

## Reconciling epidemiological models with misclassified case-counts for SARS-CoV-2 with seroprevalence surveys: A case study in Delhi, India

Rupam Bhattacharyya<sup>1</sup>; Ritwik Bhaduri<sup>2</sup>; Ritoban Kundu<sup>2</sup>; Maxwell Salvatore<sup>1,3</sup>; Bhramar Mukherjee<sup>1,3\*</sup>.

<sup>1</sup>: Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109, USA

<sup>2</sup>: Indian Statistical Institute, Kolkata 700108, West Bengal, India

<sup>3</sup>: Center for Precision Health Data Science, University of Michigan, Ann Arbor, MI 48109, USA

\*: Corresponding author. Address - Department of Biostatistics, School of Public Health, University of Michigan, 1420 Washington Heights, Ann Arbor, MI 48109-2029, USA. Telephone – (734) 764-6544. Email – [bhramar@umich.edu](mailto:bhramar@umich.edu).

## ABSTRACT

Underreporting of COVID-19 cases and deaths is a hindrance to correctly modeling and monitoring the pandemic. This is primarily due to limited testing, lack of reporting infrastructure and a large number of asymptomatic infections. In addition, diagnostic tests (RT-PCR tests for detecting current infection) and serological antibody tests for IgG (to assess past infections) are imperfect. In particular, the diagnostic tests have a high false negative rate. Epidemiologic models with a latent compartment for unascertained infections like the Susceptible-Exposed-Infected-Removed (SEIR) models can provide predictions for unreported cases and deaths under certain assumptions. Typically, the number of unascertained cases is unobserved and thus we cannot validate these estimates for a real study except for simulation studies. Population-based seroprevalence studies can provide a rough estimate of the total number of infections and help us check epidemiologic model projections. In this paper, we develop a method to account for high false negative rates in RT-PCR in an extension to the classic SEIR model. We apply this method to Delhi, the national capital region of India, with a population of 19.8 million and a COVID-19 hotspot of the country, obtaining estimates of underreporting factor for cases at **34-53 times** and that for deaths at **8-13 times**. Based on a recently released serological survey for Delhi with an estimated 22.86% seroprevalence, we compute adjusted estimates of the true number of infections reported by the survey (after accounting for misclassification of the antibody test results) which is largely consistent with the model outputs, yielding an underreporting factor for cases from **30-42**. Together with the model and the serosurvey, this implies approximately **96-98% cases in Delhi remained unreported** and whereas only 109,140 cases were reported on July 10, the true number of infections varied somewhere between **4.4-4.6 million** across different estimates. While repeated serological monitoring is resource intensive, model-based adjustments, run with the most up to date data, can provide a viable option to keep track of the unreported cases and deaths and gauge the true extent of transmission of this insidious virus.

## INTRODUCTION

COVID-19 was first diagnosed in Wuhan, China in December 2019 and was quickly declared a pandemic by the World Health Organization on March 11<sup>1</sup>. The first case in India was declared on January 30, and as of July 28, there have been 1,531,783 cases and 34,224 deaths reported<sup>2</sup>. India responded quickly, instituting a nationwide lockdown on March 25, when there were only 657 cases and 11 deaths<sup>2-3</sup>. However, given the high rate of asymptomatic individuals and limited and imperfect diagnostic testing, it is unclear how many people have actually been infected.

Classical epidemiologic models, like a susceptible – exposed – infected – removed (SEIR) compartmental model, have been used to predict the growth of the COVID-19 pandemic. For example, a modification of the standard SEIR model – which accounts for pre-symptomatic infectiousness, time-varying ascertainment rates, transmission rates and population movements – applied on data from Wuhan, China, identified that the outbreak had high covertness and high transmissibility<sup>4</sup>. They estimated that 87% (with a lower bound of 53%) of the infections in Wuhan before March 8 were unascertained<sup>4</sup>. However, traditional models, including the one used in this paper, do not take into account the underreporting and potential misclassification via imperfect testing.

In the context of identification of cases, there are two classes of tests that are being discussed in the literature: diagnostic tests and antibody tests. A diagnostic test (typically an **RT-PCR test**) is used to identify the presence of SARS-CoV-2, indicating an *active* infection<sup>5</sup>. An antibody test, i.e., a **serology test**, looks for the presence of antibodies, the body's immune response to fight off SARS-CoV-2, indicating a *past* infection<sup>6</sup>. Figure 1 exhibits a timeline in terms of when these tests are done during the time course of an infection. Due to a large number of asymptomatic cases and limited number of tests, many infections do not get detected. Population-based seroprevalence surveys, therefore, give us an idea about the “true number of infections” including reported and unreported cases, and consequently, the ascertainment rate<sup>6</sup>. Thus, adjusted estimates of total number of cases and ascertainment rates based on serological surveys when available provide an option to validate model-based estimates of unreported cases and underreporting factors, which would usually be impossible to validate (except for a simulation study) since these numbers are not observable the real data.

In an attempt to understand the spread of the virus in Delhi, the national capital region of India and one of the hotspots of COVID-19 in the country, the National Centre for Disease Control (NCDC) in India performed a serological survey in Delhi. While limited on reported details, the Delhi Serology Study collected 21,387 random samples across 11 districts in Delhi between June 27 and July 10 and found COVID-19 antibodies present in 22.86% of samples<sup>7-9</sup>. A simple proportional estimate would tell us that Delhi, with approximately 19.8 million people, had a total number of cases standing somewhere around 4.6 Million by July 10. This contrasts sharply both with the 109,140 cumulative cases (3,300 total deaths) reported in Delhi as of July 10, which represents roughly 0.55% of Delhi's population, indicates that roughly, only 2.4% of cases are being detected (underreporting factor of about 42), and also implies that the infection fatality rate (IFR) for Delhi is of the order of 0.07% or 717 per million. This IFR seems low compared to estimates worldwide<sup>10</sup> and as such it may be reasonable to argue that COVID related deaths are also possibly unreported, or the cause of death misclassified. The doubt regarding death data is further substantiated as a very small fraction of deaths in India are medically reported<sup>11</sup> and the IFR estimates for SARS-CoV-2 from other studies in the world<sup>10</sup> appear to be higher than influenza (infection fatality rate of influenza, as of 2018-19, is at 961 per million or around 0.1%)<sup>12</sup>

Both diagnostic and antibody tests suffer from the issue of false negatives and false positives, and depending on which test is being talked about, one or the other of these errors are more crucial. For the RT-PCR test, false negatives are more worrisome since that means allowing an infected person to go about freely, potentially spreading the virus around. Similarly, we worry about the false positives of an antibody test more, since it gives the false impression that the person has been infected in the past, gained immunity, and is unlikely to be infected again. The RT-PCR test is quoted to have a high false negative rate, ranging from 15-30% (i.e., low sensitivity, 85-70%), and a low false positive rate around 1-4% (i.e., high specificity, 99-96%)<sup>13</sup>. The antibody test assays are more precise - the commercial assays have sensitivity around 97.6% and specificity of 99.3% (DiaSorin) at about 15 days after infection<sup>14</sup>. The ELISA assay used in the Delhi serosurvey is a customized assay, about which no official information was not publicly released, but some discussions on imperfections of the test are available in public media domain<sup>15</sup>. In light of these imperfections and the high rate of asymptomatic COVID-19 cases, we develop an extension of a standard SEIR model incorporating misclassification due to imperfect diagnostic testing to predict both the number of unreported cases and deaths and to estimate the rate at which COVID-19 cases

are being underreported. An alternative approach that has been discussed in contemporary literature is to model test activities in a way such that symptomatic individuals are identified and successfully isolated with a given average delay from the onset of symptoms<sup>16</sup>. This approach can handle both real-time RT-PCR tests and lab-based sero-surveillance.

Because the Delhi Serology Study provides a seroprevalence estimate, this is a unique opportunity to help validate predictions for latent unreported infections for a SEIR model. We perform adjustments of the reported case counts (and hence corresponding metrics such as the underreporting factor, infection-fatality rate and case-fatality rate) under different sensitivity and specificity assumptions for the diagnostic and antibody tests and potential underreporting of the death counts, and validate the model-based estimates of the extent of underreporting to those obtained from the seroprevalence-based calculations. We apply this framework to Delhi, using reported COVID-19 data from [covid19india.org](https://covid19india.org)<sup>2</sup>. This framework can be adapted to and applied outside of Delhi and in other contexts where imperfect and limited testing exists.

