BMJ Open

Estimating under-reporting of COVID-19 cases in Indian states: an approach using a delay-adjusted case fatality ratio

Jayakrishnan Unnikrishnan,[1] Sujith Mangalathu ,[2] Raman V Kutty[3]

To cite: Unnikrishnan J, Mangalathu S, Kutty RV. Estimating under-reporting of COVID-19 cases in Indian states: an approach using a delay-adjusted case fatality ratio. *BMJ Open* 2021;**11**:e042584. doi:10.1136/bmjopen-2020-042584

► Prepublication history for this paper is available online. To view these files, please visit the journal online (http://dx.doi.org/10.1136/bmjopen-2020-042584).

Received 10 July 2020
Revised 20 December 2020
Accepted 30 December 2020

[1]Independent researcher, Jersey City, New Jersey, USA
[2]Data Analytics, Mangalathu, Cumming, Georgia, USA
[3]Amala Cancer Research Center, Amalanagar, Thrissur, Kerala, India

Correspondence to
Dr Sujith Mangalathu; sujithmangalath@ucla.edu

ABSTRACT

Objectives The COVID-19 pandemic has spread to all states in India. Due to limitations in testing coverage, the true extent of the spread may not be fully reflected in the reported cases. In this study, we obtain time-varying estimates of the fraction of COVID-19 infections reported in the different states.

Methods Following a methodology developed in prior work, we use a delay-adjusted case fatality ratio to estimate the true fraction of cases reported in different states. We also develop a delay adjusted test positivity estimation method and study the relationship between the estimated test positivity rate for each state and the estimated fraction of cases reported.

Setting We apply this method of analysis to all Indian states reporting at least 100 deaths as of 10 October 2020.

Results Our analysis suggests that delay-adjusted case fatality ratios observed in different states range from 0.47% to 3.55%. The estimated fraction of cases reported in different states ranges from 39% to 100% for an assumed baseline case fatality ratio of 1.38%, from 18.6% to 100% for an assumed baseline case fatality ratio of 0.66%, and from 2.8% to 19.7% for an assumed baseline case fatality ratio of 0.1%. We also demonstrate a statistically significant negative relationship between the fraction of cases reported in each state and the testing positivity rate.

Conclusions The estimates provide a means to quantify and compare the trends of reporting and the true level of current infections in different states. This information may be used to guide policies for prioritising testing in different states, and also to analyse the time-varying effects of different quarantine measures adopted in different states.

Strengths and limitations of this study

- By quantifying the time-varying estimate of under-reporting, this study provides a method to quantify the true extent of the infection, and the temporal trend in the occurrence of new infections in different states.
- By accounting for delay from case reporting to death, this method provides a method to estimate the case fatality rate in a region more accurately.
- Unlike methods based on expensive serological tests that provide cumulative estimates for the total number of infections over the course of the pandemic, the proposed method provides an inexpensive alternative to obtain time-varying estimates of the rate of new infections.
- The accuracy of these results depends greatly on the value of the true baseline case fatality rate of COVID-19, which is still not known with certainty.
- The accuracy of these results depends on the assumption that the number of deaths are correctly reported.

BACKGROUND

The first case of COVID-19 in India was reported in the state of Kerala in a student returning from Wuhan, China, on 30 January 2020. Since then, the infection has spread throughout the country, with every state reporting at least one case positive case of COVID-19 as of 10 October 2020. However, the reported cases may not give the full picture of the extent of the infection as testing coverage has not been complete. Data from covid19india.org[1] suggest that the number of tests conducted up to 10 October 2020 in various states were in the range of 29 to 182 per 1000 residents. Although patients hospitalised with symptoms are typically tested, those who develop mild symptoms at home and those who do not develop symptoms are unlikely to be tested. The testing protocols used in different states have also changed significantly over the duration of the pandemic. Nevertheless, knowing the true extent of the prevalence of infection throughout the country is critical for policy making around handling the outbreak, including determining the required level of deployment of testing and treatment infrastructure and personnel. Estimating the time-varying level of under-reporting existing in different states can help in determining the true time-varying extent of the infection.

