Estimation of undetected COVID-19 infections in India

Siuli Mukhopadhyay[1] & Debraj Chakraborty[2]

[1] *Department of Mathematics,* [2] *Department of Electrical Engineering*
Indian Institute of Technology Bombay, India

Background and Objectives: **While the number of detected COVID-19 infections are widely available, an understanding of the extent of undetected COVID-19 cases is urgently needed for an effective tackling of the pandemic and as a guide to lifting the lockdown. The aim of this work is to estimate and predict the true number of COVID-19 (detected and undetected) infections in India for short to medium forecast horizons. In particular, using publicly available COVID-19 infection data upto 16th April 2020, we predict the true number of infections in India during and upto the end of the formal lockdown period (21st April 2020).**

Methods: **The high death rate observed in most COVID-19 hit countries is suspected to be a function of the undetected infections existing in the population. An estimate of the age weighted infection fatality rate (IFR) of the disease of** 0.41%**, specifically calculated by taking into account the age structure of Indian population, is already available in the literature. In addition, the recorded case fatality rate (CFR=** 0.70%**) of Kerala, the only state in India to report single digit new infections over the second week of April, is used as a second estimate of the IFR. These estimates are used to formulate a relationship between deaths recorded and the true number of infections. The estimated undetected and detected cases time series based on these two IFR estimates are then used to fit a discrete time multivariate infection model to predict the total infections at the end of the formal lockdown period.**

Results: **In two consecutive fortnights during the lockdown, it was noted that the rise in detected infections has decreased by 2.7 times. For an IFR of** 0.41%**, the rise in undetected infections decreased by 3.2 times and the predicted number of total infections in India is 3.14 lakhs. While for an IFR of** 0.70%**, the rise in undetected cases decreased by 3.3 times and the total number of infections predicted on 21st April is 1.75 lakhs.**

Interpretation and Conclusions: **The behaviour of the undetected cases over time effectively illustrates the effects of lockdown and increased testing. From**

our estimates, it is found that the lockdown has brought down the undetected to detected cases ratio, and has consequently dampened the increase in the number of total cases. However, even though the rate of rise in total infections has fallen, the lifting of the lockdown should be done keeping in mind that 1.75 to 3 lakhs undetected cases will already exist in the population on 21st April.

Key Word: Discrete time - infection model - infection fatality rate - lockdown

In this article, we propose a discrete time multivariate infection model for predicting the total true number of COVID-19 infections for a short to medium forecast horizon. Using publicly available COVID-19 data for India upto 16th April 2020, we predict the true number of infections in India during and upto the end of active lockdown (21st April 2020) period. For successful prediction of the total infections, we require an estimate of the extent of cases escaping detection. An estimate of the infection fatality rate (ratio of total deaths to total infections) for the Indian population has been calculated recently (Bommer and Vollmer (2020)). Assuming this rate to be constant, we determine estimates of the undetected infections for each day during the lockdown period. This time series data of undetected infections, along with recorded data for infected, recovered and deceased cases available from https://www.covid19india.org/, is used to fit a multivariate discrete time auto regressive (AR(1)) infection model. Using data upto 16th April, this model is used to predict both to-be-detected and to-be-undetected infections in India upto 21st April.

The low detected infection numbers reported by a densely populated country like India is a highly debated subject with no known clear explanation. Many sources are attributing the low infection rates to the low number of tests being conducted for a country with 1.38 billion population. Bommer and Vollmer (2020) assessed the Indian COVID-19 data and suggested that India is detecting only 1.68% of the total number of infections. While, Srinivas and James (2020) concludes a 3.6% detection rate with wide variation among the states. Their assessment of low detection rate is based on the fact that the Indian case fatality rate (CFR = total number of deaths divided by total detected infections) is poorly estimating the true infection fatality rate, due to a large number of undetected infections in India. Using a recent study by Verity et al. (2020) based on age stratified fatality data from mainland China and international Wuhan residents returning on repatriation flights, Bommer and Vollmer (2020) calculate an infection fatality rate (IFR) of 0.41% specifically for India. The IFR was calculated for India using population data from UN to correct for differences in age distributions in China and India. In our study we propose to use 0.41% as the first estimate of the IFR. We obtain a second estimate of the IFR for India from cumulative death and infections data for the state of Kerala. Kerala has been exceptionally successful in reducing the number of new infections. This is only possible through successful tracking and isolation of every COVID-19 infection in the state. From this observations we argue, that Kerala does not have any undetected

infections and consequently, for Kerala the CFR is equal to the IFR. The CFR for Kerala is calculated at 0.70%. Hence this number is used as our second estimate of national IFR.