## RESULTS

### Extended SEIR Model Adjusted for Misclassification

Under low (0.7), medium (0.85) and perfect (1) sensitivity, and perfect (1) specificity assumptions for the RT-PCR diagnostic test, we perform predictions of total (reported and unreported) cases and deaths for Delhi using the proposed extended SEIR model. Using data till June 30, this model estimates 4.8 million cases and 33,165 deaths on July 10 if we assume the RT-PCR test has a sensitivity of 0.85, and those predicted counts become 4.2 million and 28,499, respectively, if the sensitivity is assumed to be 1.0. Compare to the observed case and death counts of 109,140 and 3,300 reported in Delhi as of July 10.<sup>2</sup> The model predictions under the different scenarios considered and the performances in terms of fitting the daily observed case and death counts are summarized in Figures 2-3. Looking at the ratio of predicted total number of cases and the predicted number of reported cases on July 10, it appears that the underreporting factor for cases reported by the model is within the range of **34-53** and the same for the deaths is between **8-13** (Table 1 and Figure 4). This implies according to the model **97-98%** Delhi's cases remain undetected.

### Naïve Corrections to Reported Test Results using Known Misclassification Rates for Tests

Since the unreported number of cases and subsequently, the underreporting factor, are not part of the observed data and therefore cannot directly be validated, we validate these estimates using adjustments regarding the imperfection of the tests and estimated number of true infections predicted by the serosurvey data. We also consider the hypothetical scenario considering 10-fold underreporting of deaths, as suggested by the model outputs. However, we are not able to perform any validation for the estimated underreporting factor for deaths as we do not have estimate of true death rates or excess deaths.

Using varying (low to perfect) sensitivities and specificities for the diagnostic and antibody tests, we estimate that the true case count in Delhi as of July 10 lies between **4.4 and 4.6 million**, which represents **30 to 42 times** the number of reported cases (Table 2), *these estimates being greatly in*

agreement with the model outputs, as reported in the previous subsection. This indicates **96-97%** cases in Delhi were underreported.

**Case fatality rate (CFR) and Infection fatality rate (IFR):** The sensitivity and specificity of the diagnostic test impact our estimate of the case-fatality rate ( $\frac{\#deaths}{\#reported\ cases}$ ), but not the infection-fatality rate ( $\frac{\#deaths}{\#true\ infections}$ ). We estimate that the CFR lies between 2.24-3.02%. On the other hand, the sensitivity and specificity of the antibody test impact our estimates of the IFR. We estimate that the IFR lies between 0.07-0.08% based on the reported death counts. If we consider a 10-fold underreporting in death counts, the infection-fatality rate estimate increases to 0.7-0.8% (Table 2).

## DISCUSSION

We developed an extension of the standard SEIR compartmental model to adjust for the misclassifications due to imperfect diagnostic testing. Applying our model on publicly available infection and death data for Delhi, we estimated the underreporting factor for cases to be somewhere between **34 and 53** and that for deaths to be somewhere between **8 and 13** on July 10. Further, using adjustments under different imperfection scenarios for the diagnostic and antibody tests, we came up with adjusted estimates of the underreporting factor which agreed greatly with and validated those estimated from the model. Having an accurate idea about the underreporting factor and the extent of spread is extremely helpful in terms of tracking the growth of the pandemic and determining policies. Repeated serological surveys to track the ever-evolving seroconversion scenario are often not viable options due to being expensive both in terms of cost, resources and time, the model estimates, updated regularly with new incoming data, provide an option to keep track of the underreporting factor and unreported cases and deaths.

The Delhi Serology Study is one of several COVID-19 serology studies that have been conducted across the world (for a summary of such studies, please refer to Table 3)<sup>8, 17-28</sup>. The seroprevalence found in the Delhi Serology Study (22.86%) is the highest among these studies but is similar to that found in New York City (22.70%), another large, densely populated area<sup>26</sup>. This indicates that Delhi is definitely on the higher side in terms of seroprevalence, even within worldwide epicenters and hotspots of COVID-19. Another serosurvey conducted in the city of Mumbai, India found strikingly different seroprevalences in slum (57.8%) vs non-slum areas (16%), the overall estimated seroprevalence also being staggeringly high (40.5%) compared to other studies across the world (Table 3)<sup>19</sup>.

Extensive and long-drawn discussions have already taken place in relation to potential community transmission of COVID-19 in India. While even without a serological survey it is possible to comment on this based on other information available in terms of the number of tests and test positive rates, our results confirm that at least for Delhi, there is undoubtedly community transmission with regards to the classical definition of the term<sup>29</sup>. With more than 500,000 active cases in India and more than 10,000 active cases in Delhi as of July 28, many of the cases are potentially not being tracked to an identifiable source of infection. Along these lines, there have been debates about achieving herd immunity, and estimated range for the herd immunity threshold lies within 44-73% (based on worldwide estimated basic reproduction number of 1.8-3.8)<sup>30-31</sup>.

Both for Delhi and more so possibly for other parts of India, herd immunity will potentially take some time to be attained and is definitely not a panacea we can rely on. Even based on the IFR obtained without adjusting for potential death underreporting and trusting the reported death counts (Table 2), if 50% people in India, with a population of 1.38 billion, get infected (a concept that many proponents of herd immunity have suggested), it would imply an estimated **550,000** deaths, which skyrockets to an estimate of a staggering 5.5 million deaths if we believe the estimated underreporting factor from death from our proposed model.

There are several factors that we need to take into account about the Delhi Serological Study and consequently, these factors also shape the potential implications of our results. A large set of important information isn't well-known or wasn't publicly reported in the NCDC serology survey, such as the response and positivity rates stratified by age, sex, job type, district; sampling design; sensitivity or specificity of the customized assay – and so on. Releasing a single magic number without a complete report is definitely not the best practice for science and policy and is dissatisfying and potentially misleading. Further, we do not know if individuals with antibodies are protected from re-infection or how long this protection lasts<sup>32</sup>. We need to know more about the longevity of the antibody response and the levels of it needed to protect us from re-infections, and also about the contagiousness and potential clinical severity of a person with the antibody.

Even though the appearance and spread of COVID-19 has taken the entire world by a storm, a large number of examples from all across the world clearly depict that with extensive testing, contact tracing, use of masks, hand hygiene and social distancing we can change the narrative and course of this virus. For example, Delhi has seen tremendous success in turning the corner of the virus curve with the reproduction number at 0.74 (July 28) and the R staying below unity for about a month (Figures 5-6). Rapid and significant scientific advancements in both clinical and public health aspects of the disease have been made over the past few months<sup>33</sup>, and focused and cautious tracking of the pandemic with informed policy decisions are going to be as helpful as ever at this point. In that line, our analytical framework of integrating diagnostic testing imperfections in context of estimating unreported cases using the extended SEIR model and validating against seroprevalence estimates will hopefully prove to be useful for other case-studies.

## **METHODS**

### **Extended SEIR Model Adjusted for Misclassification**

We developed an extension of a standard SEIR model. In this model, the susceptible individuals (S) become exposed (E) when they are infected, but they have not started infecting the other people. After a latency period, exposed individuals are able to infect other susceptible individuals and are either untested (U) with probability  $r$  or tested (T) with probability  $1 - r$ . Tested individuals enter either the false negative compartment (F) with probability  $f$  or the (true) positive compartment with probability  $1 - f$ . Individuals who are in the untested and the false negative compartments are considered unreported COVID-19 cases and enter either the recovered unreported (RU) or death unreported (DU) compartments. Similarly, those who tested positive move to either a recovered reported (RR) or death reported (DR) compartment. Figure 7 represents the SEIR model schematic. The corresponding system of differential equations are presented below. The parameters and their values used are described in Table 4.



