BMJ Open

Estimating under-reporting of Covid-19 cases in Indian states using a delay-adjusted case fatality ratio

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2020-042584
Article Type:	Original research
Date Submitted by the Author:	10-Jul-2020
Complete List of Authors:	Unnikrishnan, Jayakrishnan ; QUALCOMM Inc Mangalathu, Sujith; Equifax Inc; Equifax Inc, Kutty, Raman; Sree Chitra Tirunal Institute for Medical Sciences and Technology
Keywords:	INFECTIOUS DISEASES, STATISTICS & RESEARCH METHODS, Public health < INFECTIOUS DISEASES

I, the Submitting Author has the right to grant and does grant on behalf of all authors of the Work (as defined in the below author licence), an exclusive licence and/or a non-exclusive licence for contributions from authors who are: i) UK Crown employees; ii) where BMJ has agreed a CC-BY licence shall apply, and/or iii) in accordance with the terms applicable for US Federal Government officers or employees acting as part of their official duties; on a worldwide, perpetual, irrevocable, royalty-free basis to BMJ Publishing Group Ltd ("BMJ") its licensees and where the relevant Journal is co-owned by BMJ to the co-owners of the Journal, to publish the Work in this journal and any other BMJ products and to exploit all rights, as set out in our licence.

The Submitting Author accepts and understands that any supply made under these terms is made by BMJ to the Submitting Author unless you are acting as an employee on behalf of your employer or a postgraduate student of an affiliated institution which is paying any applicable article publishing charge ("APC") for Open Access articles. Where the Submitting Author wishes to make the Work available on an Open Access basis (and intends to pay the relevant APC), the terms of reuse of such Open Access shall be governed by a Creative Commons licence – details of these licences and which Creative Commons licence will apply to this Work are set out in our licence referred to above.

Other than as permitted in any relevant BMJ Author's Self Archiving Policies, I confirm this Work has not been accepted for publication elsewhere, is not being considered for publication elsewhere and does not duplicate material already published. I confirm all authors consent to publication of this Work and authorise the granting of this licence.

RESEARCH

Estimating under-reporting of Covid-19 cases in Indian states using a delay-adjusted case fatality ratio

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

[*]Correspondence:
sujithmangalath@ucla.edu
[2]Equifax Inc, 1505 Windward Concourse, 30005 Alpharetta, USA.

[1]Qualcomm Inc., 500 Somerset Corporate Blvd, Bridgewater, NJ, USA

[3]Research Director, Amala Cancer Research Centre, 680555 Thrissur, India.

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 we obtain time-varying estimates of the level of under-reporting rate of Covid-19 infections in the different states.

Methods: Following methodology developed in prior work, we use a delay-adjusted case fatality ratio to estimate the true under-reporting rate 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 under-reporting rate.

Setting: We apply this method of analysis to all Indian states reporting at least 10 deaths as of 24 June 2020.

Results: Our analysis suggests that delay-adjusted case fatality ratios observed in different states range from 0.6% to 7.6%, and that the fraction of cases reported in different states range from 18% to 100% for an assumed baseline case fatality ratio of 1.38%, and from 8.6% to 100% for an assumed baseline case fatality ratio of 0.66%. 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 prioritizing testing in different states, and also to analyze the time-varying effects of different quarantine measures adopted in different states.

Keywords: Covid-19; Under-reporting; India

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 serologic 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 that is used, and 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 20 June 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 [1] suggests that the tests conducted per million residents in various states ranges from 1465 to 45437. Although patients hospitalized with symptoms are typically tested, those who develop mild symptoms at home and those who do not develop symptoms are unlikely to be tested. 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 level of under-reporting existing in different states can help us determine the true extent of the infection.

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 [1]. This data is 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 24 June 2020. In addition, for illustration and for studying the relationship of under-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 is 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 under-counts additional deaths that may arise from the cases observed to date. This issue can be addressed by using the distribution of delay from hospitalization 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% has been reported in [2]. A different published study based on data from China puts the estimate at 0.66% [3]. More recently, the US Centers for Disease Control and Prevention reports a best estimate of 0.4% [4].

However, in regions where cases have been under-reported, we expect the adjusted-CFR to be significantly 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 [5] and apply it to data from different states of India. For completeness, we elaborate on details of the method below.

Method details

Following [5] we assume that for fatal cases, the delay from confirmation to death follows the same distribution as delay from hospitalization to death estimated in [6]. 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 standard deviation 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 reported on day s from a region. In this case, 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 cumulative sum of e_t to the cumulative number of deaths reported by day t provides an estimate of the average under-reporting in the region, over the duration of the pandemic.

We can further improve the estimate to obtain a time-varying estimate of the under-reporting rate. 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 today's death rate. Hence as we assume a mean delay of 13 days from case confirmation to death, the quantity 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 Generalize Additive Model using the pyGAM Python package. We applied this method to all states with at least 10 reported deaths.

As the root cause of under-reporting is the insufficient coverage of testing among infected people, we expect to have higher under-reporting when a larger fraction of tested people test positive. Thus we expect a negative relation between fraction of cases reported and the test positivity rate, 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 one day 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 analyze 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 of 1.38% for Covid-19. 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 and Discussion

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 24 June calculated using the Poisson regression approach. Due to lack of sufficient data, we do not estimate postivity 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 1.2 to 2.0, 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 0.66% and 1.38%. These numbers are obtained by the ratio 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 an assumed baseline CFR of 0.66% compared to that for an assumed baseline CFR of 1.38%.

The time-varying estimates of the fraction reported f_t for different regions are illustrated in Figure 1 for an assumed a baseline CFR of 1.38% for Covid-19. The red curves show the estimate of the fraction reported obtained and the shaded region represents the associated 95% confidence bounds for the Poisson regression model. For lower values of the baseline CFR, the estimate of the fraction reported would be even lower than what is shown in this figure. In the same figures, we also plot the test positivity rates obtained in each state.

In Figure 2, 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 3 shows a scatter-plot of the estimate of the fraction reported against the test positivity rate. 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 is available (24 June 2020), assuming a baseline CFR of 1.38%. The quantity plotted on the horizontal axis is the convolution of the positivity rate from past days where the filter is given by the distribution p of the delay from case to death, evaluated on the same day. We observe that the two states with the highest positivity rate are also the ones for which the estimate of fraction reported is the lowest. In the

figure, we also show a regression line t of $\log(y)$ vs x, which yields an r^2-value of 0.5 and a p-value of 0.001, indicating a statistically significant relationship. Thus an increase in test positivity rate is associated with a decrease in the fraction reported.

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 lower than what is assumed, the estimate of the fraction reported would increase proportionately. Furthermore, if the number of deaths reported is under-counted, or 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 under-reporting rate.

An alternative method one could adopt to quantify under-reporting may be to use serologic testing [7, 8] for Covid-19 antibodies among the general public. Randomized antibody testing in a general population could 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.

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 statistically significant relationship with the test positivity rate.

The estimate of under-reporting may be used to guide policies for prioritizing 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.

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 [5] for sharing their work and code online, and Timothy Russell for answering our questions on the method and code.

Contributors

JU adapted and implemented the statistical model. JU and SM wrote the paper. All authors critically reviewed the approach and the manuscript and gave approval for the publication.

Competing interests

The authors declare that they have no competing interests.

Patient and Public Involvement

Patients or the public were not involved in the design, or conduct, or reporting, or dissemination plans of our research.

Patient Consent for Publication

Not required

Ethics approval

Not required

Data availability statement

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].

Exclusive license

The Corresponding Author has the right to grant on behalf of all authors and does grant on behalf of all authors, a worldwide license to the Publishers and its licensees in perpetuity, in all forms, formats and media (whether known now or created in the future), to i) publish, reproduce, distribute, display and store the Contribution, ii) translate the Contribution into other languages, create adaptations, reprints, include within collections and create summaries, extracts and/or, abstracts of the Contribution, iii) create any other derivative work(s) based on the Contribution, iv) to exploit all subsidiary rights in the Contribution, v) the inclusion of electronic links from the Contribution to third party material where-ever it may be located; and, vi) license any third party to do any or all of the above.

References

1. COVID19-India API. https://api.covid19india.org
2. Guan, W.-j., Ni, Z.-y., Hu, Y., Liang, W.-h., Ou, C.-q., He, J.-x., Liu, L., Shan, H., Lei, C.-l., Hui, D.S.C., Du, B., Li, L.-j., Zeng, G., Yuen, K.-Y., Chen, R.-c., Tang, C.-l., Wang, T., Chen, P.-y., Xiang, J., Li, S.-y., Wang, J.-l., Liang, Z.-j., Peng, Y.-x., Wei, L., Liu, Y., Hu, Y.-h., Peng, P., Wang, J.-m., Liu, J.-y., Chen, Z., Li, G., Zheng, Z.-j., Qiu, S.-q., Luo, J., Ye, C.-j., Zhu, S.-y., Zhong, N.-s.: Brca1clinical characteristics of coronavirus disease 2019 in china. New England Journal of Medicine 382(18), 1708{1720 (2020)
3. Verity, R., Okell, L.C., Dorigatti, I., Winskill, P., Whittaker, C., Imai, N., Cuomo-Dannenburg, G., Thompson, H., Walker, P.G.T., Fu, H., Dighe, A., Gri n, J.T., Baguelin, M., Bhatia, S., Boonyasiri, A., Cori, A., Cucunuba, Z., FitzJohn, R., Gaythorpe, K., Green, W., Hamlet, A., Hinsley, W., Laydon, D., Nedjati-Gilani, G., Riley, S., van Elsland, S., Volz, E., Wang, H., Wang, Y., Xi, X., Donnelly, C.A., Ghani, A.C., Ferguson, N.M.: Estimates of the severity of coronavirus disease 2019: a model-based analysis. The Lancet Infectious Diseases 20(6), 669{677 (2020)
4. Centers for Disease Control and Prevention, USA: COVID-19 Pandemic Planning Scenarios. https://www.cdc.gov/coronavirus/2019-ncov/hcp/planning-scenarios.html
5. Russell, T.W., Hellewell, J., Abbott, S., Golding, N., Gibbs, H., Jarvis, C.I., van Zandvoort, K., CMMID nCov working group, Flasche, S., Eggo, R.M., Edmunds, W.J., Kucharski, A.J.: Using a Delay-adjusted Case Fatality Ratio to Estimate Under-reporting. https://cmmid.github.io/topics/covid19/global cfr estimates.html
6. Linton, N.M., Kobayashi, T., Yang, Y., Hayashi, K., Akhmetzhanov, A.R., Jung, S.-m., Yuan, B., Kinoshita, R., Nishiura, H.: Incubation period and other epidemiological characteristics of 2019 novel coronavirus infections with right truncation: A statistical analysis of publicly available case data. Journal of Clinical Medicine 9(2), 538 (2020)
7. Long, Q.-X., Liu, B.-Z., Deng, H.-J., Wu, G.-C., Deng, K., Chen, Y.-K., Liao, P., Qiu, J.-F., Lin, Y., Cai, X.-F., Wang, D.-Q., Hu, Y., Ren, J.-H., Tang, N., Xu, Y.-Y., Yu, L.-H., Mo, Z., Gong, F., Zhang, X.-L., Tian, W.-G., Hu, L., Zhang, X.-X., Xiang, J.-L., Du, H.-X., Liu, H.-W., Lang, C.-H., Luo, X.-H., Wu, S.-B., Cui, X.-P., Zhou, Z., Zhu, M.-M., Wang, J., Xue, C.-J., Li, X.-F., Wang, L., Li, Z.-J., Wang, K., Niu, C.-C., Yang, Q.-J., Tang, X.-J., Zhang, Y., Liu, X.-M., Li, J.-J., Zhang, D.-C., Zhang, F., Liu, P., Yuan, J., Li, Q., Hu, J.-L., Chen, J., Huang, A.-L.:

Antibody responses to SARS-CoV-2 in patients with COVID-19. Nature Medicine 26(6), 845{848 (2020)

8. Whitman, J.D., Hiatt, J., Mowery, C.T., Shy, B.R., Yu, R., Yamamoto, T.N., Rathore, U., Goldgof, G.M., Whitty, C., Woo, J.M., Gallman, A.E., Miller, T.E., Levine, A.G., Nguyen, D.N., Bapat, S.P., Balcerek, J., Bylsma, S.A., Lyons, A.M., Li, S., Wong, A.W.-Y., Gillis-Buck, E.M., Steinhart, Z.B., Lee, Y., Apathy, R., Lipke, M.J., Smith, J.A., Zheng, T., Boothby, I.C., Isaza, E., Chan, J., Acenas, n. Dante D, Lee, J., Macrae, T.A., Kyaw, T.S., Wu, D., Ng, D.L., Gu, W., York, V.A., Eskandarian, H.A., Callaway, P.C., Warrier, L., Moreno, M.E., Levan, J., Torres, L., Farrington, L.A., Loudermilk, R., Koshal, K., Zorn, K.C., Garcia-Beltran, W.F., Yang, D., Astudillo, M.G., Bernstein, B.E., Gelfand, J.A., Ryan, E.T., Charles, R.C., Iafrate, A.J., Lennerz, J.K., Miller, S., Chiu, C.Y., Stramer, S.L., Wilson, M.R., Manglik, A., Ye, C.J., Krogan, N.J., Anderson, M.S., Cyster, J.G., Ernst, J.D., Wu, A.H.B., Lynch, K.L., Bern, C., Hsu, P.D., Marson, A.: Test performance evaluation of SARS-CoV-2 serological assays. medRxiv, 2020{042520074856 (2020)

Figures

Figure 1. Curves in red show the estimates of under reporting 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. Note that due to lack of sufficient data, we do not estimate postivity rate for India and Telangana

Figure 2. 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 3 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

Table 1. Under-reporting estimates for different states

State	Deaths	Cases	Test positivity rate	nCFR	cCFR	Percentage reported (CFR of 1.38%) [%]	Percentage reported (CFR of 0.66%) [%]
India	14907	472882	-	3.15	4.61	14.31	29.92
Andhra Pradesh	129	10331	3.9	1.25	2.03	32.55	68.05
Bihar	55	8273	3.3	0.66	0.91	72.71	100.00
Chhattisgarh	12	2419	3.3	0.50	0.78	85.04	100.00
Delhi	2365	70390	25	3.36	5.71	11.55	24.15
Gujarat	1736	29001	10.8	5.99	7.64	8.64	18.07
Haryana	188	12010	10.5	1.57	2.63	25.10	52.48
Jammu and Kashmir	88	6422	2.8	1.37	1.88	35.07	73.33
Jharkhand	12	2219	3.2	0.54	0.76	86.62	100.00
Karnataka	166	10118	3.8	1.64	2.47	26.68	55.79
Kerala	23	3604	2.4	0.64	0.97	67.95	100.00
Madhya Pradesh	535	12448	2.5	4.30	5.22	12.65	26.45
Maharashtra	6738	142899	33.5	4.72	6.51	10.14	21.20
Odisha	24	5752	5.7	0.42	0.64	100.00	100.00
Punjab	114	4630	1.4	2.46	3.55	18.59	38.87
Rajasthan	375	16009	2.4	2.34	3.07	21.50	44.95
Tamil Nadu	866	67468	8.6	1.28	2.00	33.07	69.15
Telangana	225	10444	-*	2.15	4.28	15.43	32.27
Uttar Pradesh	596	19557	4.2	3.05	4.45	14.84	31.03
Uttarakhand	35	2623	6.6	1.33	2.02	32.63	68.22
West Bengal	591	15173	5.1	3.90	5.61	11.77	24.60

* not enough data

Figures

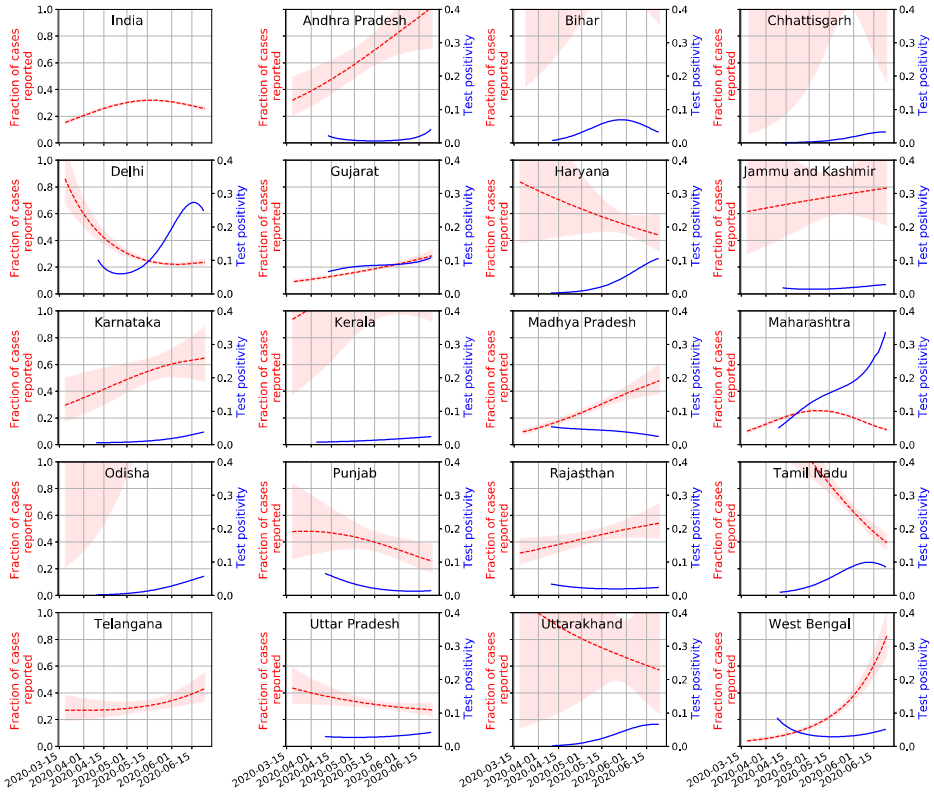


Figure 1 Curves in red show the estimates of under reporting in various regions as a function of time, assuming a baseline CFR of 1.38%. The first subplot shows the results for India and the other subplots show results for the top 19 states with most reported deaths. Curves in blue show the smoothed estimate of test positivity rate. Note that due to lack of sufficient data, we do not estimate postivity rate for India and Telangana.

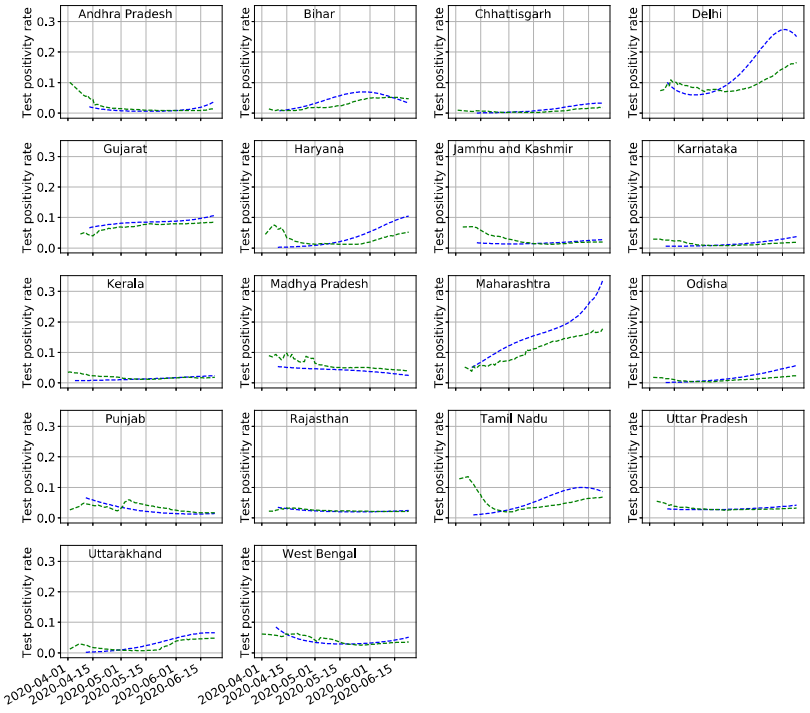


Figure 2 Curves in blue shows the test positivity rate estimated via the Poisson regression method. Curves in green show the ratio of cumulative positive cases to cumulative tests performed.

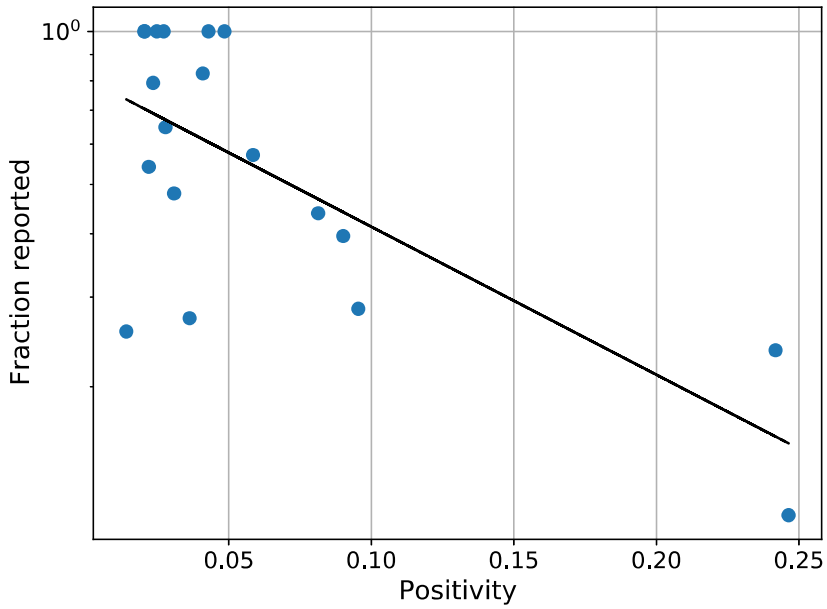


Figure 3 Scatter plot of the estimate of the fraction f_t of cases reported from different states evaluated on the last date considered, against the corresponding test positivity rate

BMJ Open

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

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2020-042584.R1
Article Type:	Original research
Date Submitted by the Author:	13-Oct-2020
Complete List of Authors:	Unnikrishnan, Jayakrishnan ; QUALCOMM Inc Mangalathu, Sujith; Equifax Inc; Equifax Inc, Kutty, Raman; Sree Chitra Tirunal Institute for Medical Sciences and Technology
Primary Subject Heading:	Epidemiology
Secondary Subject Heading:	Epidemiology, Global health, Health policy, Infectious diseases
Keywords:	INFECTIOUS DISEASES, STATISTICS & RESEARCH METHODS, Public health < INFECTIOUS DISEASES

I, the Submitting Author has the right to grant and does grant on behalf of all authors of the Work (as defined in the below author licence), an exclusive licence and/or a non-exclusive licence for contributions from authors who are: i) UK Crown employees; ii) where BMJ has agreed a CC-BY licence shall apply, and/or iii) in accordance with the terms applicable for US Federal Government officers or employees acting as part of their official duties; on a worldwide, perpetual, irrevocable, royalty-free basis to BMJ Publishing Group Ltd ("BMJ") its licensees and where the relevant Journal is co-owned by BMJ to the co-owners of the Journal, to publish the Work in this journal and any other BMJ products and to exploit all rights, as set out in our licence.