We propose a 2-equation discrete time AR(1)/state space model for the infection, death and recovered population dynamics. This is similar to the discrete time SIR model of Allen (1994) however with two major changes. Firstly, we ignore the S (susceptible population) equation. Since S is significantly larger as compared to current infection proliferation, we assume S is almost constant. Secondly, we let the coefficients of the model to vary linearly with time. We use these coefficient variations to model various forms of interventions such as lockdown, increased testing, etc.

Materials and Methods

The prediction of total infections till the end of active lockdown is broken into two steps. Firstly, we estimate the total number of undetected cases using the IFR for the period for which data on deaths are available. In the second step, this time series data of undetected infections, along with recorded data for infected, recovered and deceased are used in a multivariate discrete time auto regressive (AR(1)) infection model for predicting the total infections into the future.

Step 1: Estimation of undetected cases based on IFR estimates upto the present time: We describe an estimation procedure for the total number of undetected infections upto present time t (in our case 16th April). We denote,
I_t: detected infections upto time t,
D_t: deaths upto time t
R_t: recoveries upto time t
A_t: undetected infections upto time t.
Thus, the total number of infections recorded till time t is $(I_t + A_t)$. Note, from our data set, we have observed values of I_t, R_t and D_t. However A_t is unknown for all values of t.

Verity et al. (2020) reports that the average time from symptom onset in COVID-19 to death to be approximately 14 - 18 days. Assuming, total deaths till t depends on the total infections recorded up to two weeks ago, we use a similar relation as used by Bommer and Vollmer (2020),

$$I_t + A_t = \frac{100 \times D_{t+14}}{IFR}, \text{ to calculate } A_t. \tag{1}$$

However, this formula cannot be used directly for dates for which we do not yet have the corresponding D_{t+14} values. For these dates, we utilise the variation of the ratio A_t/I_t, denoted by K_t hereafter, with time to estimate A_t. In the Results section we see that a linear fit to K_t, denoted by K_t^l, gives a good approximation of the variation in K_t. Using the fitted K_t^l and I_t from our data set, estimates of A_t can be computed upto the present time.

Step 2: Proposed Infection Model for Prediction of Undetected Cases: The previous step provides us with estimates of A_t upto present time. However, we would like to predict the number of cases in the near future (in particular upto the end of formal lockdown - 21 April). To make this possible, we need to have a method for predicting I_t, D_t and R_t upto 21st April. For reasons mentioned in the Introduction, for this short prediction horizon we prefer not to use a conventional SIR/SEIR type epidemiological model.

For prediction purposes, we model the relationship between the total number of infections (detected and undetected cases) and total deaths and recovered counts, using time varying coefficients upto time $t+1$ as;

$$I_{t+1} + A_{t+1} = (1+\beta_t)(I_t + A_t) - \beta_t(D_t + R_t), \quad (2)$$

$$D_{t+1} + R_{t+1} = (1-\gamma_t)(D_t + R_t) + \gamma_t(I_t + A_t), \quad (3)$$

where β_t and γ_t are unknown model coefficients changing with time. Note that the state variable $I_t + A_t$, is the cumulative count of total infections (active+dead+recovered) as opposed to cumulative active infections, which is frequently used in discrete time epidemiological models (Allen 1994). Due to our use of the IFR to estimate A_t we find this version more convenient.