One recent work attempts to estimate the level of under-reporting in the USA during the first half of March 2020 using travel data from epicentres.[2] Another study[3] uses a Bayesian analysis to get an estimate of the cumulative number of unreported cases in the USAS up to 18 April 2020.

METHODS

Data description

The primary data used in the under-reporting analysis are the daily reports of cases and deaths from various states and union territories of India, which we accessed from covid19india.org[1]. These data are crowd-sourced from different state bulletins and official and validated and maintained by a group of volunteers. We restrict to data up to and including 10 October 2020.

In addition, for illustration and for studying the relationship of the rate of reporting with testing rates, we also use the reports of testing from different states, also available at the same website.

Key assumptions and basic technique

We assume that the deaths due to COVID-19 reported in different states are accurate. Although cases may have significant under-reporting, deaths are typically reported correctly. This is because patients with severe symptoms typically report themselves to a hospital. As a result, any patient who dies from the COVID-19 disease is likely to have been tested.

A naive computation of the ratio of deaths-to-date to cases-to-date from a region gives an inaccurate estimate of the observed case fatality ratio (CFR) of the out-break in a region. This is because the deaths used in the numerator undercounts additional deaths that may arise from the cases observed to date. This issue can be addressed by using the distribution of delay from hospitalisation to deaths for cases that are fatal. With this correction, one can compute an adjusted CFR for each region being studied.

In a region where the cases and deaths have been fully reported, we expect the adjusted-CFR to match the true CFR of COVID-19 reported in published studies that have accounted for reporting biases. For example, a value of 1.4% for the true CFR has been reported in Guan et al. (2020)[4]. A different published study based on data from China puts the estimate at 0.66%.[5] More recent reports based on seroprevalence studies provide much lower estimates as low as 0.1%.[6]

However, in regions where cases have been under-reported, we expect the adjusted-CFR to be significantly

Table 1 Estimates of fraction of cases reported in different states

State	Deaths	Cases	Test positivity rate(%)	*nCFR*(%)	*cCFR*(%)	Percentage reported (CFR of 1.38%)(%)	Percentage reported (CFR of 0.66%)(%)	Percentage reported (CFR of 0.10%)(%)
India	106 863	6 976 461	–	1.53	1.78	77.62	37.12	5.62
Andhra Pradesh	6159	744 864	6.1	0.83	0.93	100.00	71.07	10.77
Assam	807	192 314	3.6	0.42	0.47	100.00	100.00	21.11
Bihar	934	193 826	0.6	0.48	0.53	100.00	100.00	18.92
Chhattisgarh	1196	137 570	14.1	0.87	1.14	100.00	57.86	8.77
Delhi	5692	303 693	7.0	1.87	2.13	64.85	31.01	4.70
Gujarat	3549	149 193	2.2	2.38	2.68	51.59	24.67	3.74
Haryana	1562	139 932	6.5	1.12	1.29	100.00	51.13	7.75
Jammu and Kashmir	1306	82 429	4.1	1.58	1.84	74.84	35.79	5.42
Karnataka	9200	690 269	10.7	1.33	1.60	86.35	41.30	6.26
Kerala	956	268 101	14.8	0.36	0.51	100.00	100.00	19.53
Madhya Pradesh	2575	143 629	7.2	1.79	2.14	64.57	30.88	4.68
Maharashtra	39 731	1 506 018	19.3	2.64	3.02	45.67	21.84	3.31
Odisha	1044	246 839	7.8	0.42	0.51	100.00	100.00	19.70
Punjab	3774	122 462	3.9	3.08	3.55	38.88	18.59	2.82
Rajasthan	1621	154 785	9.4	1.05	1.25	100.00	52.81	8.00
Tamil Nadu	10 120	646 128	6.1	1.57	1.75	78.80	37.69	5.71
Telangana	1208	208 025	4.1	0.58	0.66	100.00	100.00	15.18
Uttar Pradesh	6293	430 666	2.1	1.46	1.66	83.16	39.77	6.03
West Bengal	5501	287 603	6.6	1.91	2.23	61.89	29.60	4.49

CFR, case fatality ratio.

higher than the true-CFR. Hence, computing the ratio of the true-CFR to the adjusted CFR gives an estimate of the fraction of cases that have been reported.