1.  $\frac{\partial S}{\partial t} = -\beta \frac{S(t)}{N} (\alpha_P P(t) + \alpha_U U(t) + F(t)) + \lambda - \mu S(t).$
2.  $\frac{\partial E}{\partial t} = \beta \frac{S(t)}{N} (\alpha_P P(t) + \alpha_U U(t) + F(t)) - \frac{E(t)}{D_e} - \mu E(t).$
3.  $\frac{\partial U}{\partial t} = \frac{(1-r)E(t)}{D_e} - \frac{U(t)}{\beta_1 D_r} - \delta_1 \mu_c U(t) - \mu U(t).$
4.  $\frac{\partial P}{\partial t} = \frac{r(1-f)E(t)}{D_e} - \frac{P(t)}{D_r} - \mu_c P(t) - \mu P(t).$
5.  $\frac{\partial F}{\partial t} = \frac{rfE(t)}{D_e} - \frac{\beta_2 F(t)}{D_r} - \frac{\mu_c F(t)}{\delta_2} - \mu F(t).$
6.  $\frac{\partial RU}{\partial t} = \frac{U(t)}{\beta_1 D_r} + \frac{\beta_2 F(t)}{D_r} - \mu RU(t).$
7.  $\frac{\partial RR}{\partial t} = \frac{P(t)}{D_r} - \mu RR(t).$
8.  $\frac{\partial DU}{\partial t} = \delta_1 \mu_c U(t) + \frac{\mu_c F(t)}{\delta_2}.$
9.  $\frac{\partial DR}{\partial t} = \mu_c P(t).$

Here,  $X(t)$  denotes the number of individuals in the compartment of interest  $X$  at time  $t$ . In this setting, both  $\beta$  and  $r$  are time-varying parameters which are estimated using the Metropolis-Hastings MCMC method<sup>34</sup>. To estimate the parameters, we first need to be able to solve the differential equations, which is difficult to perform in this continuous-time setting. It is also worth noting that we do not require the values of the variables for each time point, we only need their values at discrete time steps, i.e., for each day. Thus, we approximate the above set of differential equations by a set of recurrence relations. For any compartment  $X$ , the instantaneous rate of change with respect to time  $t$  (given by  $\frac{\partial X}{\partial t}$ ) is approximated by the difference between the counts of that compartment on the  $(t + 1)^{th}$  day and the  $t^{th}$  day, that is  $X(t + 1) - X(t)$ . Starting with an initial value for each of the compartments on the Day 1 and using the discrete-time recurrence relations, we can then obtain the solutions of our interest. Some examples of these discrete-time recurrence relations are presented below.

1.  $E(t + 1) - E(t) = \beta \frac{S(t)}{N} (\alpha_P P(t) + \alpha_U U(t) + F(t)) - \frac{E(t)}{D_e} - \mu E(t),$
2.  $U(t + 1) - U(t) = \frac{(1-r)E(t)}{D_e} - \frac{U(t)}{\beta_1 D_r} - \delta_1 \mu_c U(t) - \mu U(t),$
3.  $P(t + 1) - P(t) = \frac{r(1-f)E(t)}{D_e} - \frac{P(t)}{D_r} - \mu_c P(t) - \mu P(t),$
4.  $F(t + 1) - F(t) = \frac{rfE(t)}{D_e} - \frac{\beta_2 F(t)}{D_r} - \frac{\mu_c F(t)}{\delta_2} - \mu F(t).$

The rest of the differential equations can be similarly approximated by a discrete-time recurrence relation. These parameters are estimated using training data from Delhi from March 15 to June 30. The training data is divided into 7 periods, in accordance with the lockdown and unlock procedures employed by the government of India, as described in Table 5. Using these, we performed our predictions for the dates ranging from June 1 through July 26.

## Naïve Corrections to Reported Test Results using Known Misclassification Rates

Let us set up a few notations first. Let  $N$  = population size,  $X$  = number of true cases in the population (hence  $N - X$  = number of non-cases in the population),  $T$  = number of people tested,  $S$  = number of true cases tested (hence  $T - S$  = number of non-cases tested,  $X - S$  = number of true cases not tested,  $N - X - T + S$  = number of non-cases not tested),  $P$  = number of positive tests (also, therefore, cumulative number of reported cases, hence  $T - P$  = number of negative tests). Note that  $X$  and  $S$  are the only two unknowns in this setting. Also, let us assume that the sensitivity of the test of interest is  $\alpha$  and the specificity of the same is  $\beta$ . With that, we can set up the following equation, because there are two ways a test can be positive, as can be seen in Figure 8.

$$P = S \times \alpha + (T - S) \times (1 - \beta) \implies \frac{P}{T} = \frac{S}{T} \times \alpha + \left(1 - \frac{S}{T}\right) \times (1 - \beta).$$

Adjusting the terms, we get the following expression for  $S$ .

$$S = T \times \frac{\frac{P}{T} + \beta - 1}{\alpha + \beta - 1}.$$

Assuming that the proportion of cases among those tested stays the same as the original population (random and hence homogenous testing), we can replace  $S$  by  $\frac{TX}{N}$ , which will lead to the following updated equation.

$$\frac{P}{T} = \frac{X}{N} \times \alpha + \left(1 - \frac{X}{N}\right) \times (1 - \beta).$$

Solving this, we get the following expression for  $X$ .

$$X = N \times \frac{\frac{P}{T} + \beta - 1}{\alpha + \beta - 1}.$$

Thus, these two expressions give us, for a given set of  $\alpha$  and  $\beta$  regarding a test, the corrected number of reported cases ( $S$ ), and also the estimated number of true (reported + unreported) cases ( $X$ ). For the computation of  $S$ , we use  $\frac{P}{T} = \frac{109,140}{747,109} \approx 0.146$ , the test positive rate of the RT-PCR tests in Delhi as of July 10<sup>2</sup>. For the computation of  $X$ , we use  $\frac{P}{T} = \frac{4,889}{21,387} \approx 0.229$ , the positive rate reported by the Delhi serological survey<sup>7-9</sup>. Once we get these two estimates, we can compute the adjusted underreporting factor as  $URF = \frac{X}{S}$ . Also, assuming that  $D$  denotes the cumulative number of deaths till a date of interest, we can compute the corrected versions of case fatality rate and infection fatality rate as  $CFR = \frac{D}{S}$  and  $IFR = \frac{D}{X}$ , respectively. Further, if we want to adjust for a potential scenario where for every  $M$  deaths due to COVID-19, we observe 1 death ( $M$ -fold underreporting for deaths), we can update these estimates as  $CFR = \frac{MD}{S+(M-1)D}$  and  $IFR = \frac{MD}{X}$ . We computed our adjusted CFR and IFR estimates for  $M = 10$ . Based on the data till July 10 from Delhi, we use  $D = 3,300^2$ . We also used a population size of  $N = 1.98 \times 10^7$ .



A critical question here is the choice of  $\alpha$  and  $\beta$  for the two tests to ensure our computations reflect adjustments made based on sensible and realistic scenarios. Based on previously reported sensitivity and specificity levels for the diagnostic test<sup>13</sup>, we used the combinations  $\alpha = \beta = 1$  (*perfect test*),  $\alpha = 0.85$  and  $\beta = 0.99$ , and  $\alpha = 0.7$  and  $\beta = 0.99$ . The serological assay used by NCDC is a customized assay, for which we could not find any officially reported numbers. Hence, we referred to existing literature on standard serological assays<sup>14</sup> and publicly available discussions on the Delhi serosurvey<sup>15</sup>, and decided to use the combinations of  $\alpha = \beta = 1$  (*perfect test*),  $\alpha = 0.976$  and  $\beta = 0.993$ , and  $\alpha = 0.92$  and  $\beta = 0.97$ .

## CODE AVAILABILITY

All our computational codes are available at [covind19.org](https://covind19.org).