The assumption of time varying model coeffcients β_t and γ_t is necessary both for a good model fit to the data, and for explaining the effect of interventions such as lockdown/increased testing. Using the relation $A_t = K_t I_t$ we can express β_t and γ_t as,

$$\beta_t = \frac{(1+K_{t+1})I_{t+1} - (1+K_t)I_t}{(1+K_t)I_t - (D_t + R_t)}$$

$$\gamma_t = \frac{(D_{t+1} + R_{t+1}) - (D_t + R_t)}{(1+K_t)I_t - (D_t + R_t)}$$

The values of β_t and γ_t over time t for the lockdown period are plotted and linear curves are fitted to both β_t and γ_t. Using these fitted values, denoted by β_t^l and γ_t^l, we find the predicted total number of infections $(I_t + A_t)$ and total deaths and recoveries $(D_t + R_t)$ by iterating (2) and (3) for required number of days into the future.

Results

Results from Step 1: We use two IFR estimates. The first estimate is taken as 0.41% from Bommer and Vollmer (2020). The CFR of Kerala is taken as the second estimate of the national IFR, which is 0.70%. This CFR is found by dividing the total number of deaths in Kerala upto 16th April (= 2) by total detected infections upto 2nd April (= 286).

Based on these two IFRs, A_t values for t (4th March - 2nd April) are obtained directly from (1) and using D_t recorded upto 16th April. Ratios of $A_t/I_t (= K_t)$ over time t (4th March

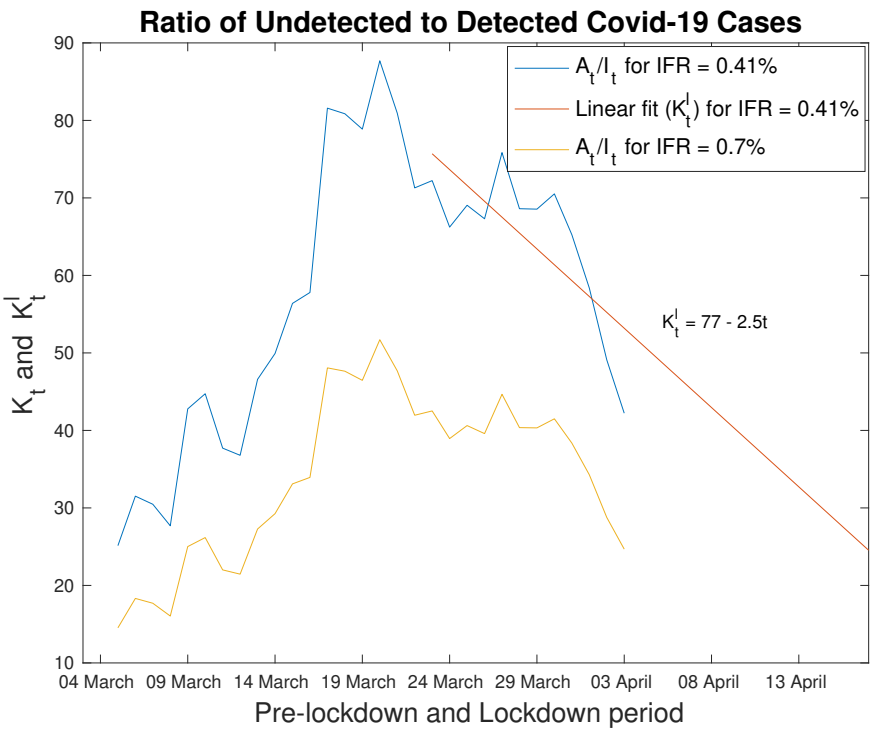

Figure 1: Estimated Ratio A_t/I_t for IFRs, 0.41% and 0.70%, and the curve fit for IFR 0.41%.

- 2nd April) for both IFRs are shown in Figure 1. However, (1) cannot be used directly for 3rd to 16th April since we do not yet have the corresponding D_{t+14} values. Hence, we utilise the variation in the ratios A_t/I_t with time to estimate A_t for 3rd -16th April.