We adapt this method for estimating under-reporting developed in Russell et al. (2020)[7] and apply it to data from different states of India. We provide results for multiple choices for the baseline CFR of COVID-19. For completeness, we elaborate on the details of the method below.

Method details

Following[7] we assume that for fatal cases, the delay from confirmation to death follows the same distribution as delay from hospitalisation to death estimated in Linton et al. (2020)[8]. This estimate is based on data from the outbreak in Wuhan, China, between 17 December 2019 and 22 January 2020, and accounts for right censoring in the death numbers due to unknown disease outcomes among active cases. The fitted distribution is a lognormal distribution p with a mean delay of 13 days and a SD of 12.7 days. Let p_s represent the probability that an eventually fatal case leads to death during the s-th day from the day of confirmation. Let c_s denote the number of new cases and d_s denote the number of new deaths reported on day s from a region. With these definitions, we can now calculate the adjusted CFR $cCFR$ for the region as the ratio of the total deaths to the expected number of eventually fatal cases among the reported cases

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

where T is last date for which data are available. Moreover, disagreement between the $cCFR$ and the true CFR of COVID-19 can be used to get an estimate of the fraction of total cases that have been reported. If CFR is the true CFR of COVID-19, the total number of deaths that we expect to occur among the reported cases on day t can be calculated as

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

where CFR is the true CFR of COVID-19. The ratio of the total number of deaths reported by day T to the cumulative sum of e_t up to T provides an estimate of the average fraction of true cases that have been reported in the region, over the duration of the pandemic.

We can further improve the estimate to obtain a time-varying estimate of the fraction of cases reported. We model the daily deaths as a time-varying Poisson process. The deaths on day t is a random variable with mean given by

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

where f_t is the fraction of cases reported. To be precise f_t represents the fraction reporting as reflected in the death rate on day t. Hence as we assume a mean delay of 13 days from case confirmation to death, the quantity

