	0.73 [0.724, 0.73]
		Predicted Total (Millions)	4.19 [3.93, 4.54]	20.04 [19.97, 20.1]
		Observed (Millions)	0.309	0.73
		Under-Reporting Factor	13.6 [12.7, 14.7]	43.3 [43.2, 43.5]
	Deaths	Predicted Reported (Thousands)	11.99 [11.76, 12.23]	15.52 [15.39, 15.65]
		Predicted Total (Thousands)	43.9 [41.9, 46.34]	115 [114.6, 115.4]
		Observed (Thousands)	11.4	15.5
		Under-Reporting Factor	3.86 [3.69, 4.08]	9.06 [8.97, 9.16]

Table S.8: Predicted Cumulative Cases and Deaths (Reported and Total) of Delhi and Mumbai along with observed counts and predicted underreporting factors on 2 different dates, 1st February, 2021 & 1st July, 2021

References

[1] S. Abbott, K. Sherratt, M. Gerstung, and S. Funk. Estimation of the test to test distribution as a proxy for generation interval distribution for the omicron variant in england. *medRxiv*, 2022. doi: 10.1101/2022.01.08.22268920. URL `https://www.medrxiv.org/content/early/2022/01/10/2022.01.08.22268920`.

[2] J. Backer, D. Klinkenberg, and J. Wallinga. Incubation period of 2019 novel coronavirus (2019-ncov) infections among travellers from wuhan, china, 20-28 january 2020. *Euro Surveill*, 25:1–6, 2020. doi: 10.2807/1560-7917.ES.2020.25.5.2000062.

[3] M. Despotovic, V. Nedic, D. Despotovic, and S. Cvetanovic. Evaluation of empirical models for predicting monthly mean horizontal diffuse solar radiation. *Renewable and Sustainable Energy Reviews*, 56:246 – 260, 2016. ISSN 1364-0321. doi: 10.1016/j.rser.2015.11.058.

[4] W. Dhouib, J. Maatoug, I. Ayouni, N. Zammit, R. Ghammem, S. B. Fredj, and H. Ghannem. The incubation period during the pandemic of covid-19: a systematic review and meta-analysis. *Systematic reviews*, 10(1):1–14, 2021.

[5] X. Fernández-i Marín et al. ggmcmc: Analysis of mcmc samples and bayesian inference. *Journal of Statistical Software*, 70(9):1–20, 2016.

[6] A. Gelman and D. B. Rubin. Inference from iterative simulation using multiple sequences. *Statistical Science*, 7 (4):457–472, 1992.

[7] X. Hao, S. Cheng, D. Wu, T. Wu, X. Lin, and C. Wang. Reconstruction of the full transmission dynamics of covid-19 in wuhan. *Nature*, pages 1–5, 2020. doi: 10.1038/s41586-020-2554-8.

[8] L. Jansen, B. Tegomoh, K. Lange, K. Showalter, J. Figliomeni, B. Abdalhamid, P. C. Iwen, J. Fauver, B. Buss, and M. Donahue. Investigation of a sars-cov-2 b. 1.1. 529 (omicron) variant cluster—nebraska, november–december 2021. *Morbidity and Mortality Weekly Report*, 70(5152):1782, 2021.

[9] L. M. Kucirka, S. A. Lauer, O. Laeyendecker, D. Boon, and J. Lessler. Variation in false-negative rate of reverse transcriptase polymerase chain reaction–based sars-cov-2 tests by time since exposure. *Annals of Internal Medicine*, 2020. doi: 10.7326/M20-1495.

[10] S. A. Lauer and K. H. Grantz. The incubation period of coronavirus disease 2019 (covid-19) from publicly reported confirmed cases: estimation and application. *Annals of internal medicine*, 14, 2020. doi: 10.7326/M20-0504.

[11] B. Li, A. Deng, K. Li, Y. Hu, Z. Li, Y. Shi, Q. Xiong, Z. Liu, Q. Guo, L. Zou, et al. Viral infection and transmission in a large, well-traced outbreak caused by the sars-cov-2 delta variant. *Nature Communications*, 13(1):1–9, 2022.

[12] R. Thiruvengadam, A. Binayke, and A. Awasthi. Sars-cov-2 delta variant: a persistent threat to the effectiveness of vaccines. *The Lancet Infectious Diseases*, 2021.

[13] P. van den Driessche. Reproduction numbers of infectious disease models. *Infectious Disease Modelling*, 2(3): 288–303, 2017. doi: 10.1016/j.idm.2017.06.002.

[14] H. Xin, Y. Li, P. Wu, Z. Li, E. H. Lau, Y. Qin, L. Wang, B. J. Cowling, T. Tsang, and Z. Li. Estimating the latent period of coronavirus disease 2019 (covid-19). *Clin Infect Dis*, 2021.