## REFERENCES

1. Hui, D. et al. The continuing 2019-nCoV epidemic threat of novel coronaviruses to global health — The latest 2019 novel coronavirus outbreak in Wuhan, China. *International Journal of Infectious Diseases* 91, 264-266 (2020).
2. Coronavirus in India: Latest Map and Case Count. *Covid19india.org* (2020). at <https://covid19india.org/>
3. Chauhan, N. After Covid-19 lockdown, plan to unlock India in phases. *Hindustan Times* (2020). at <https://www.hindustantimes.com/india-news/after-lockdown-plan-to-unlock-india-in-phases/story-vsK1wGQ7moLTMjIkKueIHP.html>
4. Hao, X. et al. Reconstruction of the full transmission dynamics of COVID-19 in Wuhan. *Nature* (2020). doi:10.1038/s41586-020-2554-8
5. Udugama, B. et al. Diagnosing COVID-19: The Disease and Tools for Detection. *ACS Nano* 14, 3822-3835 (2020).
6. Peeling, R. et al. Serology testing in the COVID-19 pandemic response. *The Lancet Infectious Diseases* (2020). doi:10.1016/s1473-3099(20)30517-x
7. Babu, N. Percentage of people with antibodies high, shows Delhi serological survey. *The Hindu* (2020). at <https://www.thehindu.com/news/cities/Delhi/percentage-of-people-with-antibodies-high/article32156162.ece>
8. Saxena, A. Explained: Here are the key takeaways from Delhi's serological survey. *The Indian Express* (2020). at <https://indianexpress.com/article/explained/delhi-serological-survey-shows-antibodies-in-23-participants-what-does-this-mean-6516512/>
9. Saxena, A. 1 in 5 in Delhi sero-survey developed antibodies, indicating spread of virus. *The Indian Express* (2020). at <https://indianexpress.com/article/cities/delhi/delhi-covid-19-sero-survey-antibodies-6517379/>

10. Ioannidis, J. The infection fatality rate of COVID-19 inferred from seroprevalence data. (2020). doi:10.1101/2020.05.13.20101253
11. Roy, L. Infected India: The True Toll Of Coronavirus In The World's 2nd-Most Populated Country. *Forbes* (2020). at <https://www.forbes.com/sites/lipiroy/2020/06/25/infected-india-the-true-toll-of-coronavirus-in-the-worlds-2nd-most-populated-country/#4cf904c850fb>
12. Burden of Influenza. *Centers for Disease Control and Prevention* (2020). at <https://www.cdc.gov/flu/about/burden/index.html>
13. Woloshin, S., Patel, N. & Kesselheim, A. False Negative Tests for SARS-CoV-2 Infection — Challenges and Implications. *New England Journal of Medicine* (2020). doi:10.1056/nejmp2015897
14. Tran, N., Cohen, S., Waldman, S. & May, L. Review of COVID-19 Testing Methods. *Laboratory Best Practice Blog* (2020). at <https://blog.ucdmc.ucdavis.edu/labbestpractice/index.php/2020/06/16/review-of-covid-19-testing-methods/>
15. The Print India. *Serosurveys - Pure Science*. (2020). at <https://www.facebook.com/1733495223546925/posts/3183938748502558/>
16. Mandal, S., Das, H., Deo, S. & Arinaminpathy, N. When to relax a lockdown? A modelling-based study of testing-led strategies coupled with sero-surveillance against SARS-CoV-2 infection in India. (2020). doi:10.1101/2020.05.29.20117010
17. Xu, X. et al. Seroprevalence of immunoglobulin M and G antibodies against SARS-CoV-2 in China. *Nature Medicine* (2020). doi:10.1038/s41591-020-0949-6
18. Korth, J. et al. SARS-CoV-2-specific antibody detection in healthcare workers in Germany with direct contact to COVID-19 patients. *Journal of Clinical Virology* 128, 104437 (2020).
19. Barnagarwala, T. Coronavirus: What Mumbai serosurvey shows about gender differences in infection, mortality and herd immunity. *The Indian Express* (2020). at <https://indianexpress.com/article/explained/mumbais-serosurvey-what-it-shows-about-gender-differences-in-infection-mortality-and-herd-immunity-6529186/>
20. Shakiba, M. et al. Seroprevalence of COVID-19 virus infection in Guilan province, Iran. (2020). doi:10.1101/2020.04.26.20079244
21. Doi, A. et al. Estimation of seroprevalence of novel coronavirus disease (COVID-19) using preserved serum at an outpatient setting in Kobe, Japan: A cross-sectional study. (2020). doi:10.1101/2020.04.26.20079822

22. Pollán, M. et al. Prevalence of SARS-CoV-2 in Spain (ENE-COVID): a nationwide, population-based seroepidemiological study. *The Lancet* (2020). doi:10.1016/s0140-6736(20)31483-5
23. Public Health Agency Sweden. *Första resultaten från pågående undersökning av antikroppar för covid-19-virus*. (2020).
24. Stringhini, S. et al. Seroprevalence of anti-SARS-CoV-2 IgG antibodies in Geneva, Switzerland (SEROCoV-POP): a population-based study. *The Lancet* (2020). doi:10.1016/s0140-6736(20)31304-0
25. Sood, N. et al. Seroprevalence of SARS-CoV-2–Specific Antibodies Among Adults in Los Angeles County, California, on April 10-11, 2020. *JAMA* 323, 2425 (2020).
26. Rosenberg, E. et al. Cumulative incidence and diagnosis of SARS-CoV-2 infection in New York. *Annals of Epidemiology* (2020). doi:10.1016/j.annepidem.2020.06.004
27. Ng, D. et al. SARS-CoV-2 seroprevalence and neutralizing activity in donor and patient blood from the San Francisco Bay Area. (2020). doi:10.1101/2020.05.19.20107482
28. Bendavid, E. et al. COVID-19 Antibody Seroprevalence in Santa Clara County, California. (2020). doi:10.1101/2020.04.14.20062463
29. Dutt, A. Community transmission of Covid-19 for Centre to declare: Delhi health minister. *Hindustan Times* (2020). at <https://www.hindustantimes.com/india-news/community-transmission-has-begun-but-up-to-centre-to-declare-satyendar-jain/story-GPOYz6P2CFECADgfbjmeO.html>
30. Britton, T., Ball, F. & Trapman, P. A mathematical model reveals the influence of population heterogeneity on herd immunity to SARS-CoV-2. *Science* (2020). doi: 10.1126/science.abc6810
31. Randolph, H. & Barreiro, L. Herd Immunity: Understanding COVID-19. *Immunity* 52, 737-741 (2020).
32. Kirkcaldy, R., King, B. & Brooks, J. COVID-19 and Postinfection Immunity. *JAMA* 323, 2245 (2020).
33. COVID-19 Vaccine and Therapeutic Drugs Tracker. *COVID-19 Vaccine and Therapeutic Drugs Tracker* (2020). at <https://biorender.com/covid-vaccine-tracker>
34. Hastings, W. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* 57, 97-109 (1970).

## **ACKNOWLEDGEMENTS**

The authors would like to thank the Center for Precision Health Data Sciences at the University of Michigan School of Public Health, The University of Michigan Rogel Cancer Center and the Michigan Institute of Data Science for internal funding that supported this research. The authors are grateful to Professors Eric Fearon, Aubree Gordon and Parikshit Ghosh for useful conversations that helped formulating the ideas in this manuscript.

## **AUTHOR CONTRIBUTIONS**

Rupam Bhattacharyya prepared the initial draft and carried out the naïve misclassification correction to case counts and took leadership of composing the final draft. Ritwik Bhaduri, Ritoban Kundu and Bhramar Mukherjee developed the extended SEIR model with misclassification. Ritwik Bhaduri and Ritoban Kundu implemented the extended SEIR model. Maxwell Salvatore carried out extensive literature search and visualization and participated in writing of the manuscript. Bhramar Mukherjee conceptualized the project and oversaw the research.

## **COMPETING INTERESTS**

No competing interest.

## **MATERIALS AND CORRESPONDENCE**

All correspondences should be directed to Bhramar Mukherjee [bhramar@umich.edu](mailto:bhramar@umich.edu). All data and code are available at [covind19.org](http://covind19.org).

## TABLES

**Table 1. Summary of extended SIR model results for Delhi.** Predicted cumulative case and death counts and corresponding underreporting factors with respect to the observed data are presented for July 10. The specificity of the RT-PCR test is assumed to be 1. The observed number of cumulative cases and deaths in Delhi on July 10 were taken to be 109,140 and 3,300 respectively, according to [covid19india.org](https://covid19india.org).