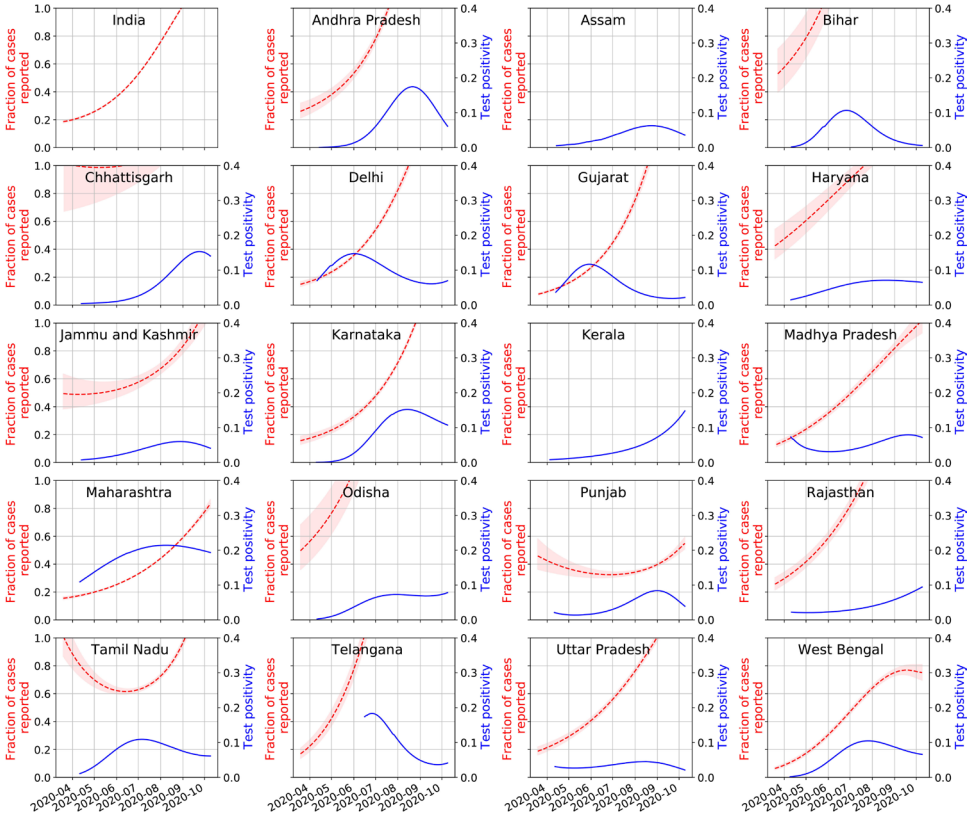

Figure 1 Curves in red show the estimates of the fraction of cases reported in various regions as a function of time, assuming a baseline CFR of 1.38%. The first subplot shows the results for India and the other subplots show results for the top 19 states with most reported deaths. Curves in blue show the smoothed estimate of test positivity rate. CFR, case fatality ratio.

f_t is reflective of the under-reporting that existed around day $t-13$.

We estimate $1/f_t$ by performing Poisson regression on the reported deaths using the aforementioned model for the mean function λ_t. To ensure a smooth estimate, we estimate $1/f_t$ as a spline by fitting a generalise additive model using the pyGAM Python package. We applied this method to all states with at least 100 reported deaths.

Under-reporting of cases occurs when infected people have not been tested. In regions with insufficient testing, the fraction of cases reported is expected to be low. Moreover, in regions with low testing coverage, testing tends to be performed only on people who are most at risk of having contracted the infection. Consequently, in such regions, a larger fraction of the tests conducted also tend to turn out positive. Therefore, we expect a negative correlation between the fraction of cases reported in a region and the test positivity observed in a region, defined as the fraction of tests that are positive. In order to test this hypothesis, we also computed the test positivity rate of the different states. As testing rates are time-varying, we again use a Poisson model to estimate the positivity rate. We assume that the result of test performed on 1 hour is obtained with equal probability on the same day, the next day, or the day after. We model the number of positives reported on a particular day t as a Poisson random variable with the mean given by the product of the positivity rate and the average number of tests performed on days $t-2$, $t-1$ and t. We then perform Poisson regression on the data on reported positives and tests performed to obtain a smoothened estimate for the positivity rate of each state. We further analyse the relationship between the under-reporting estimated by our method and the test positivity rate.

Summary of assumptions

- We assume that deaths are accurately reported.
- The estimates of under-reporting obtained are a function of the assumed base line CFR for COVID-19. We provide results for baseline CFRs of 1.38%, 0.66% and 0.1%. These estimates will vary if the true baseline is different.
- We assume that for eventually fatal cases, the delay from reporting of cases to death follows the lognormal distribution with parameters described above.

RESULTS

In table 1, we list the estimates obtained for all states that report at least 10 deaths. The test positivity is the test positivity on 10 October calculated using the Poisson regression approach. Due to lack of sufficient data, we do not estimate