The Submitting Author accepts and understands that any supply made under these terms is made by BMJ to the Submitting Author unless you are acting as an employee on behalf of your employer or a postgraduate student of an affiliated institution which is paying any applicable article publishing charge ("APC") for Open Access articles. Where the Submitting Author wishes to make the Work available on an Open Access basis (and intends to pay the relevant APC), the terms of reuse of such Open Access shall be governed by a Creative Commons licence – details of these licences and which Creative Commons licence will apply to this Work are set out in our licence referred to above.

Other than as permitted in any relevant BMJ Author's Self Archiving Policies, I confirm this Work has not been accepted for publication elsewhere, is not being considered for publication elsewhere and does not duplicate material already published. I confirm all authors consent to publication of this Work and authorise the granting of this licence.

RESEARCH

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*], and Raman V Kutty[3]

[*]Correspondence:
sujithmangalath@ucla.edu
[2]Equifax Inc, 1505 Windward Concourse, 30005 Alpharetta, USA.

[1]Qualcomm Inc., 500 Somerset Corporate Blvd, Bridgewater, NJ, USA

[3]Research Director, Amala Cancer Research Centre, 680555 Thrissur, India.

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 prioritizing testing in different states, and also to analyze the time-varying effects of different quarantine measures adopted in different states.

Keywords: Covid-19; Under-reporting; India

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 serologic 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 [1] suggests that the tests conducted up to October 10, 2020, in various states range from 29 to 182 per thousand residents. Although patients hospitalized with symptoms are typically tested, those who develop mild symptoms at home and those who do not develop symptoms are unlikely to be tested. 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 level of under-reporting existing in different states can help us determine the true extent of the infection.

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 [1]. This data is 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 is 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 under-counts additional deaths that may arise from the cases observed to date. This issue can be addressed by using the distribution of delay from hospitalization 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 [2]. A different published study based on data from China puts the estimate at 0.66% [3]. More recent reports based on seroprevalence studies provide much lower estimates as low as 0.1% [4].

However, in regions where cases have been under-reported, we expect the adjusted-CFR to be significantly 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 [5] 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 [5] we assume that for fatal cases, the delay from confirmation to death follows the same distribution as delay from hospitalization to death estimated in [6]. 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 standard deviation 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 is 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 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 Generalize 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 one day 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 analyze 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 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 to 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%

confidence 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 is available (10 October 2020), assuming a baseline CFR of 0.1%. As mentioned earlier, f_t 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 highest positivity rate also tend to have low estimates of the fraction of cases reported. The Spearman's rank correlation coefficient [7] between these two quantities is -0.4 with a p-value of 0.03 indicating a statistically significant negative relation. In the figure, we also show a regression line fit of $\log(y)$ vs x, which yields an r^2-value of 0.17 and a p-value of 0.04. Thus, an increase in test positivity rate is associated with a decrease in the fraction 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 serologic testing [8, 9] for Covid-19 antibodies among the general public. Randomized 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 under-counted, 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 statistically significant relationship with the test positivity rate.

The estimate of under-reporting may be used to guide policies for prioritizing 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.

Author Affiliations

[1]Qualcomm Inc., 500 Somerset Corporate Blvd, Bridgewater, NJ, USA
[2]Equifax Inc, 1505 Windward Concourse, 30005 Alpharetta, USA.
[3]Research Director, Amala Cancer Research Centre, 680555 Thrissur, India.

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 [5] 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, 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
The authors declare that they have no competing interests.

Patient and Public Involvement
Patients or the public were not involved in the design, or conduct, or reporting, or dissemination plans of our research.

Patient Consent for Publication
Not required

Ethics approval
Not required

Data availability statement
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].

References

1. COVID19-India API. https://api.covid19india.org
2. Guan, W.-j., Ni, Z.-y., Hu, Y., Liang, W.-h., Ou, C.-q., He, J.-x., Liu, L., Shan, H., Lei, C.-l., Hui, D.S.C., Du, B., Li, L.-j., Zeng, G., Yuen, K.-Y., Chen, R.-c., Tang, C.-l., Wang, T., Chen, P.-y., Xiang, J., Li, S.-y., Wang, J.-l., Liang, Z.-j., Peng, Y.-x., Wei, L., Liu, Y., Hu, Y.-h., Peng, P., Wang, J.-m., Liu, J.-y., Chen, Z., Li, G., Zheng, Z.-j., Qiu, S.-q., Luo, J., Ye, C.-j., Zhu, S.-y., Zhong, N.-s.: Brca1clinical characteristics of coronavirus disease 2019 in china. New England Journal of Medicine 382(18), 1708{1720 (2020)
3. Verity, R., Okell, L.C., Dorigatti, I., Winskill, P., Whittaker, C., Imai, N., Cuomo-Dannenburg, G., Thompson, H., Walker, P.G.T., Fu, H., Dighe, A., Gri n, J.T., Baguelin, M., Bhatia, S., Boonyasiri, A., Cori, A., Cucunuba, Z., FitzJohn, R., Gaythorpe, K., Green, W., Hamlet, A., Hinsley, W., Laydon, D., Nedjati-Gilani, G., Riley, S., van Elsland, S., Volz, E., Wang, H., Wang, Y., Xi, X., Donnelly, C.A., Ghani, A.C., Ferguson, N.M.: Estimates of the severity of coronavirus disease 2019: a model-based analysis. The Lancet Infectious Diseases 20(6), 669{677 (2020)
4. Ioannidis, J., The infection fatality rate of COVID-19 inferred from seroprevalence data. medRxiv. doi: https://doi.org/10.1101/2020.05.13.20101253 July 14, 2020.
5. Russell, T. W., Hellewell, J., Abbott, S., Jarvis, C. I., van Zandvoort, K., CMMID nCov working group, ... & Kucharski, A. J. (2020). Using a delay-adjusted case fatality ratio to estimate under-reporting. Centre for Mathematical Modeling of Infectious Diseases Repository
6. Linton, N.M., Kobayashi, T., Yang, Y., Hayashi, K., Akhmetzhanov, A.R., Jung, S.-m., Yuan, B., Kinoshita, R., Nishiura, H.: Incubation period and other epidemiological characteristics of 2019 novel coronavirus infections with right truncation: A statistical analysis of publicly available case data. Journal of Clinical Medicine 9(2), 538 (2020)
7. Spearman, C. “The Proof and Measurement of Association between Two Things.” The American Journal of Psychology, vol. 15, no. 1, 1904, pp. 72–101. JSTOR, www.jstor.org/stable/1412159. Accessed 11 Oct. 2020.
8. Long, Q.-X., Liu, B.-Z., Deng, H.-J., Wu, G.-C., Deng, K., Chen, Y.-K., Liao, P., Qiu, J.-F., Lin, Y., Cai, X.-F., Wang, D.-Q., Hu, Y., Ren, J.-H., Tang, N., Xu, Y.-Y., Yu, L.-H., Mo, Z., Gong, F., Zhang, X.-L., Tian, W.-G., Hu, L., Zhang, X.-X., Xiang, J.-L., Du, H.-X., Liu, H.-W., Lang, C.-H., Luo, X.-H., Wu, S.-B., Cui, X.-P., Zhou, Z., Zhu, M.-M., Wang, J., Xue, C.-J., Li, X.-F., Wang, L., Li, Z.-J., Wang, K., Niu, C.-C., Yang, Q.-J., Tang, X.-J., Zhang, Y., Liu, X.-M., Li, J.-J., Zhang, D.-C., Zhang, F., Liu, P., Yuan, J., Li, Q., Hu, J.-L., Chen, J., Huang, A.-L.: Antibody responses to SARS-CoV-2 in patients with COVID-19. Nature Medicine, 2020, 1-4
9. Whitman, J.D., Hiatt, J., Mowery, C.T., Shy, B.R., Yu, R., Yamamoto, T.N., Rathore, U., Goldgof, G.M., Whitty, C., Woo, J.M., Gallman, A.E., Miller, T.E., Levine, A.G., Nguyen, D.N., Bapat, S.P., Balcerek, J., Bylsma, S.A., Lyons, A.M., Li, S., Wong, A.W.-Y., Gillis-Buck, E.M., Steinhart, Z.B., Lee, Y., Apathy, R., Lipke, M.J., Smith, J.A., Zheng, T., Boothby, I.C., Isaza, E., Chan, J., Acenas, n. Dante D, Lee, J., Macrae, T.A., Kyaw, T.S., Wu, D., Ng, D.L., Gu, W., York, V.A., Eskandarian, H.A., Callaway, P.C., Warrier, L., Moreno, M.E., Levan, J., Torres, L., Farrington, L.A., Loudermilk, R., Koshal, K., Zorn, K.C., Garcia-Beltran, W.F., Yang, D., Astudillo, M.G., Bernstein, B.E., Gelfand, J.A., Ryan, E.T., Charles, R.C., Iafrate, A.J., Lennerz, J.K., Miller, S., Chiu, C.Y., Stramer, S.L., Wilson, M.R., Manglik, A., Ye, C.J., Krogan, N.J., Anderson, M.S., Cyster, J.G., Ernst, J.D., Wu, A.H.B., Lynch, K.L., Bern, C., Hsu, P.D., Marson, A.: Test performance evaluation of SARS-CoV-2 serological assays. medRxiv, 2020

Figures

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.

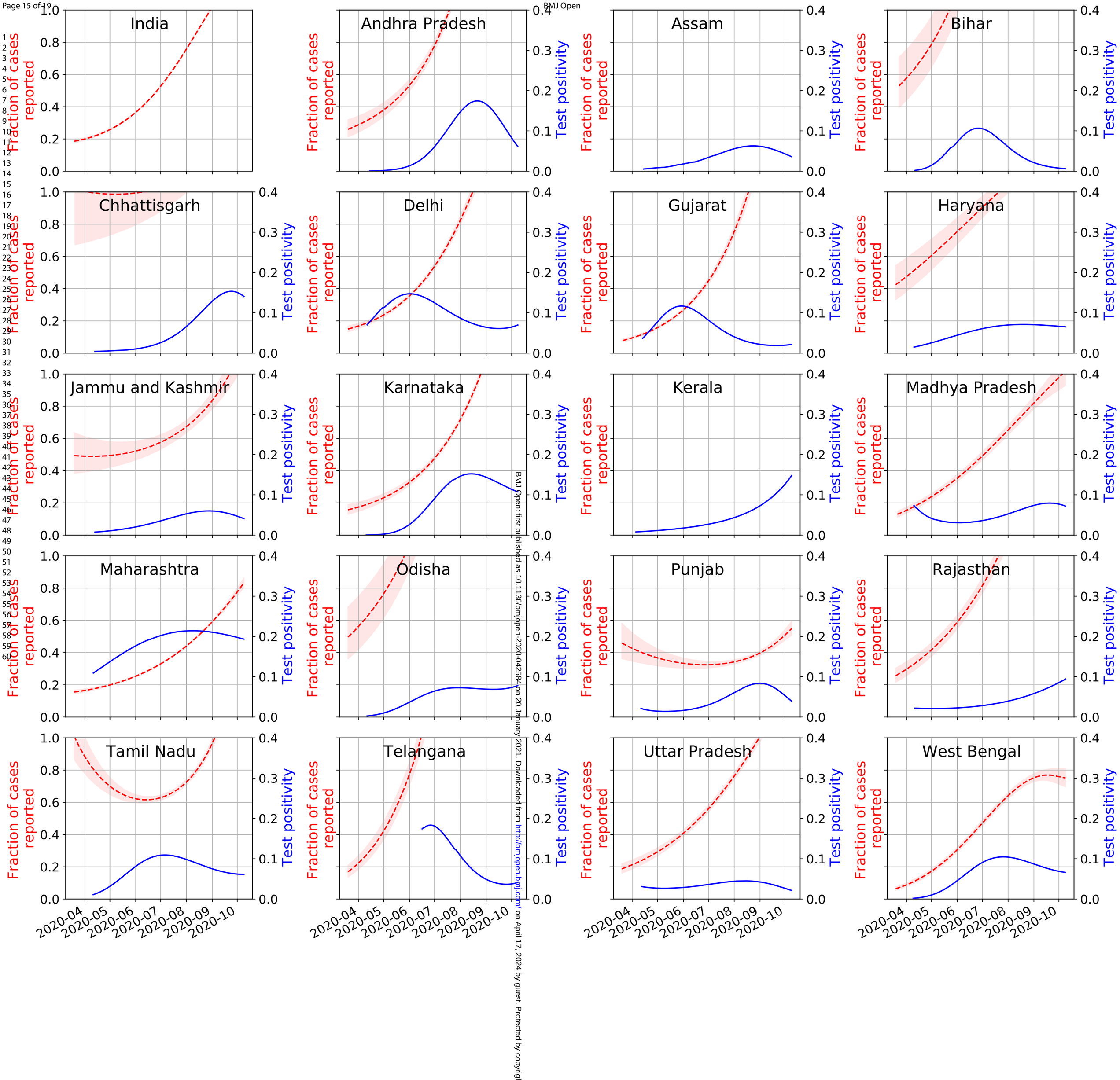
Figure 2. Curves in red show the estimates of the fraction of cases reported in various regions as a function of time, assuming a baseline CFR of 0.1%. The first subplot shows the results for India and the other subplots show results for the top 19 states with most reported deaths. Curves in blue show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

