Estimating under-reporting of COVID-19 cases in Indian states

MA4240: Applied Statistics

Name: Adepu Adarsh Sai
Roll no: AI20BTECH11001

Name: Ananthoju Pranav Sai
Roll no: AI20BTECH11004

Name: Murarisetty Adhvik
Roll no: AI20BTECH11015

Name: Perambuduri Srikaran
Roll no: AI20BTECH11018

Name: Arun Siddhardha
Roll no: AI20BTECH11019

Name: Yashas Tadikamalla
Roll no: AI20BTECH11027

***Abstract*— Due to limitations of testing coverage, it is known that the extent of spread of COVID-19 was under reported. We try to obtain both, an average as well as a time varying estimate for the fraction of COVID-19 cases reported in each state of India. The former is calculated using a delay adjusted case fatality ratio (CFR) approach for different states, while the latter is obtained by modelling the problem as a time-varying Poisson process. We show that the estimated average fraction of cases reported ranges from 2.7% to 19.7%, for an assumed baseline CFR of 0.1%. We also plot the time-varying fraction for different states. We observed that "more" underestimation was done specifically during the peak of second wave of COVID-19 in India. The utility of this work is to estimate the true spread of infection which can help in policy making and taking other measures that can curb the pandemic.**

***Index Terms*—CFR, delay-adjusted CFR, Poisson regression, time-varying poisson process, COVID-19, confidence interval.**

I. Introduction

COVID-19 pandemic has certainly affected the world in many ways. To understand its spread and to take the right decisions, it is important to estimate its true spread. However, we know that there has been significant under-reporting in India. Some of the possible reasons for this are:

- Low testing rate: As of October 10, 2020, the number of tests conducted in different states ranged only from 29 to 182 per 1000 residents.
- Asymptomatic cases: Typically, only symptomatic patients would get tested and hospitalised. Asymptomatic or mild symptomatic patients are unlikely to be tested.

Section II discusses about the data and the key assumptions used for this work. Sections III and IV describe how to find an average and time varying estimate of fraction of cases reported respectively. Section V discusses the confidence interval bounds. Section VI outlines all our findings and results, in two subsections VI-A, VI-B. Section VII talks about the strengths and weaknesses of this work.

II. Data and Key Assumptions

The data is taken from covid19india.org. We studied the data till October 10, 2020, for all the states of India and showed the results for a few of them. Some of the assumptions we assumed to perform the analysis are:

- Deaths due to COVID-19 are reported accurately i.e., no under-reporting of deaths.
- For fatal cases, the distribution of delay from confirmation to death is the same as the distribution of hospitalisation to death. We assume it is a log-normal distribution with mean of 13 days and standard deviation of 12.7 days. [3]

We will use the true CFR value of 0.1%, as estimated by [2].

III. Average estimate of fraction of cases reported

A naive estimate for case fatality ratio could be,

$$nCFR = \frac{\text{deaths-to-date}}{\text{cases-to-date}} \tag{1}$$

This is clearly inaccurate as the additional deaths that may arise from the cases observed to date will not be counted. Hence, we can use the distribution of delay from hospitalisation to calculate the delay-adjusted CFR as ratio of deaths to the expected number of eventually fatal cases.

$$cCFR = \frac{\sum_{t=0}^{T} d_t}{\sum_{t=0}^{T} \sum_{s<t} p_{t-s} c_s} \tag{2}$$

In the above equation, d_t and c_t are the number of new deaths and the number of new cases reported on day t from a certain region respectively. p_s is the probability that an eventually fatal case will lead to death on the s^{th} day from the day of reporting. T is the last day for which the data is available.

In regions where the cases have been under reported, the adjusted CFR will be higher than the true CFR. So, calculating the ratio of the true CFR to the adjusted CFR will give an estimate of the fraction of cases reported.

$$f_{avg} = \frac{CFR}{cCFR} \tag{3}$$

Now, the expected total number of deaths to occur among the reported cases on day t can be written as,

$$e_t = \sum_{s<t} p_{t-s} c_s CFR \tag{4}$$

Here, CFR is the true CFR. The ratio of $\sum_{t=0}^{T} e_t$ to the number of deaths reported by day T gives the average fraction of true cases that have been reported in a region over the complete time period.

$$f_{avg} = \frac{\sum_t e_t}{\sum_t d_t} \tag{5}$$

IV. Time varying estimate of fraction of cases reported

To improve the estimation, we will obtain a time-varying estimate of the fraction of cases reported. We model the number of daily deaths as a time-varying Poisson process. The number of deaths on day t is a random Poisson variable with mean,

$$\lambda_t = \frac{e_t}{f_t} \tag{6}$$

Here, f_t is the fraction of cases reported on day t.
To compute λ_t, we perform Poisson regression on the reported deaths. Then, we compute $1/f_t$ using (6). We estimated $1/f_t$ as a spline by fitting a generalise additive model using the pyGAM Python package and smoothened it. We applied this method to all the states and matched our results with the paper.