<b>Sensitivity of RT-PCR Test</b>	<b>Predicted Reported Cases</b>	<b>Predicted Total Cases</b>	<b>Under-reporting Factor for Cases</b>	<b>Predicted Reported Deaths</b>	<b>Predicted Total Deaths</b>	<b>Under-reporting Factor for Deaths</b>
<b>0.7</b>	119,920	6,318,663	52.6	3,386	43,978	12.9
<b>0.85</b>	119,879	4,780,982	39.8	3,384	33,165	9.8
<b>1</b>	119,603	4,164,568	34.8	3,376	28,499	8.4

**Table 2. Summary of corrected number of cases, estimated underreporting factor, case-fatality rate based on reported cases and infection-fatality rate across different testing scenarios.** Population size of Delhi is collected from <https://censusindia.gov.in/>, and the testing, infection, recovery and fatality data are taken from <https://covid19india.org/>.

Diagnostic test		Antibody test (past infection)					
		RT-PCR		Serology test			
		Specificity	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity
1	1	0.993	0.976	0.97	0.92		
<b>1</b>	<b>1</b>	4,526,217	4,527,984	4,418,221	<b>Est. # true infections</b>		
		109,140	109,140	109,140	<b>Corrected # reported cases</b>		
		41.5x	41.5x	40.5x	<b>URF</b>		
		0.0302	0.0302	0.0302	<b>CFR</b>		
		0.0007 (0.0073)	0.0007 (0.0073)	0.0007 (0.0075)	<b>IFR (10x adj.)</b>		
<b>0.99</b>	<b>0.85</b>	4,526,217	4,527,984	4,418,221	<b>Est. # true cases</b>		
		121,034	121,034	121,034	<b>Corrected # reported</b>		
		37.4x	37.4x	36.5x	<b>URF</b>		
		0.0273	0.0273	0.0273	<b>CFR</b>		
		0.0007 (0.0073)	0.0007 (0.0073)	0.0007 (0.0075)	<b>IFR (10x adj.)</b>		
<b>0.99</b>	<b>0.7</b>	4,526,217	4,527,984	4,418,221	<b>Est. # true cases</b>		
		147,346	147,346	147,346	<b>Corrected # reported</b>		
		30.7x	30.7x	30.0x	<b>URF</b>		
		0.0224	0.0224	0.0224	<b>CFR</b>		
		0.0007 (0.0073)	0.0007 (0.0073)	0.0007 (0.0075)	<b>IFR (10x adj.)</b>		

Abbrev: *adj.*, adjusted; *CFR*, case-fatality rate; *est.*, estimated; *IFR*, infection-fatality rate; *URF*, underreporting factor.  
Notes: The URF is the ratio of the estimated number of true cases and the corrected number of reported cases. For the IFR, we report the estimate if we adjusted for 10x death underreporting (10x adj.).



**Table 3. Summary of COVID-19 seroprevalence studies.**

Location	Study Design	Sample Size	Estimated Seroprevalence % (95% CI)	Reference
Hubei and Guangdong Provinces <i>China</i>	Cohort and location-specific surveys (Healthcare workers and their relatives, hospital outpatients, factory workers, hotel staff).	6,919 (Hospital settings) 10,449 (Community settings)	3.8 (2.6, 5.4) <i>Healthcare workers, Wuhan</i> 3.8 (2.2, 6.3) <i>Hotel staff members, Wuhan</i> 3.2 (1.6, 6.4) <i>Family members, Wuhan</i>	Xu et al. (2020) <i>Nature Medicine</i> doi: 10.1038/s41591-020-0949-6
Essen <i>Germany</i>	Prospective cross-sectional monocentric study recruiting healthcare workers from University Hospital Essen.	316	1.6	Korth et al. (2020) <i>Journal of Clinical Virology</i> doi: 10.1016/j.jcv.2020.104437
India	Pilot survey in 83 districts across 21 states.	Unknown	0.73 <i>overall</i> 1.09 <i>urban</i>	<i>The Indian Express</i> (2020) url: <a href="https://indianexpress.com/article/explained/delhi-serological-survey-shows-antibodies-in-23-participants-what-does-this-mean-6516512/">https://indianexpress.com/article/explained/delhi-serological-survey-shows-antibodies-in-23-participants-what-does-this-mean-6516512/</a>
Mumbai <i>India</i>	Consent-based survey across three wards with high COVID-19 growth and proximity to hotspots.	6,936 (Out of 8,800 invited)	40.5 <i>overall</i> 57.8 <i>slum areas</i> 16.0 <i>non-slum areas</i>	<i>The Indian Express</i> (2020) url: <a href="https://indianexpress.com/article/explained/mumbai-serosurvey-what-it-shows-about-gender-differences-in-infection-mortality-and-herd-immunity-6529186/">https://indianexpress.com/article/explained/mumbai-serosurvey-what-it-shows-about-gender-differences-in-infection-mortality-and-herd-immunity-6529186/</a>
Guilan Province <i>Iran</i>	Population-based cluster random sampling through phone call invitations.	552 (196 households)	0.22 (0.19, 0.26) <i>unadjusted</i> 0.33 (0.28, 0.39) <i>adjusted for imperfect testing</i> 0.21 (0.14, 0.29) <i>adjusted by population weights</i>	Shakiba et al. (2020) <i>medRxiv</i> doi: 10.1101/2020.04.26.20079244
Kobe City <i>Japan</i>	Cross-sectional study on hospital outpatients.	1,000	3.3 (2.3, 4.6)	Doi et al. (2020) <i>medRxiv</i> doi: 10.1101/2020.04.26.20079822
Spain	Two-stage random sampling of households stratified by province and municipality size.	61,075 (Point-of-care test) 51,958 (Immunoassay) (35,883 households)	5.0 (4.7, 5.4) <i>point-of-care test</i> 4.6 (4.3, 5.0) <i>immunoassay</i>	Pollán et al. (2020) <i>The Lancet</i> doi: 10.1016/S0140-6736(20)31483-5
Sweden (9 Regions)	Consecutive weekly region-specific surveys.	1,200 (Per week).	7.3 <i>Stockholm</i> 4.2 <i>Skåne</i> 3.7 <i>Västra Götaland</i>	Public Health Agency Sweden (2020) url: <a href="https://www.folkhalsomyndigheten.se/nyheter-och-press/nyhetsarkiv/2020/maj/forsta-resultaten-fran-pagaende-undersokning-av-antikroppar-for-covid-19-virus/">https://www.folkhalsomyndigheten.se/nyheter-och-press/nyhetsarkiv/2020/maj/forsta-resultaten-fran-pagaende-undersokning-av-antikroppar-for-covid-19-virus/</a>
Geneva <i>Switzerland</i>	Series of 5 consecutive weekly serosurveys among randomly selected participants from a previous population-representative survey, and their household members aged 5 years and older.	2,766 (1,339 households; 341, 469, 577, 604 and 775 samples respectively in weeks 1-5.)	4.8 (2.4, 8.0) <i>week 1</i> 8.5 (5.9, 11.4) <i>week 2</i> 10.9 (7.9, 14.4) <i>week 3</i> 6.6 (4.3, 9.4) <i>week 4</i> 10.8 (8.2, 13.9) <i>week 5</i>	Stringhini et al. (2020) <i>The Lancet</i> doi: 10.1016/S0140-6736(20)31304-0
LA County, California <i>USA</i>	Invited enrollment, based on demographic match and geographical proximity to the testing centers.	863 (Out of 1952 invited)	4.06 (2.84, 5.60) <i>unadjusted</i> 4.34 (2.76, 6.07) <i>adjusted for imperfect testing</i>	Sood et al. (2020) <i>JAMA</i> doi: 10.1001/jama.2020.8279

New York State USA	Convenience sampling of New Yorkers attending 99 grocery stores across 26 counties, containing 87.3% of the state's population, located all across the state.	15,101	14.0 (13.3, 14.7) <i>overall</i> 22.7 (21.5, 24.0) <i>New York City</i>	Rosenberg et al. (2020) <i>Annals of Epidemiology</i> doi: 10.1016/j.annepidem.2020.06.004
San Francisco Bay Area USA	Cohort-based recruitment of non-COVID patients and blood donors.	387 (Non-COVID patients) 1,000 (Blood donors)	0.26 (0.00, 0.76) <i>non-COVID patients</i> 0.10 (0.00, 0.56) <i>blood donors</i>	Ng et al. (2020) <i>medRxiv</i> doi: 10.1101/2020.05.19.20107482
Santa Clara County, California USA	Ad-based recruitment, matched on geographic location and demographics.	3,330	1.5 (1.1, 2.0) <i>unadjusted</i> 1.2 (0.7, 1.8) <i>adjusted for imperfect testing</i> 2.8 (1.3, 4.7) <i>adjusted for county demographics</i>	Bendavid et al. (2020) <i>medRxiv</i> doi: 10.1101/2020.04.14.20062463