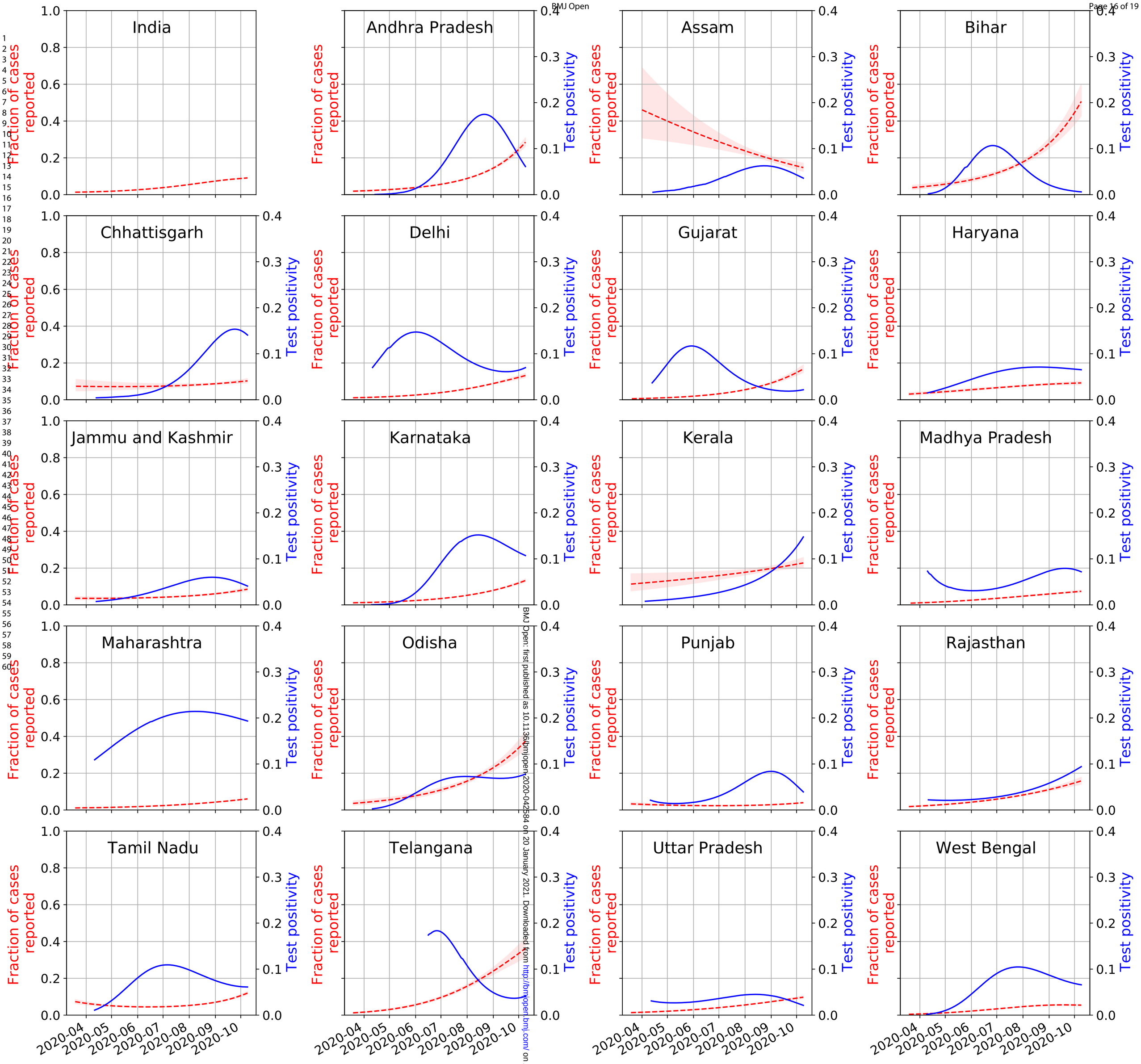
Figure 3. Curves in blue shows the test positivity rate estimated via the Poisson regression method. Curves in green show the ratio of cumulative positive cases to cumulative tests performed.

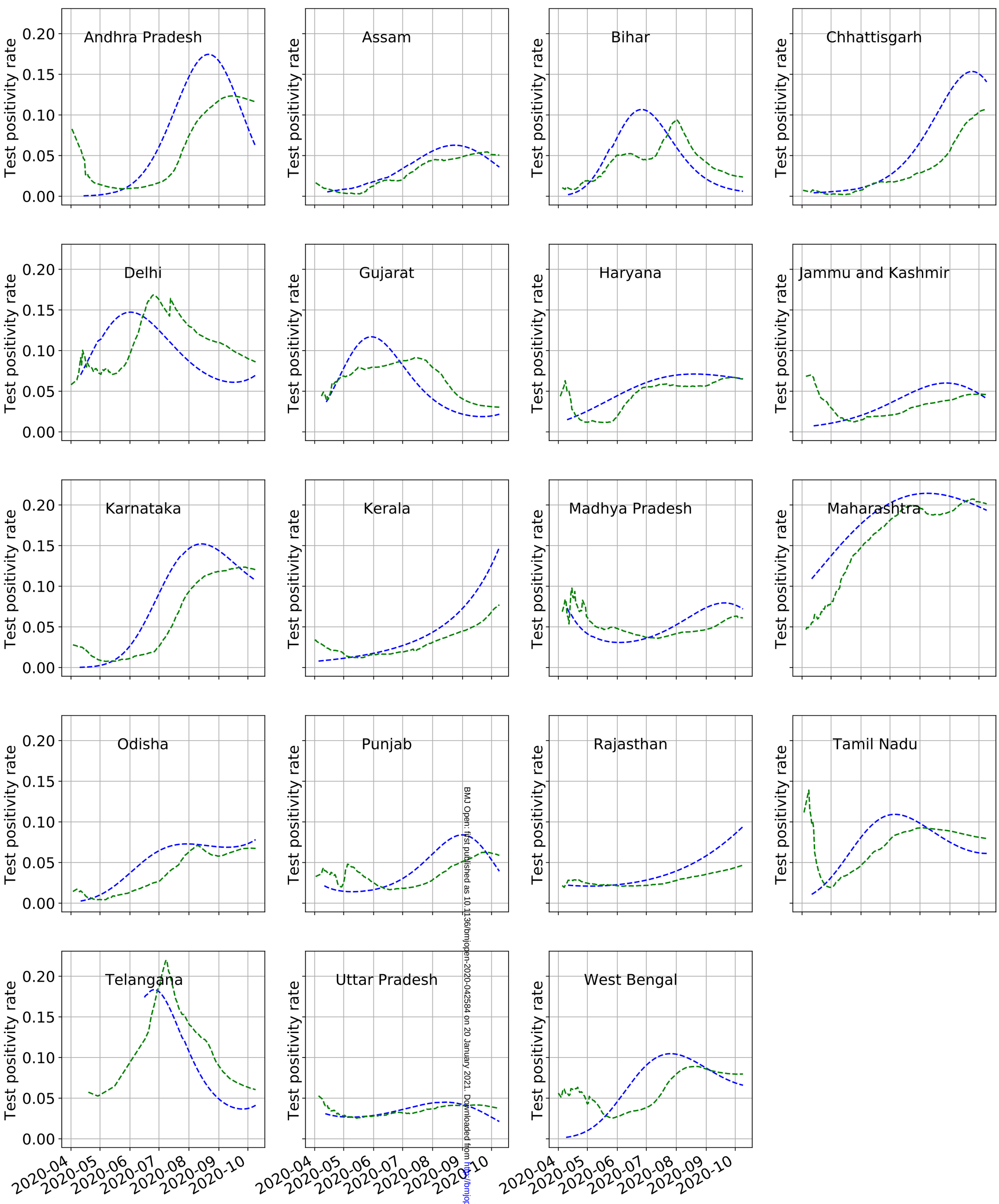
Figure 4. Scatter plot of the estimate of the fraction f_i of cases reported from different states evaluated on the last date considered, against the corresponding test positivity rate

