Correction in Active Cases Data of COVID-19 for the US States by Analytical study

Ravi Solanki\(^1\), Anubhav Varshney\(^2\), Raveesh Gourishetty\(^3\), Saniya Minase\(^4\), Namitha Sivadas\(^5\), Ashutosh Mahajan\(^5\)*

\(^1\) Visvesvaraya National Institute of Technology, Nagpur-440 010, India
\(^2\) Swami Vivekanand College of Engineering, Indore-452 020, India
\(^3\) Indian Institute of Technology Bombay, Mumbai-400076, India
\(^4\) Birla Institute of Technology and Science Pilani, Goa-403 726, India
\(^5\) Vellore Institute of Technology, Vellore-632 014, India

*E-mail: ashutosh.mahajan@vit.ac.in
Abstract—The total COVID-19 (SARS-CoV-2) infected cases have reached 139 million worldwide and nearing 3 million deaths on April 16, 2021. The availability of accurate data is crucial as it makes it possible to analyze correctly the infection trends and make better forecasts. The reported recovered cases for many US states are surprisingly low. This could be due to difficulties in keeping track of recoveries and resulted in higher numbers for the reported active cases than the actual numbers on the ground. In this work, based on the typical range of recovery rate for COVID-19, we estimate the active data from the total cases and death cases and bring out a correction for the data for all the US states reported on worldometer.

Keywords—COVID-19, Mathematical Model, epidemiology

BACKGROUND

The availability of accurate data of an epidemic is important as the data provides key insights on the disease spread and enable the authorities to take a decision on control measures. Worldometer is one of the very popular sources of the global COVID-19 data and it is also trusted and used by many government bodies and agencies [1]. The available data for the COVID-19 cases can be used for the prediction and analysis of hospitalization and meeting the demands of health care facilities and setting up the critical care systems for the patients. The active cases represent the number of infected people whether symptomatic or asymptomatic detected through self-reporting or testing. This number is important for Public Health authorities to estimate the current status of the disease spread and can be calculated by subtracting death and recovered cases from the total confirmed cases.

METHOD

A compartmental predictive mathematical model SIPHERD for COVID-19 disease dynamics is described in [2], [3] where recovery rate is a model parameter and is fixed by optimizing the model with the actual data. The data for the total, death, and active cases for 364 days from March 4, 2020, is taken from Worldometer [1] and found close to total cases and death data from [4]. After running the SIPHERD model for the USA, as reported in [2], the recovery rate of the active category was found to be 0.015 (corresponding to 66 days of mean recovery time) which is very low compared to other countries like Germany and India [2] [3].
where it is found 0.065. The low recovery rate in the USA may be attributed to either incorrect reporting of the Active cases [5] or the testing of serious cases only and longer recovery time in hospitals compared to quarantined with mild symptoms. Secondly, keeping the record of recoveries is difficult because some of the infected people are asked to quarantine while only critical patients are hospitalized. Sometimes the reporting of those recoveries is not accurate or incomplete. This has led to inconsistent data for active cases.

The number of mild cases is reported to be 81% in a Chinese study [6]. Covid-19 data reported from 49 states, the District of Columbia, and three U.S. territories to CDC from February 12–March 16 shows that 20.7 reported cases were severe and were hospitalized [7]. COVID-NET regions show this number to be 21.4% till April 4 [8] and IHME data March 5–April 4 shows that to be 20.3% [9] According to WHO recovery time for mild cases is of 2 weeks and for severe cases that recovery time of 3-6 weeks. Considering 80% mild cases, the recovery rate cannot be as low as appears from the data for the states listed in column 2 in Table 1.

The correct estimation of the active cases can be done by subtracting the death and recovery cases, with appropriate recovery rate, from the total cases. The Worldometer data for total cases and death cases are assumed to be true as the test positive results and count of diseased cases are done more stringently as compared to recovery counting. The active cases can be obtained by solving the following differential equation,

\[ \frac{dH(t)}{dt} = \frac{dT(t)}{dt} - \frac{dD(t)}{dt} - \sigma H(t - t_R) \]

where, \(H, T,\) and \(D\) are the active, total, and death cases, respectively. The recovery from the infected category is defined by the two parameters: delay in recovery \(t_R\) and the recovery rate \(\sigma\). As these values are dependent upon the immune system of the community and the hospital facilities, it should not vary much within the USA. We have taken \(t_R\) as 10 days and \(\sigma\) as 0.048 (21 days of mean recovery time considering both mild and severe cases).