**Table 4. Description of extended SEIR model parameters.**

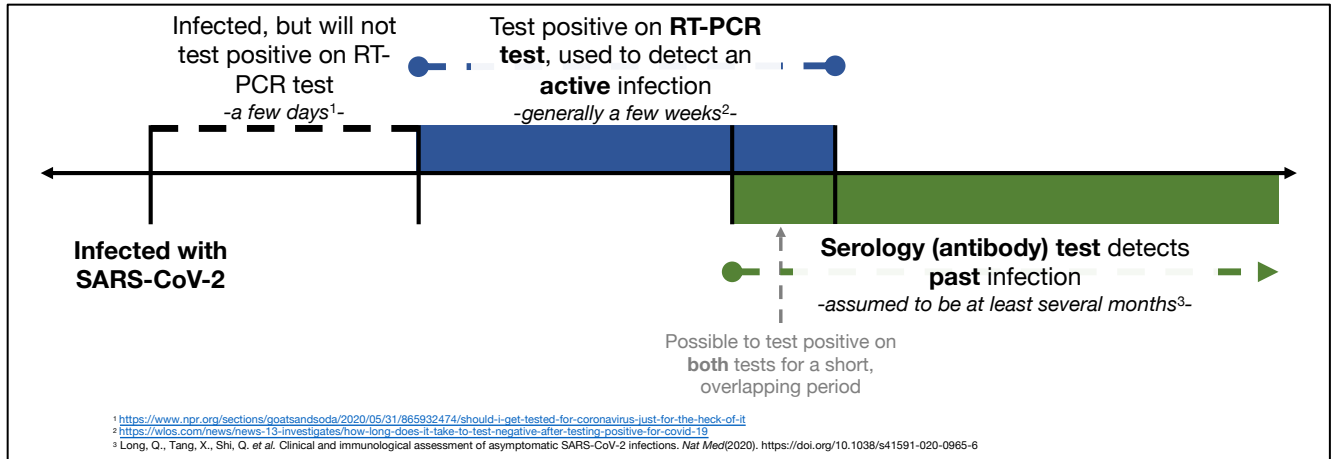
Parameter	Value	Description
$\beta$	<i>Time-varying</i>	Rate of infectious transmission by infected, tested individuals with false negative results.
$\alpha_p$	0.5	Ratio of rate of spread of infection by tested positive patients to that by false negatives. $\alpha_p < 1$ represents the scenario where individuals who test positive are infecting susceptible individuals at a lower rate than infected individuals with false negative test results.
$\alpha_u$	0.7	Scaling factor for the rate of spread of infection by untested individuals. $\alpha_u$ is assumed to be $< 1$ as U mostly consists of asymptomatic or mildly symptomatic cases who are known to spread the disease at a much lower rate than those with higher levels of symptoms.
$D_e$	5.2	Incubation period (in days).
$D_r$	17.8	Means number of days until recovery for infected individuals.
$D_t$	0	Mean number of days for the test result to come after a person is tested. Under the assumption of instantaneous test results, this is taken to be zero.
$\mu_c$	0.0562	Death rate attributable to COVID-19 which is equivalent to inverse of the average number of days for death starting from the onset of disease times the probability of death of an infected individual.
$\lambda, \mu$	$3.95 \times 10^{-5}$	Natural birth and death rates (assumed to be equal).
$r$	<i>Time-varying</i>	Probability of being tested for infectious individuals.
$f$	0.3, 0.15, 0	Probability of a false negative RT-PCR diagnostic test result.
$\beta_1, \frac{1}{\beta_2}$	0.6 ( $\beta_1$ ) 0.7 ( $\beta_2$ )	Scaling factors for rate of recovery for undetected and false negative individuals respectively. Both $\beta_1$ and $\beta_2$ are assumed to be less than 1. It is assumed that the recovery rate is slower than the detected ones for the False Negative ones because they are not getting any hospital treatments. The condition of Untested individuals is not so severe as they consist of mostly asymptomatic people. So, they are assumed to recover faster than the Current Positive Ones.
$\delta_1, \frac{1}{\delta_2}$	0.3 ( $\delta_1$ ) 0.7 ( $\delta_2$ )	Scaling factors for death rate for undetected and false negative individuals respectively. Both $\delta_1$ and $\delta_2$ are assumed to be less than 1. Same as before, the death rate for False Negative ones are assumed to be higher than the Current detected Positive as they are not receiving proper treatment. While, for the Untested ones, the death rate is taken to be lesser because they are mostly asymptomatic. So, their probability of dying is much less.

**Table 5. Training data periods according to interventions.**

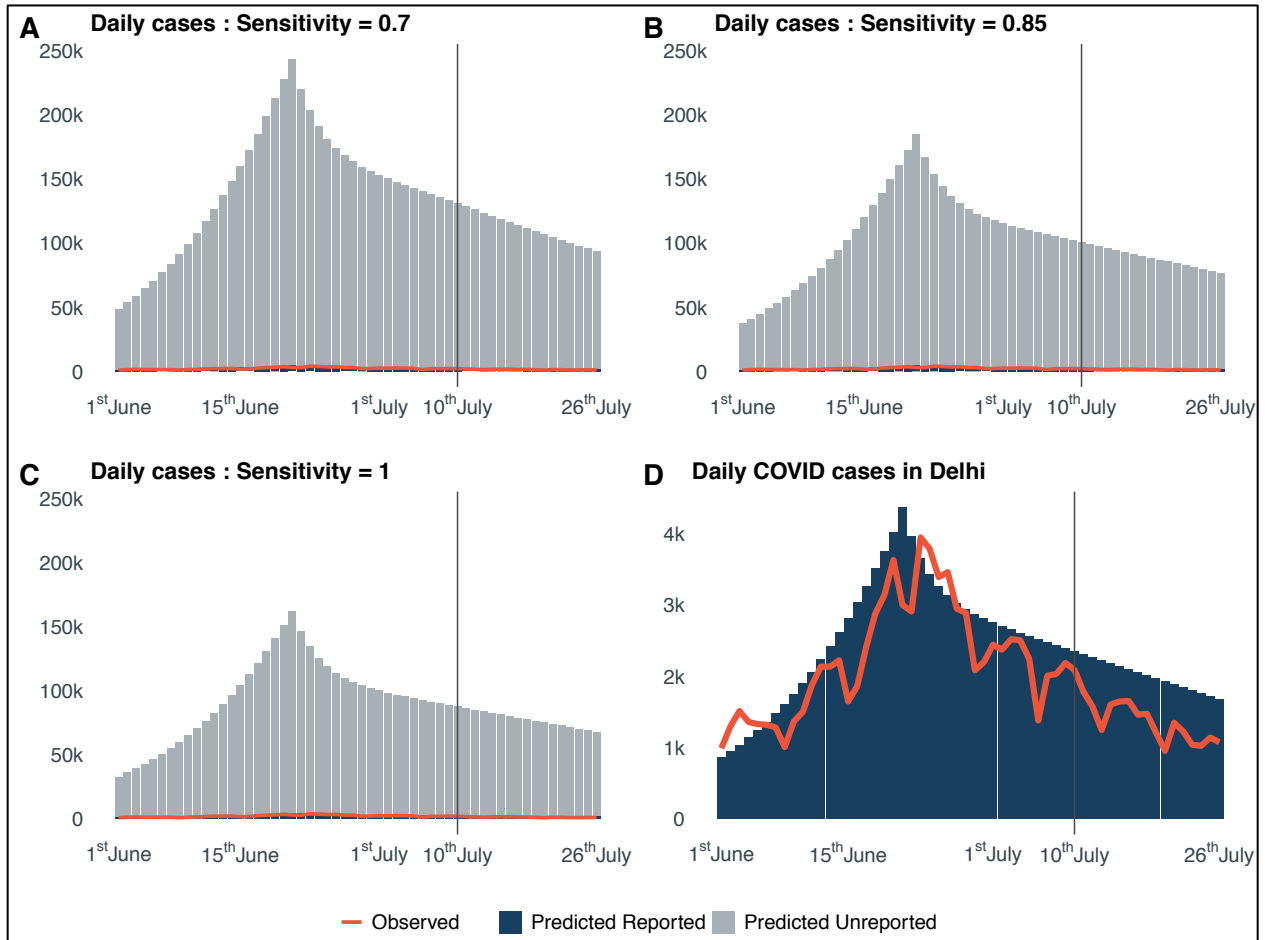
Pre-lockdown	March 15 – March 24
Lockdown 1.0	March 25 – April 14
Lockdown 2.0	April 15 – May 3
Lockdown 3.0	May 4 – May 17
Lockdown 4.0	May 18 – May 31
Unlock 1.0	June 1 – June 30 (This period is divided into two parts: June 1 – June 19 and June 20 – June 30, since there was a high increase in the number of tests on June 20.)