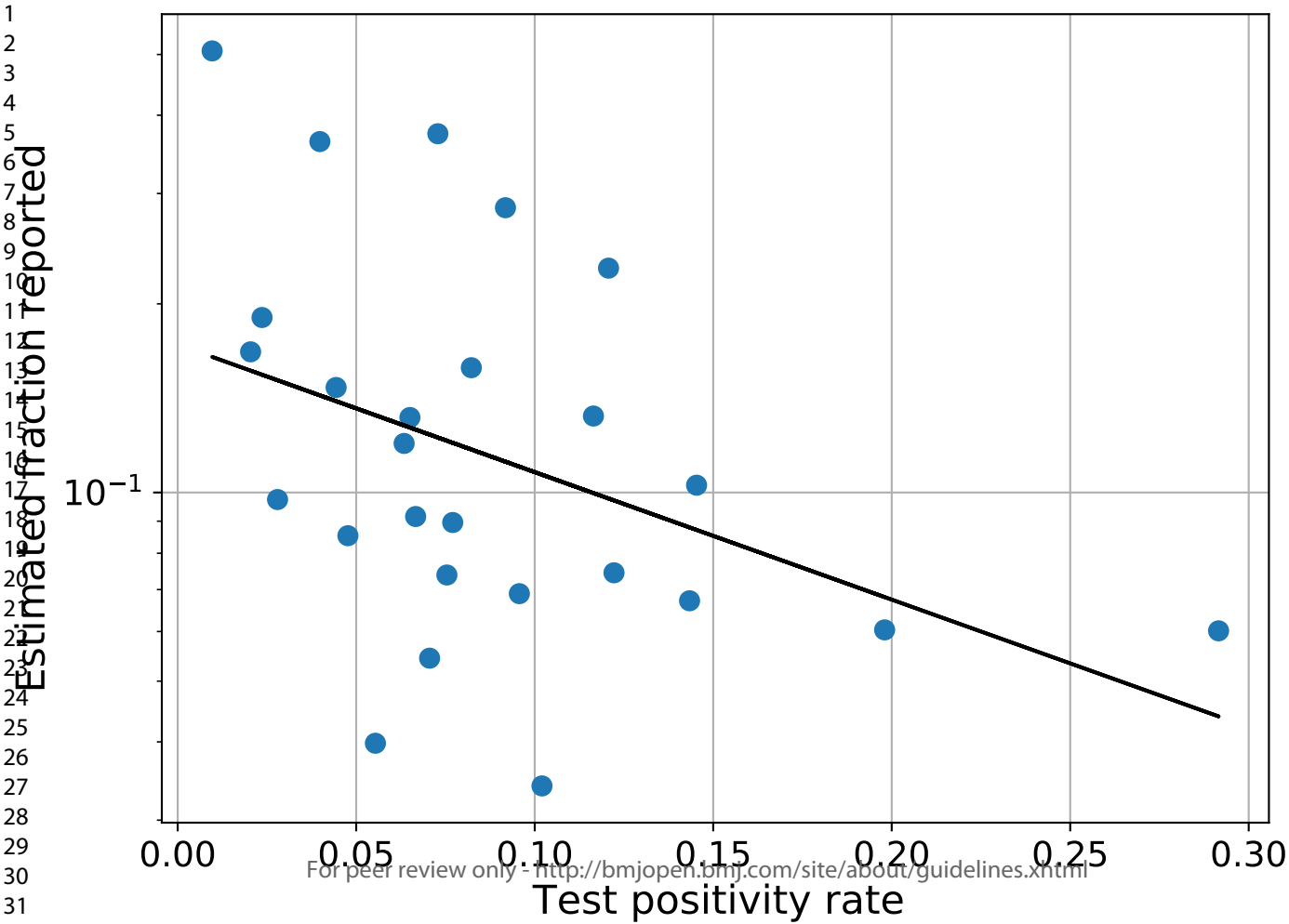
Table 1. Estimates of fraction of cases reported in different states

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	106863	6976461	-	1.53	1.78	77.62	37.12	5.62
Andhra Pradesh	6159	744864	6.1	0.83	0.93	100.00	71.07	10.77
Assam	807	192314	3.6	0.42	0.47	100.00	100.00	21.11
Bihar	934	193826	0.6	0.48	0.53	100.00	100.00	18.92
Chhattisgarh	1196	137570	14.1	0.87	1.14	100.00	57.86	8.77
Delhi	5692	303693	7.0	1.87	2.13	64.85	31.01	4.70
Gujarat	3549	149193	2.2	2.38	2.68	51.59	24.67	3.74
Haryana	1562	139932	6.5	1.12	1.29	100.00	51.13	7.75
Jammu and Kashmir	1306	82429	4.1	1.58	1.84	74.84	35.79	5.42
Karnataka	9200	690269	10.7	1.33	1.60	86.35	41.30	6.26
Kerala	956	268101	14.8	0.36	0.51	100.00	100.00	19.53
Madhya Pradesh	2575	143629	7.2	1.79	2.14	64.57	30.88	4.68
Maharashtra	39731	1506018	19.3	2.64	3.02	45.67	21.84	3.31
Odisha	1044	246839	7.8	0.42	0.51	100.00	100.00	19.70
Punjab	3774	122462	3.9	3.08	3.55	38.88	18.59	2.82
Rajasthan	1621	154785	9.4	1.05	1.25	100.00	52.81	8.00
Tamil Nadu	10120	646128	6.1	1.57	1.75	78.80	37.69	5.71
Telangana	1208	208025	4.1	0.58	0.66	100.00	100.00	15.18
Uttar Pradesh	6293	430666	2.1	1.46	1.66	83.16	39.77	6.03
West Bengal	5501	287603	6.6	1.91	2.23	61.89	29.60	4.49









STROBE Statement—Checklist of items that should be included in reports of ***cross-sectional studies***

	Item No	Page No.	Recommendation
Title and abstract	1	1	(*a*) Indicate the study's design with a commonly used term in the title or the abstract
		1	(*b*) Provide in the abstract an informative and balanced summary of what was done and what was found
			Introduction
Background/rationale	2	2	Explain the scientific background and rationale for the investigation being reported
Objectives	3	2	State specific objectives, including any prespecified hypotheses
			Methods
Study design	4	2,3	Present key elements of study design early in the paper
Setting	5	2,3	Describe the setting, locations, and relevant dates, including periods of recruitment, exposure, follow-up, and data collection
Participants	6	3	(*a*) Give the eligibility criteria, and the sources and methods of selection of participants
Variables	7	3	Clearly define all outcomes, exposures, predictors, potential confounders, and effect modifiers. Give diagnostic criteria, if applicable
Data sources/ measurement	8*	*3*	For each variable of interest, give sources of data and details of methods of assessment (measurement). Describe comparability of assessment methods if there is more than one group
Bias	9	3,4	Describe any efforts to address potential sources of bias
Study size	10	3,4	Explain how the study size was arrived at
Quantitative variables	11	3.4	Explain how quantitative variables were handled in the analyses. If applicable, describe which groupings were chosen and why
Statistical methods	12	4	(*a*) Describe all statistical methods, including those used to control for confounding
		4	(*b*) Describe any methods used to examine subgroups and interactions
		NA	(*c*) Explain how missing data were addressed
		NA	(*d*) If applicable, describe analytical methods taking account of sampling strategy
		4,5	(*e*) Describe any sensitivity analyses
			Results
Participants	13*	6,7	(a) Report numbers of individuals at each stage of study—eg numbers potentially eligible, examined for eligibility, confirmed eligible, included in the study, completing follow-up, and analysed
		NA	(b) Give reasons for non-participation at each stage
		NA	(c) Consider use of a flow diagram
Descriptive data	14*	6,7	(a) Give characteristics of study participants (eg demographic, clinical, social) and information on exposures and potential confounders
		NA	(b) Indicate number of participants with missing data for each variable of interest

Outcome data	15*	NA	Report numbers of outcome events or summary measures
Main results	16	6,7	(*a*) Give unadjusted estimates and, if applicable, confounder-adjusted estimates and their precision (eg, 95% confidence interval). Make clear which confounders were adjusted for and why they were included
			(*b*) Report category boundaries when continuous variables were categorized
			(*c*) If relevant, consider translating estimates of relative risk into absolute risk for a meaningful time period
Other analyses	17	7	Report other analyses done—eg analyses of subgroups and interactions, and sensitivity analyses
Discussion			
Key results	18	7,8	Summarise key results with reference to study objectives
Limitations	19	8	Discuss limitations of the study, taking into account sources of potential bias or imprecision. Discuss both direction and magnitude of any potential bias
Interpretation	20	8	Give a cautious overall interpretation of results considering objectives, limitations, multiplicity of analyses, results from similar studies, and other relevant evidence
Generalisability	21	8,9	Discuss the generalisability (external validity) of the study results
Other information			
Funding	22	10	Give the source of funding and the role of the funders for the present study and, if applicable, for the original study on which the present article is based

*Give information separately for exposed and unexposed groups.

Note: An Explanation and Elaboration article discusses each checklist item and gives methodological background and published examples of transparent reporting. The STROBE checklist is best used in conjunction with this article (freely available on the Web sites of PLoS Medicine at http://www.plosmedicine.org/, Annals of Internal Medicine at http://www.annals.org/, and Epidemiology at http://www.epidem.com/). Information on the STROBE Initiative is available at www.strobe-statement.org.

BMJ Open

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

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2020-042584.R2
Article Type:	Original research
Date Submitted by the Author:	09-Nov-2020
Complete List of Authors:	Unnikrishnan, Jayakrishnan ; QUALCOMM Inc Mangalathu, Sujith; Equifax Inc; Equifax Inc, Kutty, Raman; Sree Chitra Tirunal Institute for Medical Sciences and Technology
Primary Subject Heading:	Epidemiology
Secondary Subject Heading:	Epidemiology, Global health, Health policy, Infectious diseases
Keywords:	INFECTIOUS DISEASES, STATISTICS & RESEARCH METHODS, Public health < INFECTIOUS DISEASES

I, the Submitting Author has the right to grant and does grant on behalf of all authors of the Work (as defined in the below author licence), an exclusive licence and/or a non-exclusive licence for contributions from authors who are: i) UK Crown employees; ii) where BMJ has agreed a CC-BY licence shall apply, and/or iii) in accordance with the terms applicable for US Federal Government officers or employees acting as part of their official duties; on a worldwide, perpetual, irrevocable, royalty-free basis to BMJ Publishing Group Ltd ("BMJ") its licensees and where the relevant Journal is co-owned by BMJ to the co-owners of the Journal, to publish the Work in this journal and any other BMJ products and to exploit all rights, as set out in our licence.

The Submitting Author accepts and understands that any supply made under these terms is made by BMJ to the Submitting Author unless you are acting as an employee on behalf of your employer or a postgraduate student of an affiliated institution which is paying any applicable article publishing charge ("APC") for Open Access articles. Where the Submitting Author wishes to make the Work available on an Open Access basis (and intends to pay the relevant APC), the terms of reuse of such Open Access shall be governed by a Creative Commons licence – details of these licences and which Creative Commons licence will apply to this Work are set out in our licence referred to above.

Other than as permitted in any relevant BMJ Author's Self Archiving Policies, I confirm this Work has not been accepted for publication elsewhere, is not being considered for publication elsewhere and does not duplicate material already published. I confirm all authors consent to publication of this Work and authorise the granting of this licence.

RESEARCH

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*], and Raman V Kutty[3]

[*]Correspondence:
sujithmangalath@ucla.edu
[2]Equifax Inc, 1505 Windward Concourse, 30005 Alpharetta, USA.

[1]Qualcomm Inc., 500 Somerset Corporate Blvd, Bridgewater, NJ, USA

[3]Research Director, Amala Cancer Research Centre, 680555 Thrissur, India.

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 prioritizing testing in different states, and also to analyze the time-varying effects of different quarantine measures adopted in different states.

Keywords: Covid-19; Under-reporting; India

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 serologic 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 [1] suggests that the tests conducted up to October 10, 2020, in various states range from 29 to 182 per thousand residents. Although patients hospitalized 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 United States during the first half of March 2020 using travel data from epicenters [2]. Another study [3] uses a Bayesian analysis to get an estimate of the cumulative number of unreported cases in the United States up to April 18, 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 [1]. This data is 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 is 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 under-counts additional deaths that may arise from the cases observed to date. This issue can be addressed by using the distribution of delay from hospitalization 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 [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 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 [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 hospitalization to death estimated in [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 standard deviation 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 is 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 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 Generalize 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 one day 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 analyze 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 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 to 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% confidence 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 is available (10 October 2020), assuming a baseline CFR of 0.1%. As mentioned earlier, f_t 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 highest positivity rate also tend to have low estimates of the fraction of cases reported. The Spearman's rank correlation coefficient [9] between these two quantities is -0.4 with a p-value of 0.03 indicating a statistically significant negative relation. In the figure, we also show a regression line fit of $\log(y)$ vs x, which yields an r^2-value of 0.17 and a p-value of 0.04. Thus, an increase in test positivity rate is associated with a decrease in the fraction 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 serologic testing [10, 11] for Covid-19 antibodies among the general public. Randomized 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 under-counted, 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 statistically significant relationship with the test positivity rate.

The estimate of under-reporting may be used to guide policies for prioritizing 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.