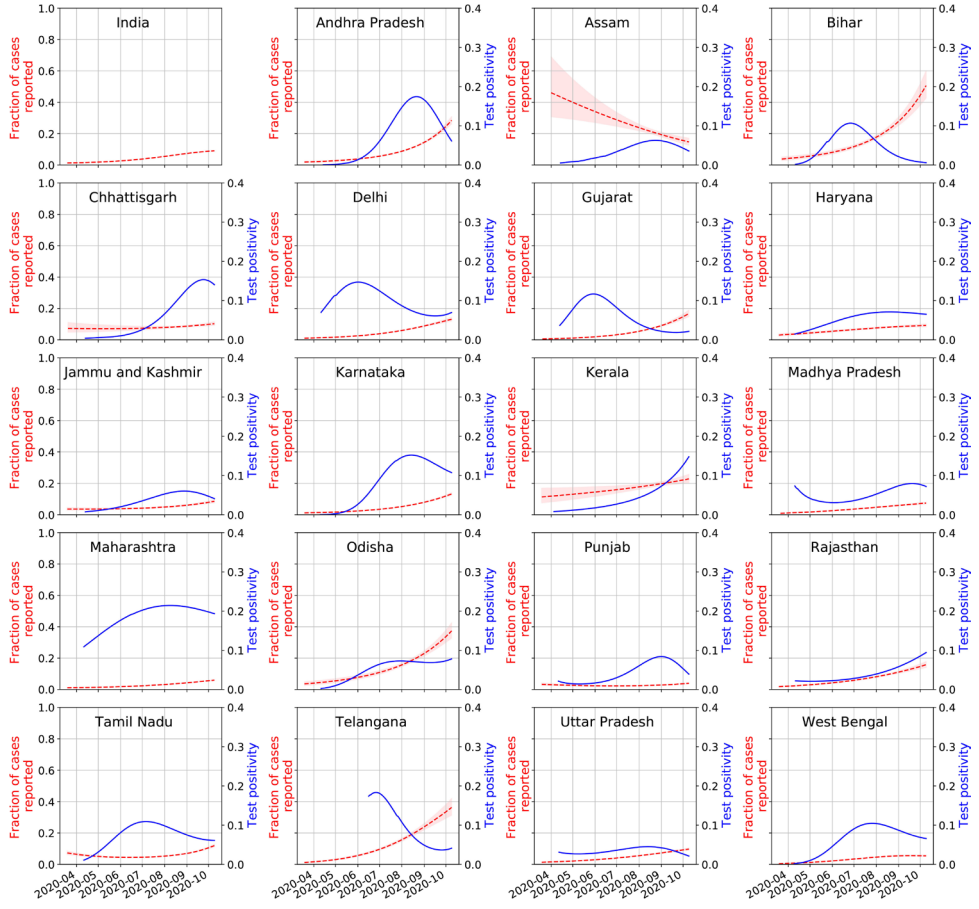

Figure 2 Curves in red show the estimates of the fraction of cases reported in various regions as a function of time, assuming a baseline CFR of 0.1%. The first subplot shows the results for India and the other subplots show results for the top 19 states with most reported deaths. Curves in blue show the smoothed estimate of test positivity rate. CFR, case fatality ratio.

positivity rate for India and Telangana. The *nCFR* column represents the naive CFR estimate one would estimate by using the ratio of total deaths to total cases, and *cCFR* gives the corrected CFR obtained after accounting for right censoring in deaths via the method described above. It can be seen that the ratio of *cCFR* to *nCFR* varies from 1.1 to 1.4, which suggests that it is important to account for the delay in reporting while estimating CFR's. In the same table, we also provide estimates of the under-reporting obtained assuming baseline CFR's of 1.38%, 0.66% and 0.1%. These numbers are the ratios of total deaths to the number of deaths that should be expected if the reported cases were accurate. As expected, the estimate for the fraction reported is significantly lower for lower values of the assumed baseline CFR compared with those for higher values of assumed baseline CFR.

The time-varying estimates of the fraction reported f_t for the whole country and for nineteen regions with most deaths are illustrated in figure 1 for an assumed baseline CFR of 1.38% for COVID-19 and in figure 2 for an assumed baseline CFR of 0.1%. The red curves show the estimate of the fraction reported and the shaded region represents the associated 95% CI bounds for the Poisson regression model. In the same figures, we also plot the test positivity rates obtained in each state.

In figure 3, we provide a comparison of the evolution of the instantaneous test positivity rate (in blue) with that of the ratio of cumulative positive cases reported to cumulative tests conducted (in green). The difference between the two curves suggests that the cumulative ratio may not accurately capture the recent test positivity rate.