From Figure 1, it is quite clear that the nature of the plot differs before (increasing with time) and during the lockdown period (decreasing with time). Since, we are interested in understanding the effect of the lockdown period we fit a linear curve to K_t from 22nd March to 2nd April. We denote the linear approximation to K_t by K_t^l. The fitted $K_t^l (= 77.74 - 2.05t)$ obtained using IFR of 0.41% (shown by the red line in Figure 1) indicates a decreasing ratio of undetected to detected infections in the lockdown period. Using an IFR of 0.70%, K_t^l turns out to be $45.79 - 1.22t$. Using the linear fit K_t^l and I_t from our data set, we estimate A_t for 3rd-16th April. These estimated undetected cases and the observed detected cases from 22nd March till 16th April for IFR (0.41%) are plotted in Figure 2. Similar calculations for A_t were done using IFR value of 0.70%, but not shown in Figure 2 for visual clarity.

Results for Step 2: The values of β_t and γ_t over time t for the lockdown period are plotted in Figures 3 and 4 respectively, for IFR of 0.41%. We fit linear curves to β_t and γ_t for both IFRs. The fitted curves for IFR (=0.41%), are $\beta_t^l = 0.19 - 0.0065t$ and $\gamma_t^l = 0.00022 + 1.7 \times 10^{-5} t$, and IFR (=0.70%), are $\beta_t^l = 0.19 - 0.0070t$ and $\gamma_t^l = 0.00037 + 2.938 \times 10^{-5} t$. The slope of γ_t^l can be explained from studying Figure 6, where we plotted the time series of D_t

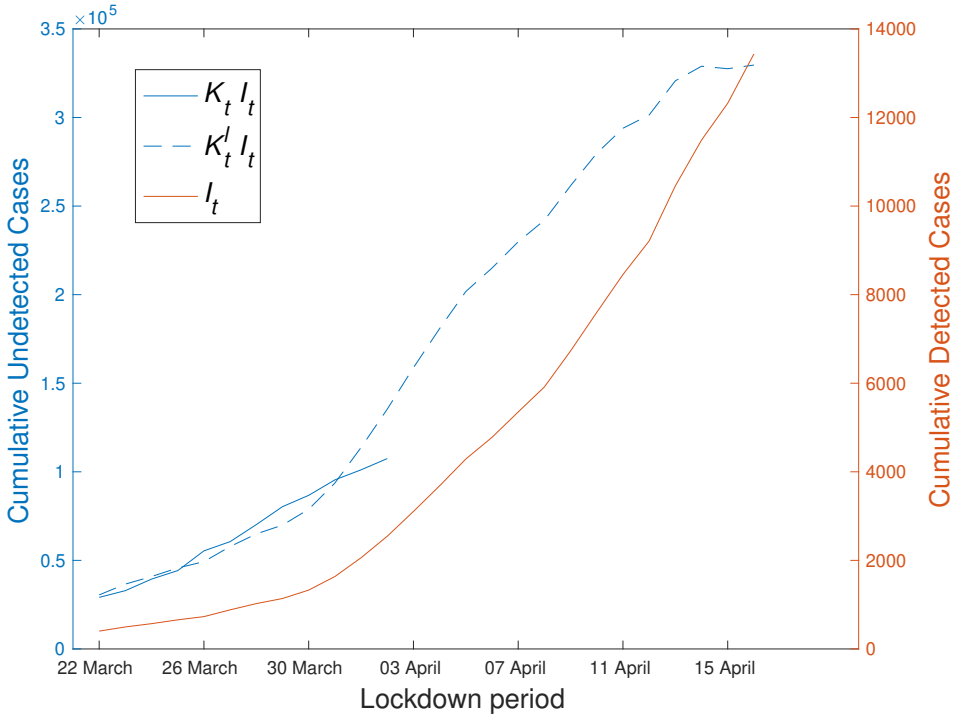

Figure 2: Detected cases, I_t (from data set) and undetected cases, A_t, estimated using K_t^l and IFR 0.41%. Note the difference in scales for I_t and A_t.