Author Affiliations

[1]Qualcomm Inc., 500 Somerset Corporate Blvd, Bridgewater, NJ, USA
[2]Equifax Inc, 1505 Windward Concourse, 30005 Alpharetta, USA.
[3]Research Director, Amala Cancer Research Centre, 680555 Thrissur, India.

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, 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
The authors declare that they have no competing interests.

Patient and Public Involvement
Patients or the public were not involved in the design, or conduct, or reporting, or dissemination plans of our research.

Patient Consent for Publication
Not required

Ethics approval
Not required

Data availability statement
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].

References

1. COVID19-India API. https://api.covid19india.org
2. Hortaçsu, Ali et al. "Estimating the fraction of unreported infections in epidemics with a known epicenter: An application to COVID-19." Journal of econometrics, 10.1016/j.jeconom.2020.07.047. 7 Sep. 2020, doi:10.1016/j.jeconom.2020.07.047
3. Wu, S.L., Mertens, A.N., Crider, Y.S. et al. "Substantial underestimation of SARS-CoV-2 infection in the United States." Nat Commun 11, 4507 (2020).
4. Guan, W.-j., Ni, Z.-y., Hu, Y., Liang, W.-h., Ou, C.-q., He, J.-x., Liu, L., Shan, H., Lei, C.-l., Hui, D.S.C., Du, B., Li, L.-j., Zeng, G., Yuen, K.-Y., Chen, R.-c., Tang, C.-l., Wang, T., Chen, P.-y., Xiang, J., Li, S.-y., Wang, J.-l., Liang, Z.-j., Peng, Y.-x., Wei, L., Liu, Y., Hu, Y.-h., Peng, P., Wang, J.-m., Liu, J.-y., Chen, Z., Li, G., Zheng, Z.-j., Qiu, S.-q., Luo, J., Ye, C.-j., Zhu, S.-y., Zhong, N.-s.: Brca1clinical characteristics of coronavirus disease 2019 in china. New England Journal of Medicine 382(18), 1708{1720 (2020)
5. Verity, R., Okell, L.C., Dorigatti, I., Winskill, P., Whittaker, C., Imai, N., Cuomo-Dannenburg, G., Thompson, H., Walker, P.G.T., Fu, H., Dighe, A., Gri n, J.T., Baguelin, M., Bhatia, S., Boonyasiri, A., Cori, A., Cucunuba, Z., FitzJohn, R., Gaythorpe, K., Green, W., Hamlet, A., Hinsley, W., Laydon, D., Nedjati-Gilani, G., Riley, S., van Elsland, S., Volz, E., Wang, H., Wang, Y., Xi, X., Donnelly, C.A., Ghani, A.C., Ferguson, N.M.: Estimates of the severity of coronavirus disease 2019: a model-based analysis. The Lancet Infectious Diseases 20(6), 669{677 (2020)
6. Ioannidis, J., The infection fatality rate of COVID-19 inferred from seroprevalence data. medRxiv. doi: https://doi.org/10.1101/2020.05.13.20101253 July 14, 2020.
7. Russell, T. W., Hellewell, J., Abbott, S., Jarvis, C. I., van Zandvoort, K., CMMID nCov working group, ... & Kucharski, A. J. (2020). Using a delay-adjusted case fatality ratio to estimate under-reporting. Centre for Mathematical Modeling of Infectious Diseases Repository
8. Linton, N.M., Kobayashi, T., Yang, Y., Hayashi, K., Akhmetzhanov, A.R., Jung, S.-m., Yuan, B., Kinoshita, R., Nishiura, H.: Incubation period and other epidemiological characteristics of 2019 novel coronavirus infections with right truncation: A statistical analysis of publicly available case data. Journal of Clinical Medicine 9(2), 538 (2020)
9. Spearman, C. "The Proof and Measurement of Association between Two Things." The American Journal of Psychology, vol. 15, no. 1, 1904, pp. 72–101. JSTOR, www.jstor.org/stable/1412159. Accessed 11 Oct. 2020.
10. Long, Q.-X., Liu, B.-Z., Deng, H.-J., Wu, G.-C., Deng, K., Chen, Y.-K., Liao, P., Qiu, J.-F., Lin, Y., Cai, X.-F., Wang, D.-Q., Hu, Y., Ren, J.-H., Tang, N., Xu, Y.-Y., Yu, L.-H., Mo, Z., Gong, F., Zhang, X.-L., Tian, W.-G., Hu, L., Zhang, X.-X., Xiang, J.-L., Du, H.-X., Liu, H.-W., Lang, C.-H., Luo, X.-H., Wu, S.-B., Cui, X.-P., Zhou, Z., Zhu, M.-M., Wang, J., Xue, C.-J., Li, X.-F., Wang, L., Li, Z.-J., Wang, K., Niu, C.-C., Yang, Q.-J., Tang, X.-J., Zhang, Y., Liu, X.-M., Li, J.-J., Zhang, D.-C., Zhang, F., Liu, P., Yuan, J., Li, Q., Hu, J.-L., Chen, J., Huang, A.-L.: Antibody responses to SARS-CoV-2 in patients with COVID-19. Nature Medicine, 2020, 1-4
11. Whitman, J.D., Hiatt, J., Mowery, C.T., Shy, B.R., Yu, R., Yamamoto, T.N., Rathore, U., Goldgof, G.M., Whitty, C., Woo, J.M., Gallman, A.E., Miller, T.E., Levine, A.G., Nguyen, D.N., Bapat, S.P., Balcerek, J., Bylsma, S.A., Lyons, A.M., Li, S., Wong, A.W.-Y., Gillis-Buck, E.M., Steinhart, Z.B., Lee, Y., Apathy, R., Lipke, M.J., Smith, J.A., Zheng, T., Boothby, I.C., Isaza, E.,

Chan, J., Acenas, n. Dante D, Lee, J., Macrae, T.A., Kyaw, T.S., Wu, D., Ng, D.L., Gu, W., York, V.A., Eskandarian, H.A., Callaway, P.C., Warrier, L., Moreno, M.E., Levan, J., Torres, L., Farrington, L.A., Loudermilk, R., Koshal, K., Zorn, K.C., Garcia-Beltran, W.F., Yang, D., Astudillo, M.G., Bernstein, B.E., Gelfand, J.A., Ryan, E.T., Charles, R.C., Iafrate, A.J., Lennerz, J.K., Miller, S., Chiu, C.Y., Stramer, S.L., Wilson, M.R., Manglik, A., Ye, C.J., Krogan, N.J., Anderson, M.S., Cyster, J.G., Ernst, J.D., Wu, A.H.B., Lynch, K.L., Bern, C., Hsu, P.D., Marson, A.: Test performance evaluation of SARS-CoV-2 serological assays. medRxiv, 2020

Figures

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.

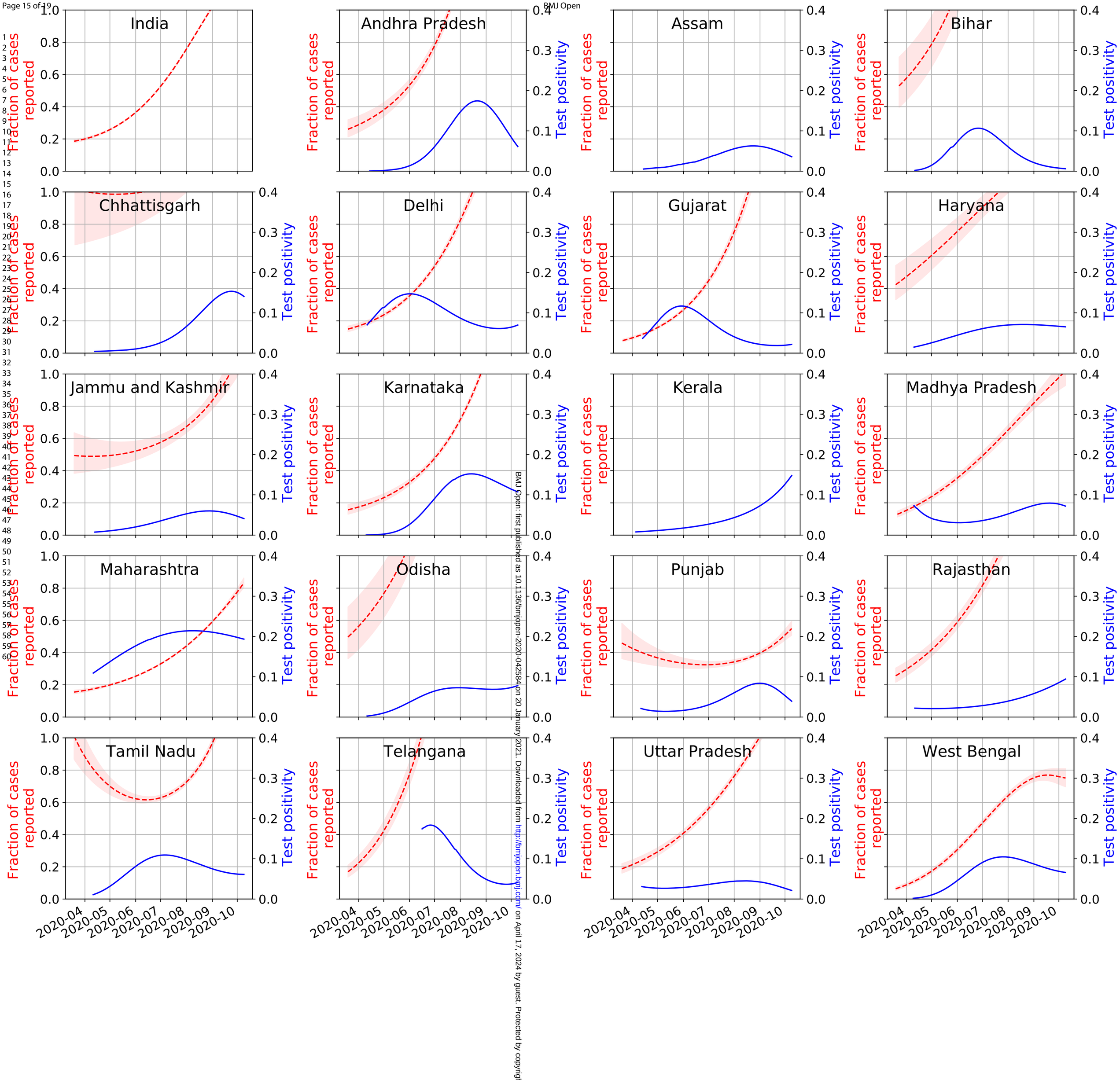
Figure 2. Curves in red show the estimates of the fraction of cases reported in various regions as a function of time, assuming a baseline CFR of 0.1%. The first subplot shows the results for India and the other subplots show results for the top 19 states with most reported deaths. Curves in blue show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