Figure 4 shows a scatter-plot of the estimate of the fraction reported against the test positivity rate for all states reporting at least 100 deaths. The quantity plotted on the vertical axis is the estimate of the fraction f_t of cases reported, estimated on the last date where data are available (10 October 2020), assuming a baseline CFR of 0.1%. As mentioned earlier, f_t

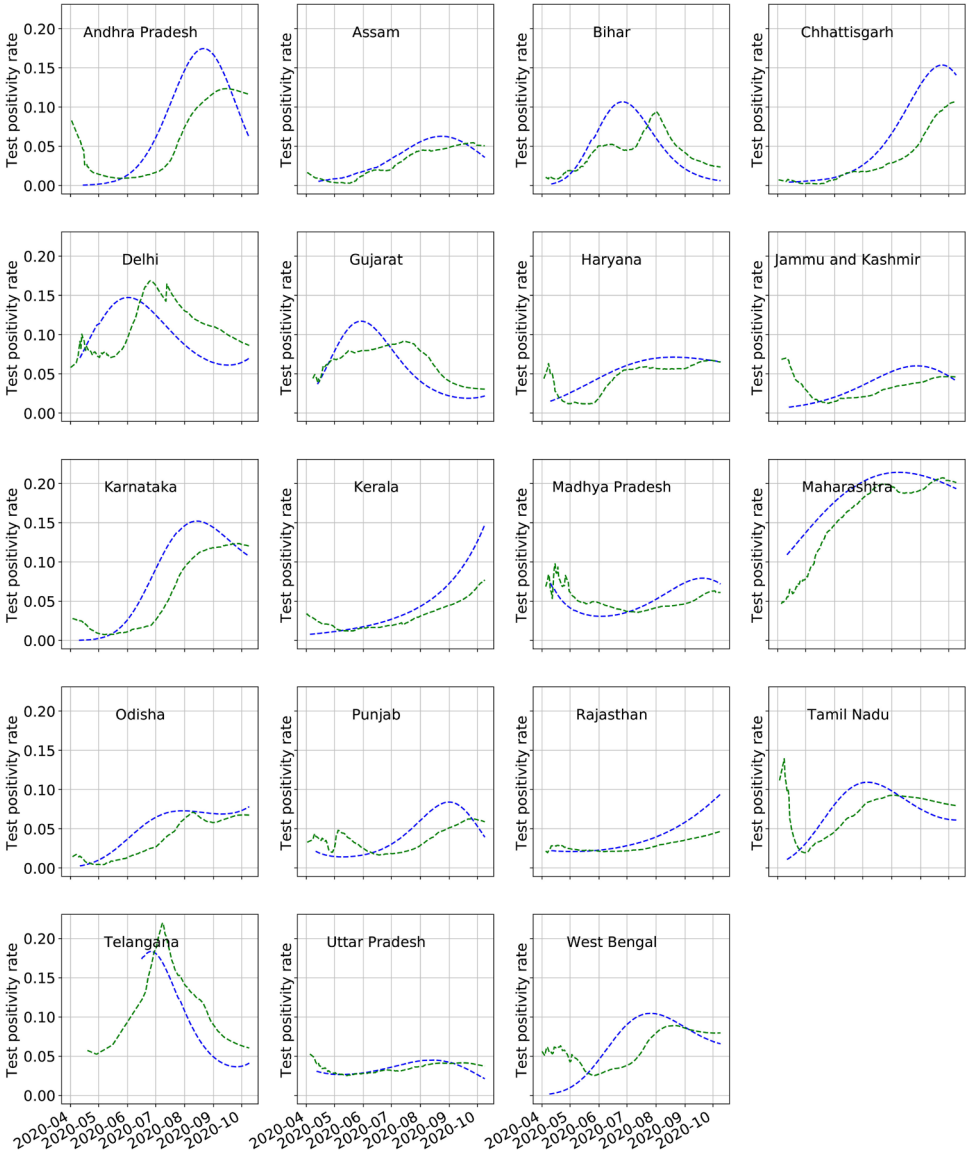

Figure 3 Curves in blue shows the test positivity rate estimated via the Poisson regression method. Curves in green show the ratio of cumulative positive cases to cumulative tests performed.