## FIGURES

**Figure 1. Timeline of COVID-19 diagnostic and antibody tests with respect to the infection and immune response time frame.**

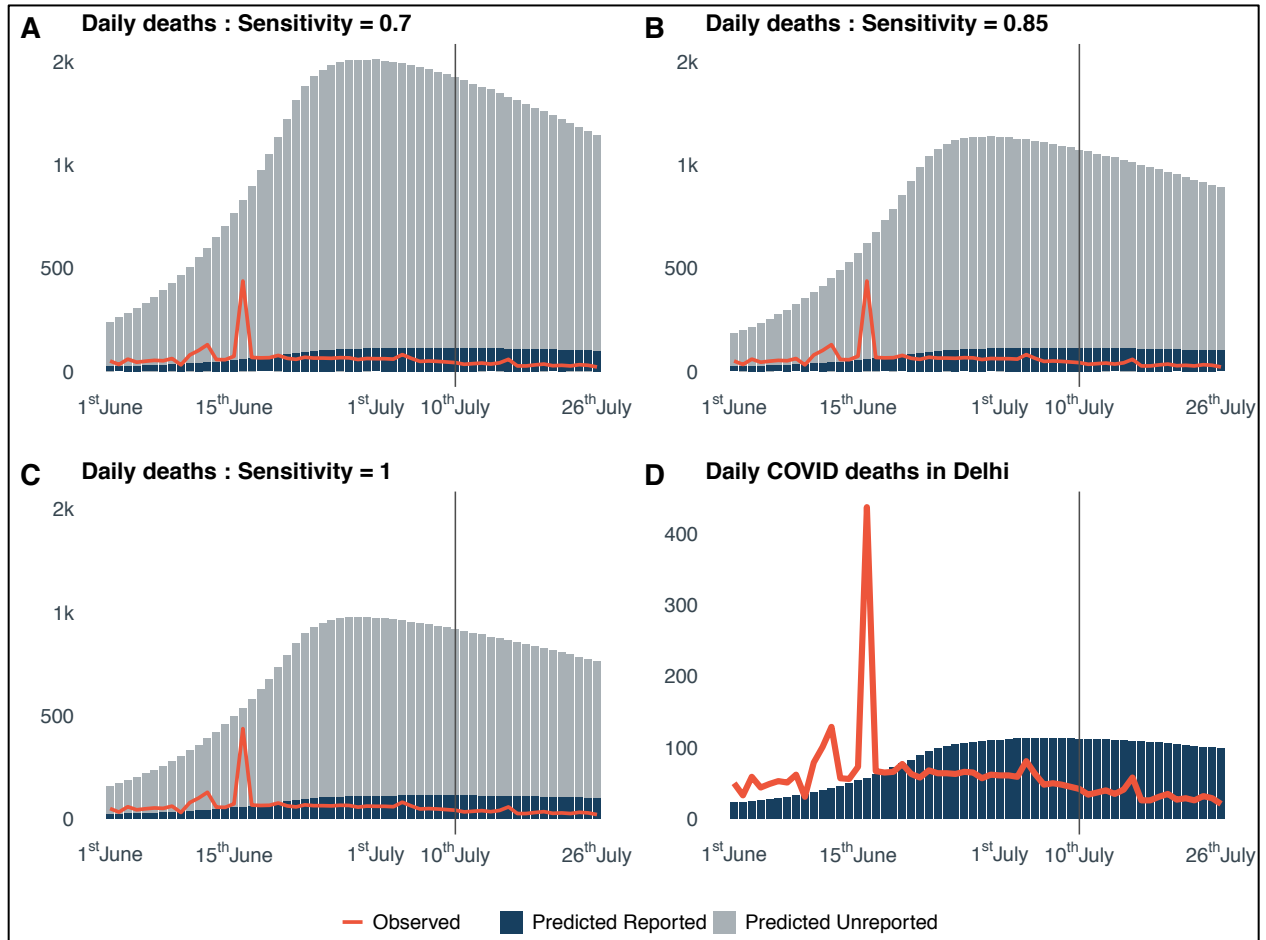


**Figure 2. Bar plots of predicted reported and unreported daily cases from June 1 to July 26.** Panels A, B and C depict the predictions under assumed sensitivity of the diagnostic test at 0.7, 0.85 and 1, respectively. Panel D shows the consistency of the predictions with the observed data. The specificity of the diagnostic test is assumed to be 1.