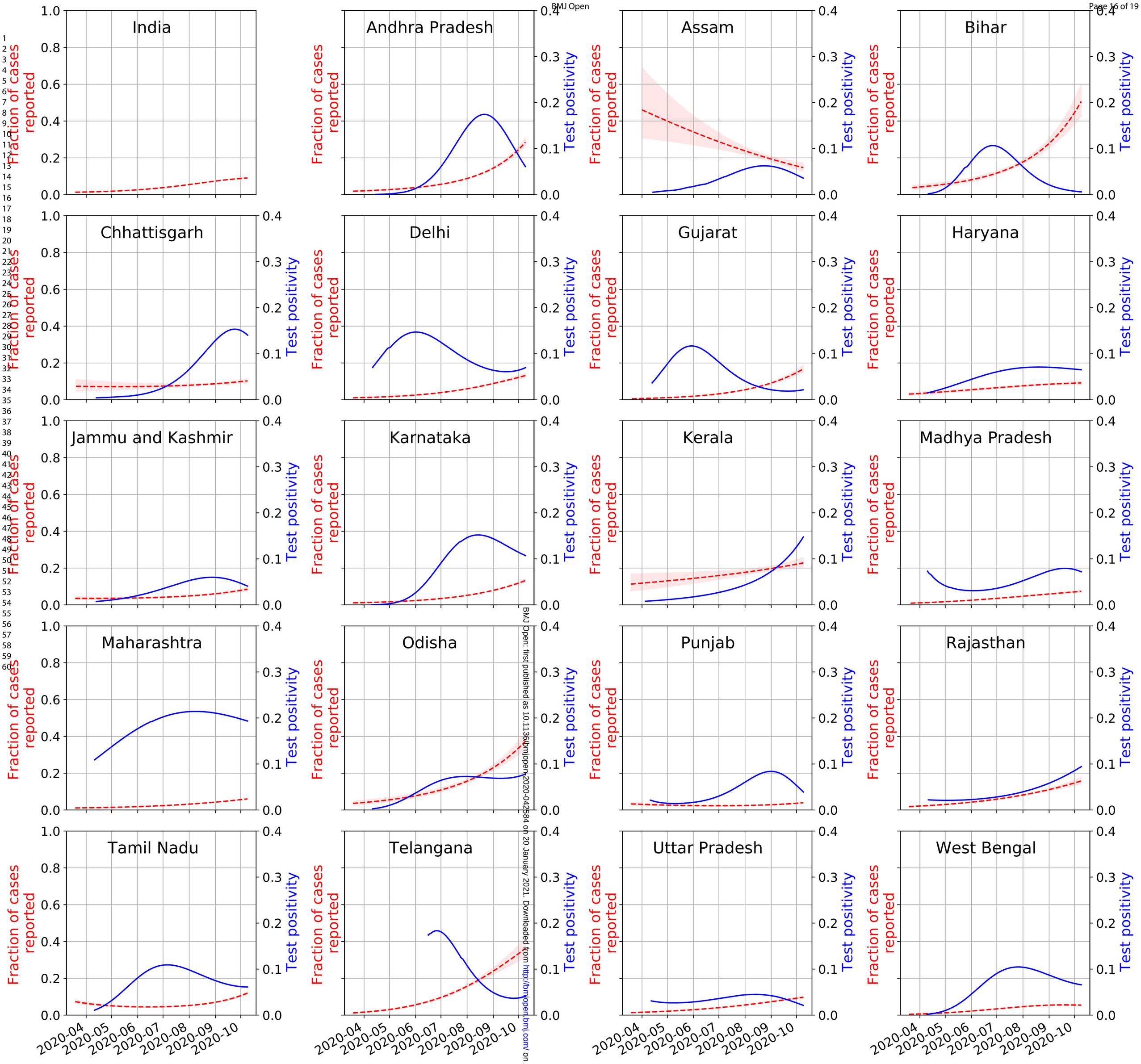
Figure 3. Curves in blue shows the test positivity rate estimated via the Poisson regression method. Curves in green show the ratio of cumulative positive cases to cumulative tests performed.

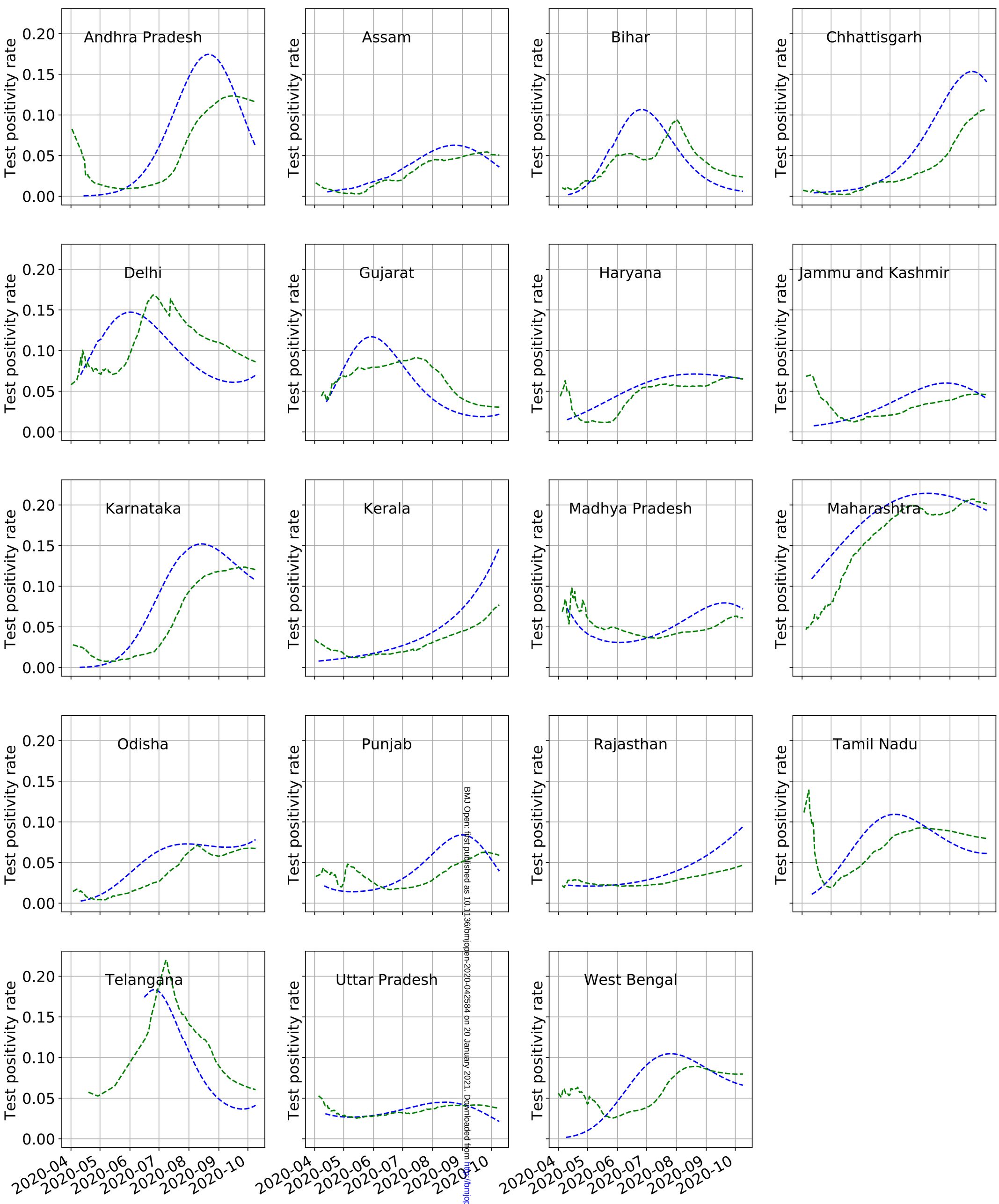
Figure 4. Scatter plot of the estimate of the fraction f_t of cases reported from different states evaluated on the last date considered, against the corresponding test positivity rate

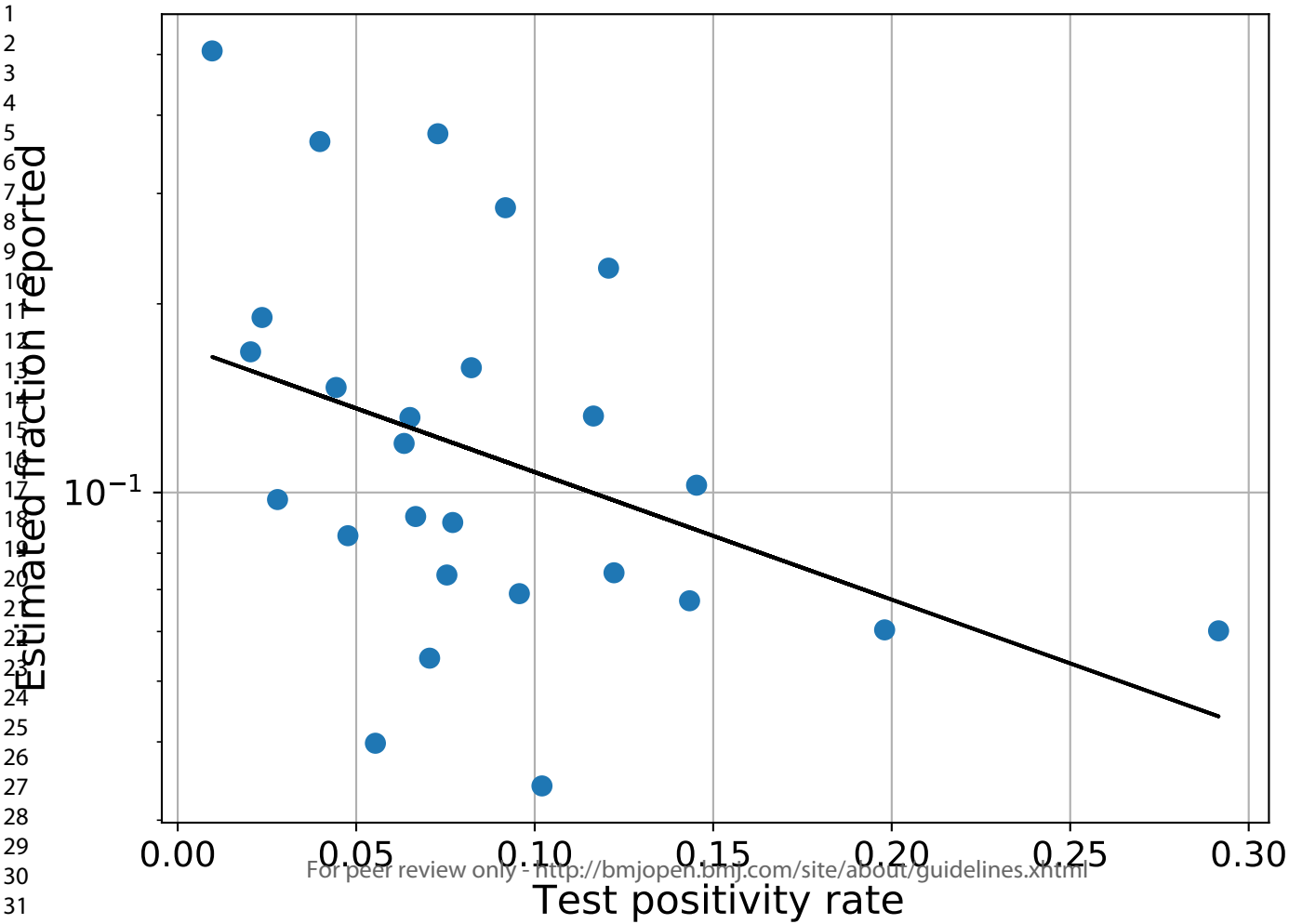
Table 1. Estimates of fraction of cases reported in different states