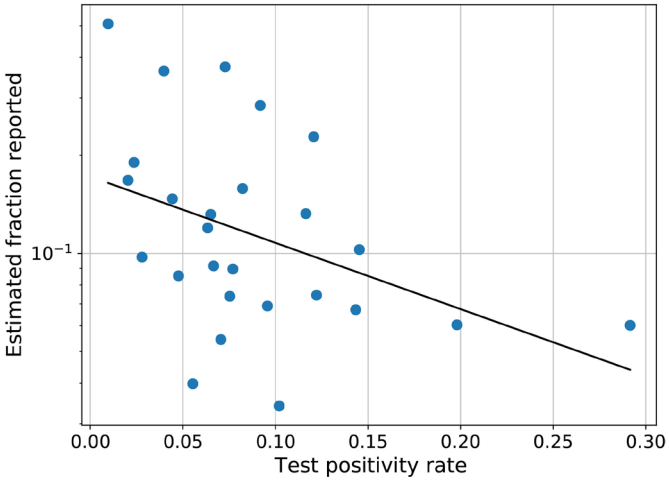

Figure 4 Scatter plot of the estimate of the fraction f_t of cases reported from different states evaluated on the last date considered, against the corresponding test positivity rate.

provides an estimate of the fraction of cases reported around day $t-13$. To account for this delay, the quantity plotted on the horizontal axis is $\sum_{s<t} p_{t-s}.P_s$, where p represents the distribution of the delay from case to death, and P_s denotes the estimated test positivity rate on day s, evaluated when t is that last day (10 October 2020). We observe that states with high values of the positivity rate also tend to have low estimates of the fraction of cases reported. In order to quantify the strength of this inverse monotonic relationship, we computed the Spearman's rank correlation coefficient[9] between these two quantities. We obtained a correlation coefficient of −0.4 with a p value of 0.03 indicating a moderately strong monotonic inverse relationship between the quantities. Thus, an increase in test positivity rate is associated with a decrease in the fraction of cases reported.

DISCUSSION

This study provides a method to estimate the fraction of COVID-19 cases reported in different states within the country. The method can be applied using only the daily reports of cases and deaths from different states. An alternative method one could adopt to quantify under-reporting may be to use results of serological testing[10 11] for COVID-19 antibodies among the general public. Randomised antibody testing in a general population may be used to estimate the fraction of the people who have the COVID-19 antibody in their system, which in turn serves as an estimate of the total population who have been exposed to the virus. This could then be used with the total cases reported to arrive at an estimate for the fraction of cases reported. An advantage of this approach is that this provides a direct way to measure past infections. However, antibody testing does not provide an estimate of when a person was infected, and hence is not sufficient to estimate the temporal variation in the under-reporting. This method therefore does not directly provide an estimate of the current prevalence of the infection in the population, which on the other hand can be obtained by the method proposed in the current study. Furthermore, in order to have accurate estimates, one would have to test a substantial portion of the population of the state and also cover a wide area of the state. This requires additional testing which could be expensive. The proposed method on the other hand uses only reports of cases and deaths, which are more readily available.

In the study, we also observed a statistical association between the estimated fraction of cases reported from a state with the test positivity rate reported from the state. It is known that one of the causes of high test positivity in a region is the lack of broad testing across the population, and hence one can expect that such regions also have higher prevalence of unreported cases. This could explain the negative correlation we observed between the estimated fraction of reported cases from a region and the test positivity from the region.

Strengths and limitations of the study

In states where extensive testing is infeasible, this study provides a method to quantify the true extent of the infection. The analysis reveals the trends in under-reporting in different states and could be useful for policy making.

The accuracy of these results depends greatly on the quality of the data and the assumptions being made. The most critical parameter assumption made here is that about the value of the true CFR of COVID-19 that we use as the baseline level in our analysis. If the true CFR is different from what is assumed, the estimate of the fraction reported would change accordingly.