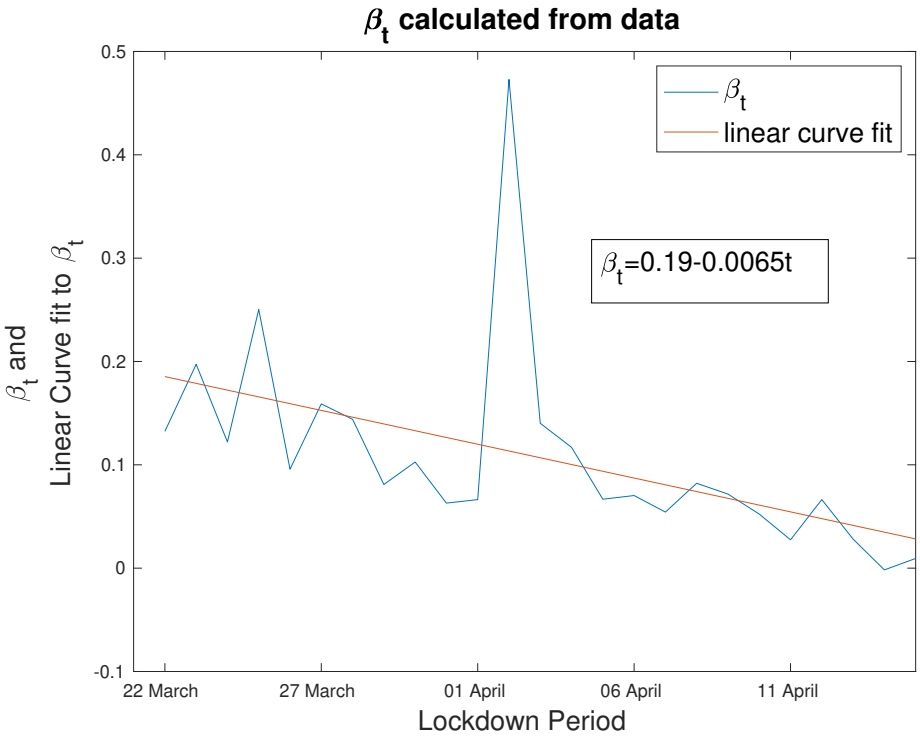

Figure 3: Estimated β_t in the lockdown period for IFR of 0.41%.

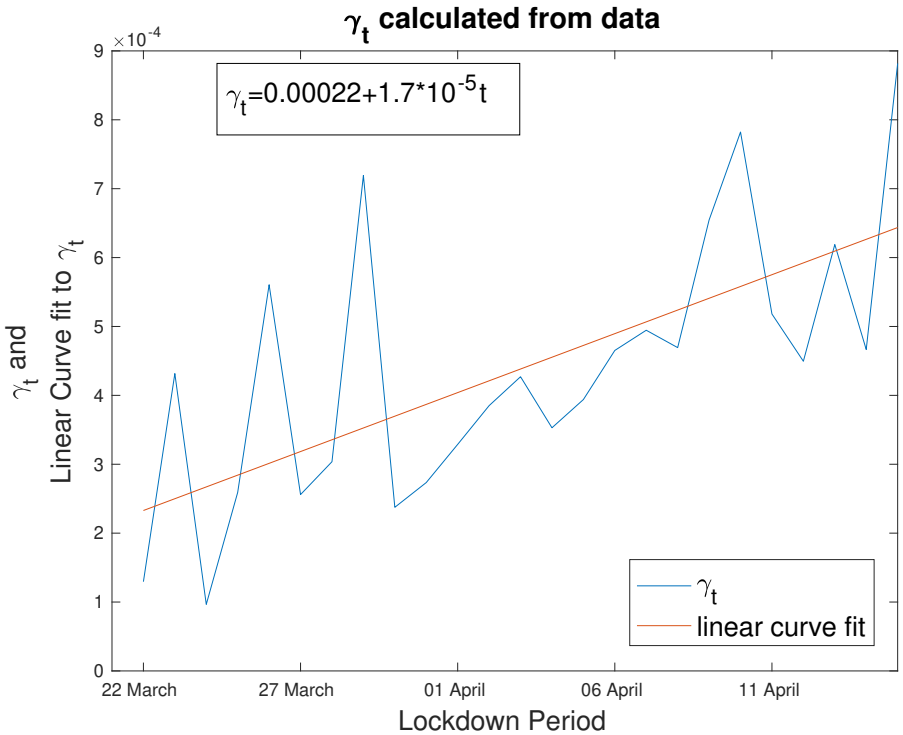

Figure 4: Estimated γ_t in the lockdown period for IFR of 0.41%.