V. Confidence Interval bounds

Let X be a Poisson random variable with mean λ and $\lambda > 0$.

$$\mathbb{E}(X) = \lambda \tag{7}$$

$$Var(X) = \lambda \tag{8}$$

An interval of values for which we can be really confident the population mean/proportion falls is called the confidence interval. The bounds for a 95% confidence interval are,

$$\hat{\lambda} \pm 1.96\sqrt{\frac{\sigma^2}{n}} \tag{9}$$

Using (8) in (9),

$$\hat{\lambda} \pm 1.96\sqrt{\frac{\hat{\lambda}}{n}} \tag{10}$$

Here, n is the total number of days and $\hat{\lambda}$ is λ_t. The confidence interval for λ_t is $\left[\lambda_t - 1.96\sqrt{\frac{\lambda_t}{n}}, \lambda_t + 1.96\sqrt{\frac{\lambda_t}{n}}\right]$.

VI. Results

The code can be found here.

State /country	Deaths	Cases	nCFR (%)	cCFR (%)	Percentage Reported (%)
India	107459	6976461	1.54	1.78	5.60
Telangana	1208	208025	0.58	0.65	15.32
Maharashtra	39731	1506018	2.64	3.01	3.32
Bihar	934	193826	0.48	0.52	19.05
Delhi	5692	303693	1.87	2.04	4.91
Kerala	3773	122459	3.08	3.66	2.73

TABLE I: The data for number of deaths, number of total positive cases reported, calculated values of nCFR, cCFR, and estimated average fraction of true cases for some states, using an assumed baseline CFR of 0.1%

A. Average and time-varying estimate of fraction of cases

In Table I, we show our results for estimated average fraction of true cases, till October 10, 2020, using a baseline CFR of 0.1%. This estimate lies in the range of 2.7% to 19.7% for all states and India.
In Fig 1, we plot our estimated fraction of true cases as a function of time, again, calculated using an assumed baseline CFR of 0.1%. We also added the 95% confidence interval bounds, to get a better idea of the estimates.

B. Observed severe under-reporting

In Fig 2, we plot both daily active cases and estimated fraction of cases reported as a function of time upto October 2021. We know that 2nd wave of COVID-19 in India, occurred between March 2021 to July 2021, which can be verified from the plots as well (The region highlighted is the span of 2nd wave). From the plots, we can observe that every state has further under-reported during the second wave (a minima for estimated fraction of cases reported during the 2nd wave). Although we show the plots for only few states, it has been verified for every state.

VII. Strengths and limitations of the study

Here are a few ways in which this study can be useful:

- In states where extensive testing is infeasible, this study provides a method to quantify the true extent of spread of COVID-19.
- It provides important information, of trends of under-reporting in different states, which could be used for policy making.

All the analysis was made, based on certain vital assumptions. The analysis has to be changed accordingly, in case any of these assumptions was not true. The following are some of assumptions and associated consequences:

- The assumption that number of deaths reported is accurate: In case the deaths reported is also under-counted, we would get a incorrectly high estimate of fraction of cases. We can overcome this, if we can leverage the results of estimate of true count of deaths, from some other study.
- The assumption of the value of true CFR: If the actual value of true CFR is different from what assumed, the estimate of fraction reported would change accordingly.