Another key limitation is the assumption that the number of deaths is accurately reported. If the number of deaths reported is undercounted, this would lead to an incorrectly high estimate for the fraction of cases reported. This limitation can be partially addressed if the under-reporting rate for deaths can be estimated by other means. For example, it may be possible to estimate the fraction of COVID-19 deaths reported based on the protocol for death-reporting followed in different regions. If it is known that only a fraction α of the actual deaths are reported, this can be used to adjust for the resulting bias in the estimation of the fraction of cases reported. In particular, the formula for the adjusted CFR $cCFR$ given in the methods section may be scaled by $1/\alpha$, and the formula for the expected deaths e_t may be scaled by factor α. These adjustments in the method will then lead to more accurate estimates for the adjusted CFR and the fraction of cases reported.

Furthermore, if the distribution of delay of eventually fatal cases from reporting to death deviates from what is assumed here, that would also have an immediate impact on the predicted fraction of cases reported.

Conclusions and future work

We have obtained an estimate of the temporal evolution of the fraction of cases reported in different Indian states. We further showed that, as expected, the estimate of fraction estimated shows a moderately strong monotonic inverse relationship with the test positivity rate.

The estimate of under-reporting may be used to guide policies for prioritising testing in different states by focusing on states with higher and increasing levels of under-reporting. The estimated reporting fraction taken together with the number of reported cases provides a means to obtain a time-varying estimate of the true number of infections in different states. As follow-up work, these estimates may be compared with timelines of different lockdown and quarantine measures to quantify their effectiveness in controlling the rate of spread of infections.

Twitter Sujith Mangalathu @sujithmss

Acknowledgements We thank the volunteers of COVID19-India [1] for making the data from all states available at a common location. We thank the authors of [7] for sharing their work and code online, and Timothy Russell for answering our questions on the method.

Contributors JU adapted and implemented the statistical model. JU and SM wrote the paper. All authors (JU, SM and RVK) critically reviewed the approach and the manuscript and gave approval for the publication. All views expressed in this publication are of the authors only.

Funding The authors have not declared a specific grant for this research from any funding agency in the public, commercial or not-for-profit sectors.

Competing interests None declared.

Patient and public involvement Patients and/or the public were not involved in the design, or conduct, or reporting, or dissemination plans of this research.

Patient consent for publication Not required.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data are available in a public, open access repository. The primary data used in the under-reporting analysis are the daily reports of cases and deaths from various states and union territories of India, which we accessed from the public website (1).

ORCID iD

Sujith Mangalathu http://orcid.org/0000-0001-8435-3919

REFERENCES

1 COVID19-India API. Available: https://api.covid19india.org
2 Hortaçsu A, Liu J, Schwieg T. Estimating the fraction of unreported infections in epidemics with a known epicenter: an application to COVID-19. *J Econom* 2021;220:106–29.
3 Wu SL, Mertens AN, Crider YS, *et al*. Substantial underestimation of SARS-CoV-2 infection in the United States. *Nat Commun* 2020;11:4507.
4 Guan W-jie, Ni Z-yi, Hu Y. Clinical characteristics of coronavirus disease 2019 in China. *N Engl J Med Overseas Ed* 2020;382:1708–20.
5 Verity R, Okell LC, Dorigatti I, *et al*. Estimates of the severity of coronavirus disease 2019: a model-based analysis. *Lancet Infect Dis* 2020;20:669–77.
6 Ioannidis J. The infection fatality rate of COVID-19 inferred from seroprevalence data. medRxiv 2020.
7 Russell TW, Hellewell J, *et al*, CMMID nCov working group. Using a delay-adjusted case fatality ratio to estimate under-reporting. *Centre for Mathematical Modeling of Infectious Diseases Repository* 2020.
8 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. *J Clin Med* 2020;9:538.
9 Spearman C. The proof and measurement of association between two things. *Am J Psychol* 1904;15:72–101 www.jstor.org/stable/1412159
10 Long Q-X, Liu B-Z, Deng H-J, *et al*. Antibody responses to SARS-CoV-2 in patients with COVID-19. *Nature Med* 2020:1–4.
11 Whitman JD, Hiatt J, Mowery CT, *et al*. Test performance evaluation of SARS-CoV-2 serological assays. *medRxiv* 2020 doi:10.1101/2020.04.25.20074856