and R_t. There, we see that though death rates are slowly increasing, the rate of increase in recoveries is quite high after 5th April, explaining the positive trend in γ_t^l. A negative slope in β_t^l is a welcome and expected result of lockdown. We note that the negative slope in β_t is modeling the effect of the intervention; lockdown, increased testing and hotspot containment, and the interventions do have a decreasing effect on the rate of rise of total infections. However, recall that β_t cannot take negative values. Hence, for the purpose of prediction we have used $\beta_t^l = \max(\beta_t^l, 0)$. Given this limitation, we advise against the use of our estimate of β_t^l for long term predictions.

Using β_t^l and γ_t^l, the proposed model (2) and (3) is initialized with the measured values of I_t, A_t, D_t and R_t on 24th March and iterated upto 21st April to simulate each variable upto this time for both values of IFR. The predicted values are plotted in Figures 5 and 6 along with observed data up to 16th April. The summary statistics of the model fit and predictions are shown in Tables 1 and 2.

From the model fitting and prediction, the important findings can be summarised as:

- From Table 1, using an IFR of 0.41%, we note that the total number of predicted infections (detected plus undetected) at the end of the active lockdown period (21st April) is 3,35,740.

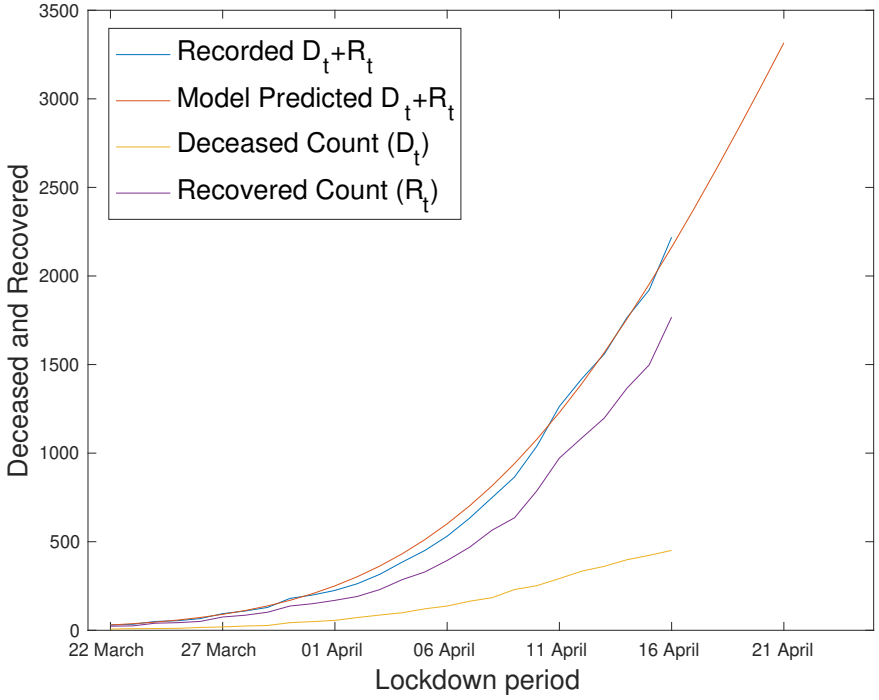

Figure 5: Predicted and observed values of D_t, R_t and $D_t + R_t$. The predictions are obtained from (3).