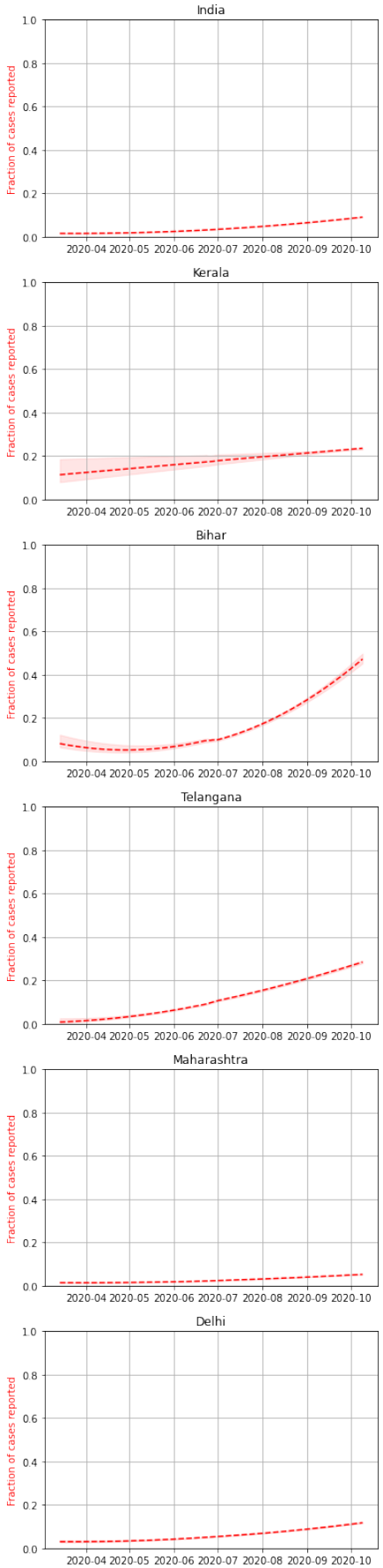

Fig. 1: Estimated fraction of cases reported vs time (till Oct 2020) , assuming a baseline CFR of 0.1% (including 95% CI bounds)

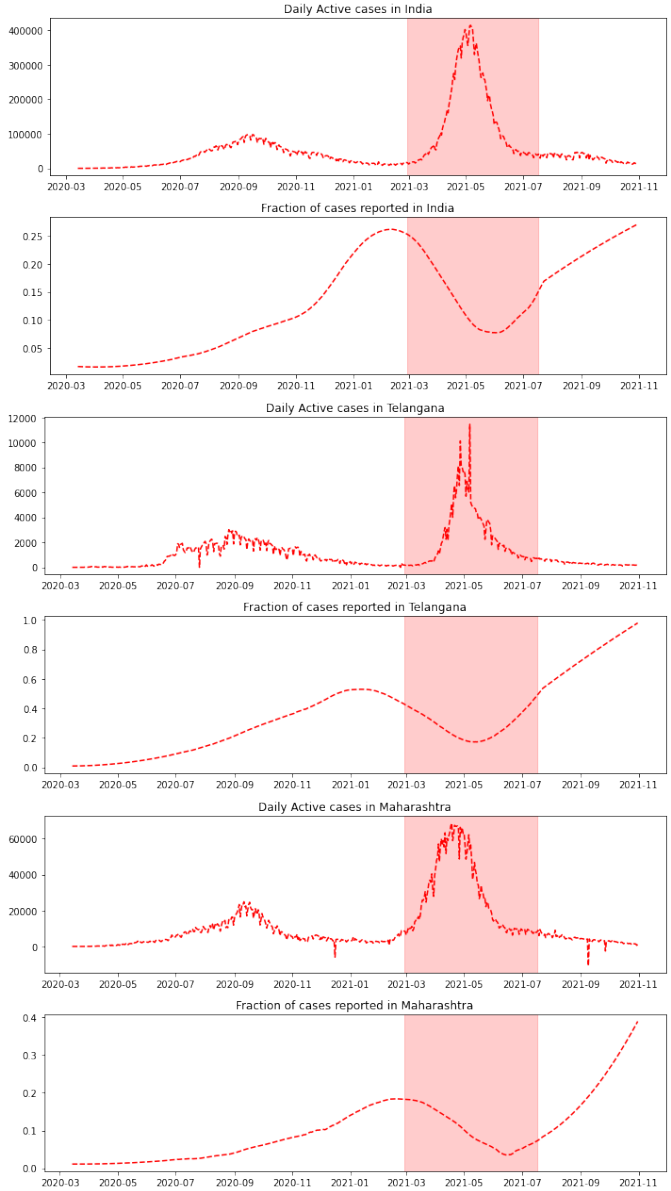

Fig. 2: Daily active cases and estimated fraction of cases reported vs time (till Oct 2021) , assuming a baseline CFR of 0.1%

- The assumption that distribution of delay from reporting to death is equal to distribution of delay from hospitalization to death: If this assumption does not hold, it would impact the fraction of cases estimated.

VIII. Contributions

- Adepu Adarsh Sai: Slides
- Ananthoju Pranav Sai: Section IV, codes for preprocessing, average and time-varying fraction of reported cases.
- Murarisetty Adhvik Mani Sai: Section VI (both VI-A, VI-B), code for observation of severe underreporting.
- Perambuduri Srikaran: Section V, code for 95% CI intervals.

- Arun Siddhardha: Slides
- Yashas Tadikamalla: Abstract, Section I, Section II, Section III, Section VII.

REFERENCES

[1] Jayakrishnan Unnikrishnan, Sujith Mangalathu and Raman V Kutty, "Estimating under-reporting of COVID-19 cases in Indian states: an approach using a delay-adjusted case fatality ratio", 2020.

[2] Ioannidis J., "The infection fatality rate of COVID-19 inferred from seroprevalence data", 2020.

[3] Linton NM, Kobayashi T, Yang Y, et al. "Incubation period and other epidemiological characteristics of 2019 novel coronavirus infections with right truncation: a statistical analysis of publicly available case data", 2020.