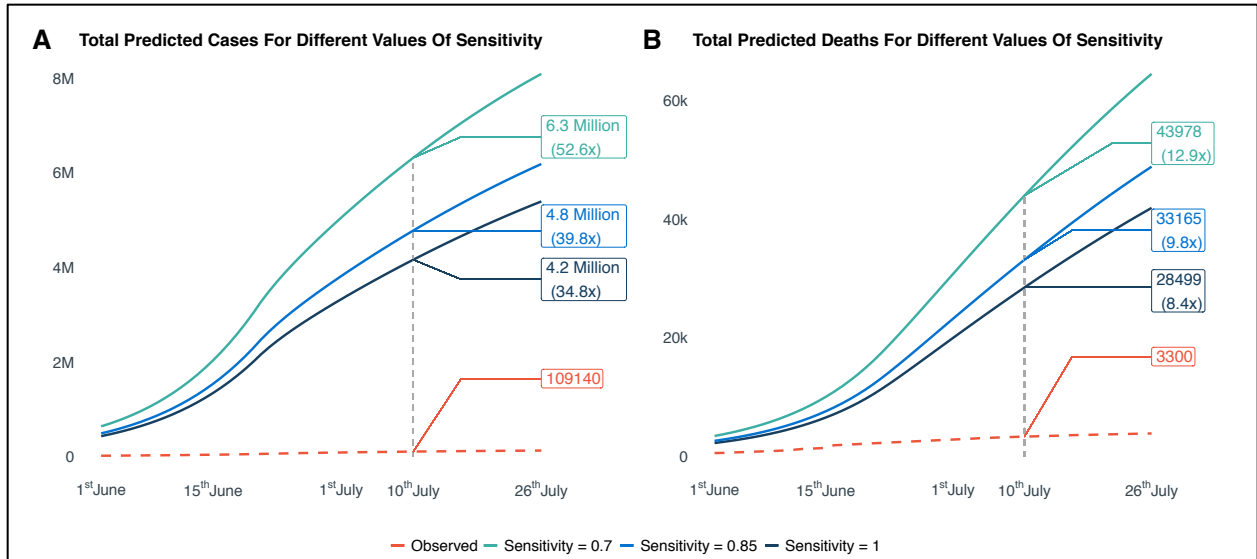




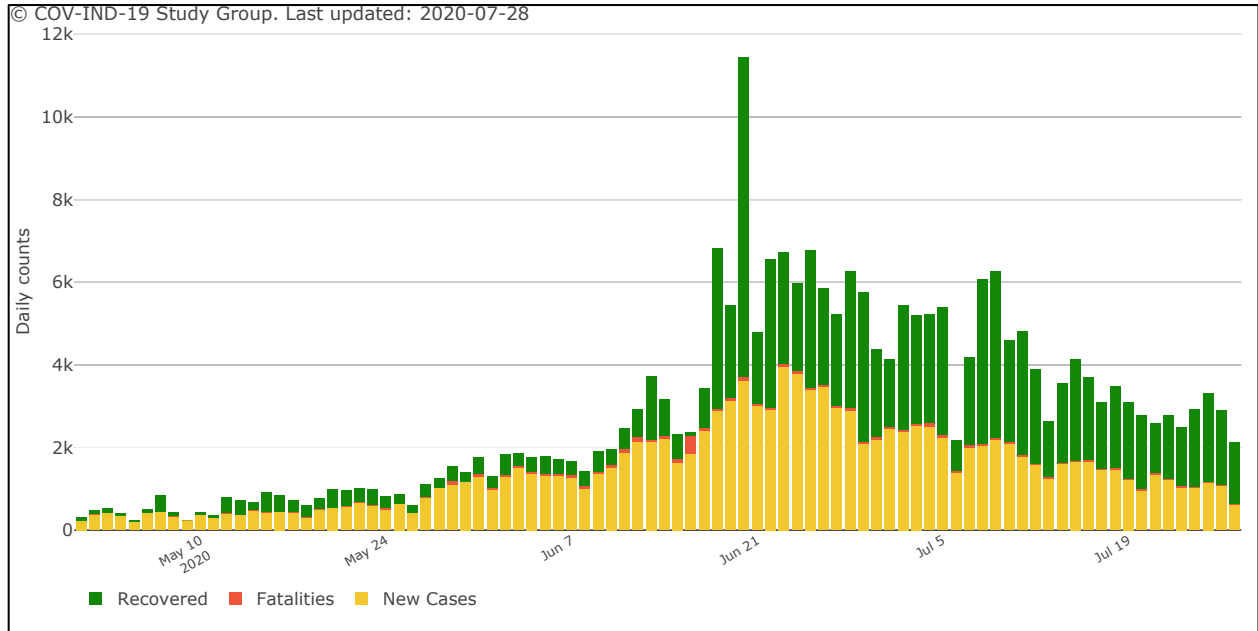
**Figure 3. Bar plots of predicted reported and unreported daily deaths from June 1 to July 26.** Panels A, B and C depict the predictions under assumed sensitivity of the diagnostic test at 0.7, 0.85 and 1, respectively. Panel D shows the consistency of the predictions with the observed data. The specificity of the diagnostic test is assumed to be 1.



**Figure 4. Summary of cumulative total (reported and unreported) cases and deaths for three different assumed values of specificity for the diagnostic test: 0.7, 0.85, 1. Panel A and B respectively summarize the cases and deaths, along with their reported observed counterparts. The specificity of the diagnostic test is assumed to be 1.**



**Figure 5. Observed daily COVID-19 case, recovery and fatality counts for Delhi during May 1 – July 28.**



**Figure 6. Estimated time-varying R (and 95% posterior credible interval) for COVID-19 in Delhi.** Computations were performed using the R package EpiEstim.

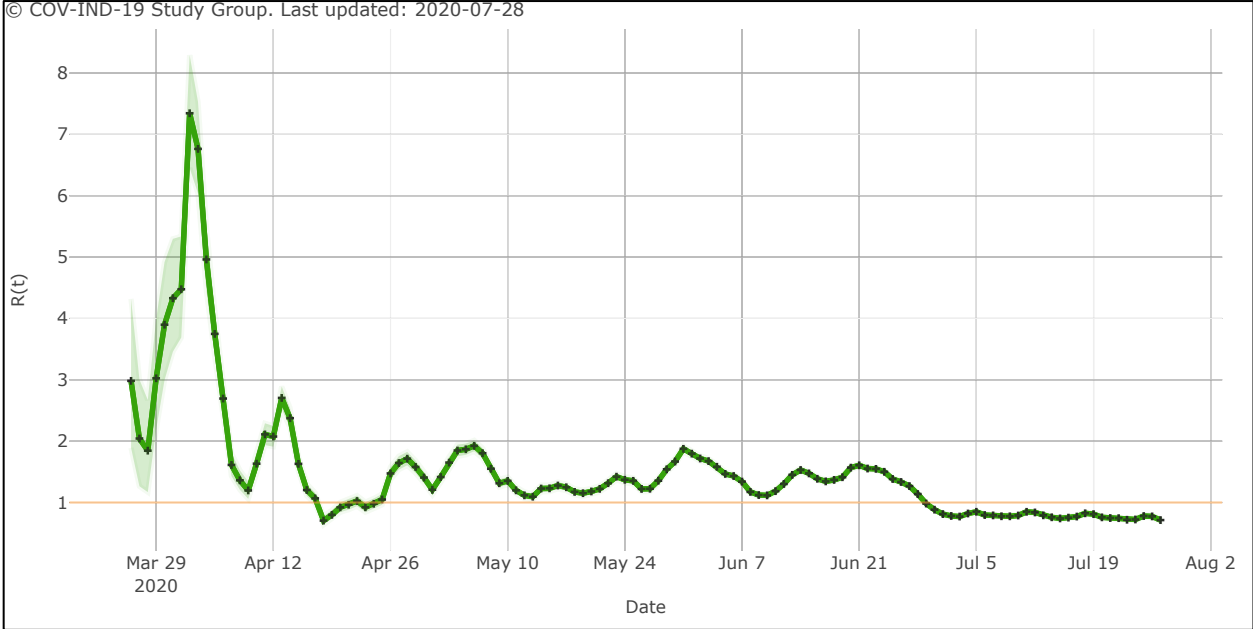
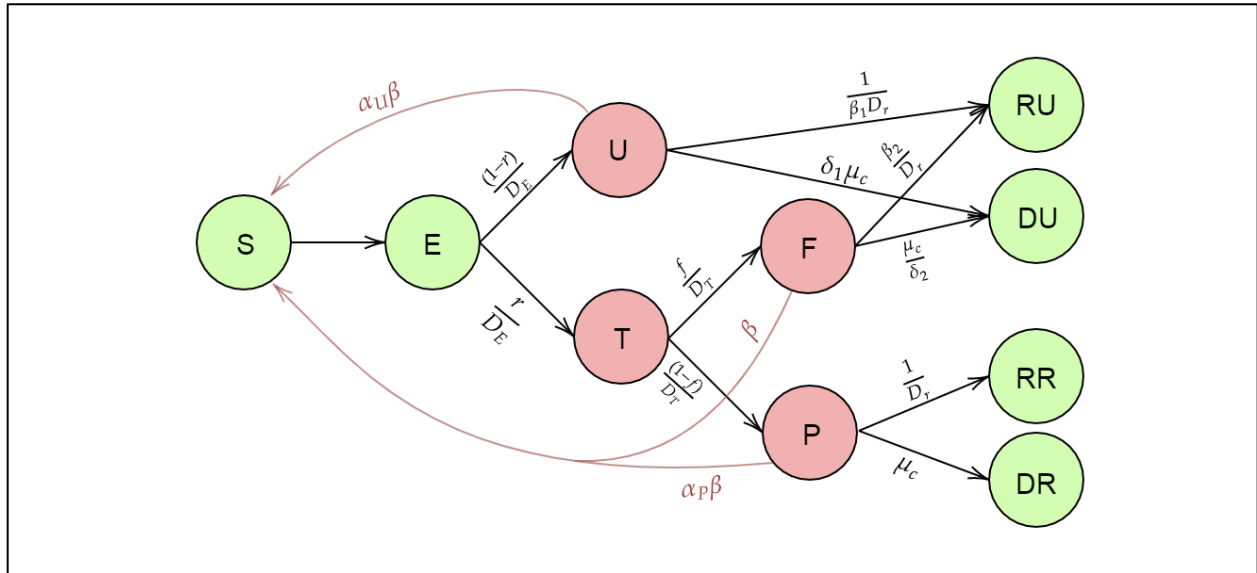


Figure 7. Diagram of model compartments and transmissions for the extended SEIR model.



**Figure 8. Diagram of testing decisions.** Dark lines indicate the break-up of the population in terms of true infection status; green and red lines indicate (correct and incorrect, respectively) decisions based on testing procedure. Here, we have referred to the diagnostic test, and hence, the active infection status. Similar framework applies to the antibody test and the past infection status.