t	A_t	I_t	$A_t + I_t$	$\frac{I_t}{(I_t+A_t)}$
24th March	0.4253 $\times 10^5$	0.0059 $\times 10^5$	0.4313 $\times 10^5$	0.0138
6th April	2.0286 $\times 10^5$ (4.7695)	0.04508 $\times 10^5$ (7.5892)	2.0737 $\times 10^5$ (4.8083)	0.0217
21st April	3.1380 $\times 10^5$ (1.5469)	0.2194 $\times 10^5$ (4.8669)	3.3574 $\times 10^5$ (1.6190)	0.0653

Table 1: Predicted Values of A_t, I_t and $A_t + I_t$ and ratio of detections from (2) using an IFR of 0.41%. The times increase are given in parenthesis.

t	A_t^{Ke}	I_t	$A_t^{Ke} + I_t$	$\frac{I_t}{(I_t+A_t^{Ke})}$
24th March	0.2467 $\times 10^5$	0.0059 $\times 10^5$	0.2526 $\times 10^5$	0.0253
6th April	1.1700 $\times 10^5$ (4.7430)	0.0451 $\times 10^5$ (7.5892)	1.21510 $\times 10^5$ (4.8099)	0.03710
21st April	1.7491 $\times 10^5$ (1.4950)	0.2194 $\times 10^5$ (4.8669)	1.9686 $\times 10^5$ (1.6201)	0.11145

Table 2: Predicted Values of A_t^{Ke}, I_t and $A_t^{Ke} + I_t$ from (2) using an IFR of 0.70%. The times increase are given in parenthesis.

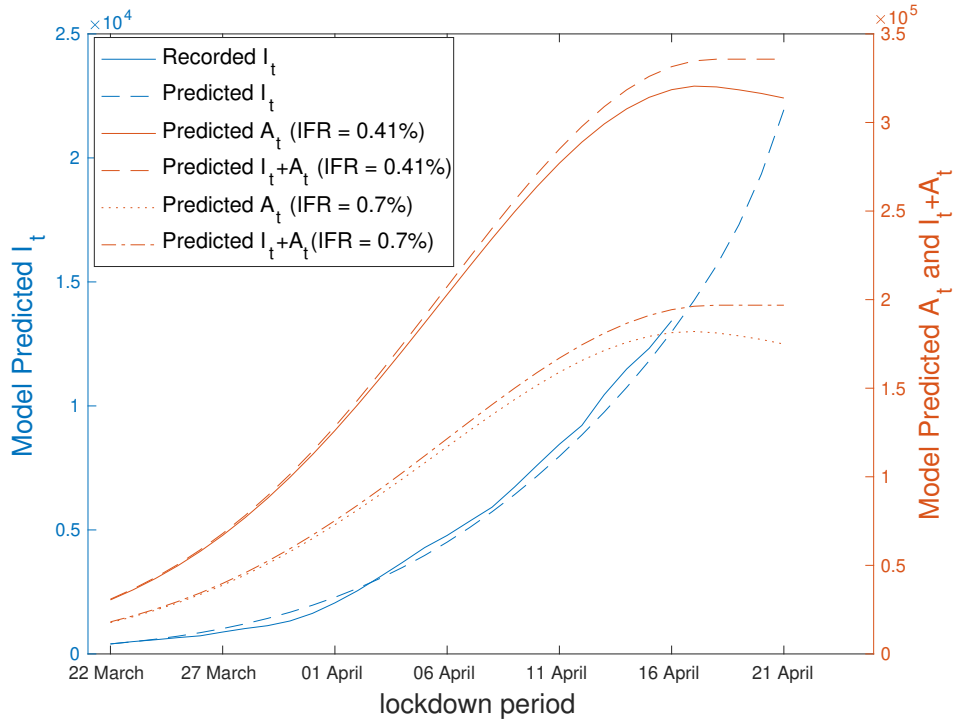

Figure 6: Predicted values of A_t, I_t, $I_t + A_t$ and observed values of I_t for both IFRs. The predictions are obtained from (2).