**RESULTS AND DISCUSSION**

The above delay differential equation is solved for all the states of the USA and we found three groups among the states according to the accuracy of the data. The active cases reported on worldometer [1] for a few states show excellent agreement with our estimation of active
cases. One example state for this group is Texas as seen in Fig. 1(a). There are few states in the second group which are largely not matching with analytical estimation indicating that reported active data is inaccurate. These states are listed in column 2 in Table 1 and one representative state is Virginia as seen in Fig. 1(b), where the current active cases are reported to be 530820 which should have been just around 31237 according to our calculation. Interestingly, in the last group, there are some states for which the reported active cases follow the estimated active cases for some days, however, the trend of the curve changes and do not follow our estimation as seen for Indiana in Fig. 1(c). In Fig. 1, the reported total and active cases with the estimated active cases for one of the states in all the three groups are shown, the figures for the remaining states are given in the Supplementary Material.

The Network Systems Science and Advanced Computing (NSSAC) division of the Biocomplexity Institute and Initiative at the University of Virginia has created a visualization tool that presents a way of examining data curated by different data sources [10]. We compare the active cases data provided by [10] and find that this independent source of active data is in close agreement with the corrected active cases data. The recovery rate in individual states may vary to an extent of ±10% depending on the variation in the number of tests performed, fraction of mild and severe cases. However, we have taken a uniform value of recovery rate as 0.048 for all the US states. Mortality rate can be calculated from the active cases data [2] as

\[ \tau = \frac{DND(t)}{H(t - t_d)} \]

where, \(DND\) are the daily new extinct cases, and \(t_d\) is the delay associate with the extinct cases as explained in [2]. In the initial phases of infection, many of the states show higher active cases than the true values that gives mortality rate lower than the actual one. Since the hospital bed estimation, ICU equipment requirement estimation depends on the active cases currently and that are expected in the near future, correction in the data facilitates better management of these entities. For the purpose of modeling and prediction, it is important that a mathematical model is validated against the data. Correct active cases data implies the right model parameters and more accurate estimation of the hospital requirements.

**CONCLUSIONS**
The reported active cases for a few states are consistent with the total detected cases and death cases for a recovery rate parameter value 0.048. For a few states, the data has been corrected recently. However, for 9 states, the active cases data is still largely incorrect. We generate the corrected active case data for all states, report in the Supplementary Material, and also keep the data available on Github.

A. Data Availability

The data for the corrected active case for all the US States can be downloaded with the github link below. [https://github.com/ravisolankigithub/covid-activcases-usa.git](https://github.com/ravisolankigithub/covid-activcases-usa.git)
Table I: The US States Active Cases Data Status

| Correct Data | Incorrect Data | Partially Correct |
|--------------|----------------|-------------------|
| Texas        | Maryland       | Nebraska          |
| Tennessee    | Virginia       | Indiana           |
| Louisiana    | Maine          | California        |
| Arizona      | Hawaii         | Oregon            |
| North Carolina | Kentucky     | Florida           |
| Illinois     | Minnesota      | Idaho             |
| Massachusetts | Rhode Island  | New Mexico        |
| Pennsylvania | South Carolina | Louisiana         |
| Ohio         | Washington     | Alabama           |
| Arkansas     |                | Michigan          |
| Connecticut  |                | Alaska            |
| Delaware     |                | Colorado          |
| Iowa         |                | Missouri          |
| Kansas       |                | New Jersey        |
Fig. 1: (a) Texas representing states in which active data is reported correctly. (b) Virginia represents the second group where data is largely incorrect and (c) Indiana from the third group where data is partially correct. The NSSAC, University of Virginia data for the active cases is in close agreement with our analytical estimation.