- From Table 2, using an IFR of 0.70%, we note that the total number of predicted infections at the end of the active lockdown period (21st April) are 1,96,860.
- From the pre-lockdown (24th March) to the mid lockdown period (6th April) there is a 7.6 times increase in the detected infections. From the mid-lockdown (6th April) to the end of active lockdown period (21st April) there is a 4.9 times increase in the detected infection. These two numbers indicate the rise in detected infections have decreased due to the lockdown and testing intervention effect.
- From the pre-lockdown (24th March) to the mid lockdown period (6th April) there is a 4.8 times increase in the undetected infections. From the mid-lockdown (6th April) to the end of active lockdown period (21st April) there is a 1.5 times increase in the undetected infection. These numbers indicate the rise in undetected infections have decreased due to the lockdown and testing intervention effect.
- The percentage of detections for IFR 0.41% increased to 6% from 1% in a month during lockdown. Modifying the IFR to 0.70%, the rate of detection grew to 11% from 2%.
- It is noted from the Figures 5 and 6 that the linear parameter varying model matches the

data satisfactorily. Significantly, the proposed model can reproduce the sub-exponential growth of infection numbers observed over the lockdown period in India.

Discussion

The focus of our work is on predicting the total number of infections in India due to COVID-19 over a short forecast horizon and also studying the effect of lockdown and increased testing on these total infections.

The important observations from the first part of the Results section, when total infections are estimated using IFRs are: From Figures 1 and 2 we see that in the pre-lockdown period, the ratio ($\frac{A_t}{I_t}$) increased with time t and the undetected cases were at least 30000. However, in the first week of the lockdown, the ratio ($\frac{A_t}{I_t}$) became almost constant over time. This slowing down effect on the rise in the undetected cases may be assumed to be the immediate effect of lockdown. From second week of the lockdown, the ratio ($\frac{A_t}{I_t}$) is noted to be decreasing with time. This leads to a further decrease in the number of undetected cases. This further decrease may be attributed to the additional effects of increased testing and active hotspot containment during the lockdown period. Thus, it seems from the data, that lockdown and increase in testing have lowered or slowed down the rate of rise in the number of undetected cases. However, one should note carefully the increasing A_t/I_t ratios prior to lockdown and consider the possible effect of increasing undetected cases before any relaxations of intervention measures. From the model fit and prediction in second step of the Results section, we note a fall in the increase in undetected cases and a simultaneous increase in detection of infections. At the end of the active lockdown we note that 21940 infections will be detected while the undetected cases will vary between 1.75-3.13 lakhs, given the two IFR values. Thus, it seems lockdown and increased testing have been effective measures in reducing the rise in infections from COVID-19 in India. However, as a word of caution, we would like to add that though the rate of increase of undetected cases seems to have slowed down with the interventions, at the end of 3 weeks into lockdown we already have 1.97 - 3.36 lakhs existing infections to combat.

We would also like to point out that the linear fit to the model coefficients (β_t and γ_t) are valid only in the short term, and the proposed infection model should not be used for long term predictions.

We understand that ignoring the susceptible population dynamics, limits the proposed model to only predict an exponential increase in the rise of infections. However, we believe that this is a reasonable assumption in the short to medium term when the susceptible population remains very high and the recovered population is negligible. As reported above, we see slight dampening in the exponentially increasing infection figures over differing time periods (during and upto the end of the formal lockdown period). We believe that this decrease in the

infection rate is due to intervention measures and gradual buildup of awareness in the general population rather than development of herd immunity or recovery dynamics. As an extension of this work, we plan to use our model for predicting state-wise total infections.

References

- Allen J S (1994). Some discrete-time SI, SIR, and SIS epidemic models. Mathematical Biosciences Volume 124, Issue 1, November 1994, Pages 83-105
- Bommer C and Vollmer S (2020) Average detection rate of SARS-CoV-2 infections is estimated around six percent. www.uni-goettingen.de/en/606540.html
- Goli S and James K S (2020) How much India detecting SARS-CoV-2 Infections? A model-based estimation. medRxiv preprint doi: https://doi.org/10.1101/2020.04.09.20059014.
- Verity et al. Estimates of the severity of coronavirus disease 2019: a model-based analysis. The Lancet Infectious Diseases 2020. DOI: 10.1016/S1473-3099(20)30243-7