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	106863	6976461	-	1.53	1.78	77.62	37.12	5.62
Andhra Pradesh	6159	744864	6.1	0.83	0.93	100.00	71.07	10.77
Assam	807	192314	3.6	0.42	0.47	100.00	100.00	21.11
Bihar	934	193826	0.6	0.48	0.53	100.00	100.00	18.92
Chhattisgarh	1196	137570	14.1	0.87	1.14	100.00	57.86	8.77
Delhi	5692	303693	7.0	1.87	2.13	64.85	31.01	4.70
Gujarat	3549	149193	2.2	2.38	2.68	51.59	24.67	3.74
Haryana	1562	139932	6.5	1.12	1.29	100.00	51.13	7.75
Jammu and Kashmir	1306	82429	4.1	1.58	1.84	74.84	35.79	5.42
Karnataka	9200	690269	10.7	1.33	1.60	86.35	41.30	6.26
Kerala	956	268101	14.8	0.36	0.51	100.00	100.00	19.53
Madhya Pradesh	2575	143629	7.2	1.79	2.14	64.57	30.88	4.68
Maharashtra	39731	1506018	19.3	2.64	3.02	45.67	21.84	3.31
Odisha	1044	246839	7.8	0.42	0.51	100.00	100.00	19.70
Punjab	3774	122462	3.9	3.08	3.55	38.88	18.59	2.82
Rajasthan	1621	154785	9.4	1.05	1.25	100.00	52.81	8.00
Tamil Nadu	10120	646128	6.1	1.57	1.75	78.80	37.69	5.71
Telangana	1208	208025	4.1	0.58	0.66	100.00	100.00	15.18
Uttar Pradesh	6293	430666	2.1	1.46	1.66	83.16	39.77	6.03
West Bengal	5501	287603	6.6	1.91	2.23	61.89	29.60	4.49









STROBE Statement—Checklist of items that should be included in reports of ***cross-sectional studies***

	Item No	Page No.	Recommendation
Title and abstract	1	1	(*a*) Indicate the study's design with a commonly used term in the title or the abstract
		1	(*b*) Provide in the abstract an informative and balanced summary of what was done and what was found
			Introduction
Background/rationale	2	2	Explain the scientific background and rationale for the investigation being reported
Objectives	3	2	State specific objectives, including any prespecified hypotheses
			Methods
Study design	4	2,3	Present key elements of study design early in the paper
Setting	5	2,3	Describe the setting, locations, and relevant dates, including periods of recruitment, exposure, follow-up, and data collection
Participants	6	3	(*a*) Give the eligibility criteria, and the sources and methods of selection of participants
Variables	7	3	Clearly define all outcomes, exposures, predictors, potential confounders, and effect modifiers. Give diagnostic criteria, if applicable
Data sources/ measurement	8*	*3*	For each variable of interest, give sources of data and details of methods of assessment (measurement). Describe comparability of assessment methods if there is more than one group
Bias	9	3,4	Describe any efforts to address potential sources of bias
Study size	10	3,4	Explain how the study size was arrived at
Quantitative variables	11	3.4	Explain how quantitative variables were handled in the analyses. If applicable, describe which groupings were chosen and why
Statistical methods	12	4	(*a*) Describe all statistical methods, including those used to control for confounding
		4	(*b*) Describe any methods used to examine subgroups and interactions
		NA	(*c*) Explain how missing data were addressed
		NA	(*d*) If applicable, describe analytical methods taking account of sampling strategy
		4,5	(*e*) Describe any sensitivity analyses
			Results
Participants	13*	6,7	(a) Report numbers of individuals at each stage of study—eg numbers potentially eligible, examined for eligibility, confirmed eligible, included in the study, completing follow-up, and analysed
		NA	(b) Give reasons for non-participation at each stage
		NA	(c) Consider use of a flow diagram
Descriptive data	14*	6,7	(a) Give characteristics of study participants (eg demographic, clinical, social) and information on exposures and potential confounders
		NA	(b) Indicate number of participants with missing data for each variable of interest

Outcome data	15*	NA	Report numbers of outcome events or summary measures
Main results	16	6,7	(*a*) Give unadjusted estimates and, if applicable, confounder-adjusted estimates and their precision (eg, 95% confidence interval). Make clear which confounders were adjusted for and why they were included
			(*b*) Report category boundaries when continuous variables were categorized
			(*c*) If relevant, consider translating estimates of relative risk into absolute risk for a meaningful time period
Other analyses	17	7	Report other analyses done—eg analyses of subgroups and interactions, and sensitivity analyses
Discussion			
Key results	18	7,8	Summarise key results with reference to study objectives
Limitations	19	8	Discuss limitations of the study, taking into account sources of potential bias or imprecision. Discuss both direction and magnitude of any potential bias
Interpretation	20	8	Give a cautious overall interpretation of results considering objectives, limitations, multiplicity of analyses, results from similar studies, and other relevant evidence
Generalisability	21	8,9	Discuss the generalisability (external validity) of the study results
Other information			
Funding	22	10	Give the source of funding and the role of the funders for the present study and, if applicable, for the original study on which the present article is based

*Give information separately for exposed and unexposed groups.

Note: An Explanation and Elaboration article discusses each checklist item and gives methodological background and published examples of transparent reporting. The STROBE checklist is best used in conjunction with this article (freely available on the Web sites of PLoS Medicine at http://www.plosmedicine.org/, Annals of Internal Medicine at http://www.annals.org/, and Epidemiology at http://www.epidem.com/). Information on the STROBE Initiative is available at www.strobe-statement.org.

BMJ Open

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

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2020-042584.R3
Article Type:	Original research
Date Submitted by the Author:	20-Dec-2020
Complete List of Authors:	Unnikrishnan, Jayakrishnan ; QUALCOMM Inc Mangalathu, Sujith; Equifax Inc; Equifax Inc, Kutty, Raman; Sree Chitra Tirunal Institute for Medical Sciences and Technology
Primary Subject Heading:	Epidemiology
Secondary Subject Heading:	Epidemiology, Global health, Health policy, Infectious diseases
Keywords:	INFECTIOUS DISEASES, STATISTICS & RESEARCH METHODS, Public health < INFECTIOUS DISEASES

I, the Submitting Author has the right to grant and does grant on behalf of all authors of the Work (as defined in the below author licence), an exclusive licence and/or a non-exclusive licence for contributions from authors who are: i) UK Crown employees; ii) where BMJ has agreed a CC-BY licence shall apply, and/or iii) in accordance with the terms applicable for US Federal Government officers or employees acting as part of their official duties; on a worldwide, perpetual, irrevocable, royalty-free basis to BMJ Publishing Group Ltd ("BMJ") its licensees and where the relevant Journal is co-owned by BMJ to the co-owners of the Journal, to publish the Work in this journal and any other BMJ products and to exploit all rights, as set out in our licence.

The Submitting Author accepts and understands that any supply made under these terms is made by BMJ to the Submitting Author unless you are acting as an employee on behalf of your employer or a postgraduate student of an affiliated institution which is paying any applicable article publishing charge ("APC") for Open Access articles. Where the Submitting Author wishes to make the Work available on an Open Access basis (and intends to pay the relevant APC), the terms of reuse of such Open Access shall be governed by a Creative Commons licence – details of these licences and which Creative Commons licence will apply to this Work are set out in our licence referred to above.

Other than as permitted in any relevant BMJ Author's Self Archiving Policies, I confirm this Work has not been accepted for publication elsewhere, is not being considered for publication elsewhere and does not duplicate material already published. I confirm all authors consent to publication of this Work and authorise the granting of this licence.

RESEARCH

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*], and Raman V Kutty[3]

[*]Correspondence:
sujithmangalath@ucla.edu
[2]Equifax Inc, 1505 Windward Concourse, 30005 Alpharetta, USA.

[1]Qualcomm Inc., 500 Somerset Corporate Blvd, Bridgewater, NJ, USA

[3]Research Director, Amala Cancer Research Centre, 680555 Thrissur, India.

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 prioritizing testing in different states, and also to analyze the time-varying effects of different quarantine measures adopted in different states.

Keywords: Covid-19; Under-reporting; India

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 serologic 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 [1] suggests that the tests conducted up to October 10, 2020, in various states range from 29 to 182 per thousand residents. Although patients hospitalized 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 United States during the first half of March 2020 using travel data from epicenters [2]. Another study [3] uses a Bayesian analysis to get an estimate of the cumulative number of unreported cases in the United States up to April 18, 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 [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 under-counts additional deaths that may arise from the cases observed to date. This issue can be addressed by using the distribution of delay from hospitalization 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 [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 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 [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 hospitalization to death estimated in [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 standard deviation 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 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 Generalize 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 one day 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 analyze 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 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 to 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% confidence 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 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 serologic testing [10, 11] for Covid-19 antibodies among the general public. Randomized 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 under-counted, 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 prioritizing 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.

Author Affiliations

[1]Qualcomm Inc., 500 Somerset Corporate Blvd, Bridgewater, NJ, USA
[2]Equifax Inc, 1505 Windward Concourse, 30005 Alpharetta, USA.
[3]Research Director, Amala Cancer Research Centre, 680555 Thrissur, India.

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, 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
The authors declare that they have no competing interests.

Patient and Public Involvement
Patients or the public were not involved in the design, or conduct, or reporting, or dissemination plans of our research.

Patient Consent for Publication
Not required

Ethics approval
Not required

Data availability statement
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].

References

1. COVID19-India API. https://api.covid19india.org
2. Hortaçsu, Ali et al. "Estimating the fraction of unreported infections in epidemics with a known epicenter: An application to COVID-19." Journal of econometrics, 10.1016/j.jeconom.2020.07.047. 7 Sep. 2020, doi:10.1016/j.jeconom.2020.07.047
3. Wu, S.L., Mertens, A.N., Crider, Y.S. et al. "Substantial underestimation of SARS-CoV-2 infection in the United States." Nat Commun 11, 4507 (2020).
4. Guan, W.-j., Ni, Z.-y., Hu, Y., Liang, W.-h., Ou, C.-q., He, J.-x., Liu, L., Shan, H., Lei, C.-l., Hui, D.S.C., Du, B., Li, L.-j., Zeng, G., Yuen, K.-Y., Chen, R.-c., Tang, C.-l., Wang, T., Chen, P.-y., Xiang, J., Li, S.-y., Wang, J.-l., Liang, Z.-j., Peng, Y.-x., Wei, L., Liu, Y., Hu, Y.-h., Peng, P., Wang, J.-m., Liu, J.-y., Chen, Z., Li, G., Zheng, Z.-j., Qiu, S.-q., Luo, J., Ye, C.-j., Zhu, S.-y., Zhong, N.-s.: Brca1clinical characteristics of coronavirus disease 2019 in china. New England Journal of Medicine 382(18), 1708{1720 (2020)
5. Verity, R., Okell, L.C., Dorigatti, I., Winskill, P., Whittaker, C., Imai, N., Cuomo-Dannenburg, G., Thompson, H., Walker, P.G.T., Fu, H., Dighe, A., Gri n, J.T., Baguelin, M., Bhatia, S., Boonyasiri, A., Cori, A., Cucunuba, Z., FitzJohn, R., Gaythorpe, K., Green, W., Hamlet, A., Hinsley, W., Laydon, D., Nedjati-Gilani, G., Riley, S., van Elsland, S., Volz, E., Wang, H., Wang, Y., Xi, X., Donnelly, C.A., Ghani, A.C., Ferguson, N.M.: Estimates of the severity of coronavirus disease 2019: a model-based analysis. The Lancet Infectious Diseases 20(6), 669{677 (2020)
6. Ioannidis, J., The infection fatality rate of COVID-19 inferred from seroprevalence data. medRxiv. doi: https://doi.org/10.1101/2020.05.13.20101253 July 14, 2020.
7. Russell, T. W., Hellewell, J., Abbott, S., Jarvis, C. I., van Zandvoort, K., CMMID nCov working group, ... & Kucharski, A. J. (2020). Using a delay-adjusted case fatality ratio to estimate under-reporting. Centre for Mathematical Modeling of Infectious Diseases Repository
8. Linton, N.M., Kobayashi, T., Yang, Y., Hayashi, K., Akhmetzhanov, A.R., Jung, S.-m., Yuan, B., Kinoshita, R., Nishiura, H.: Incubation period and other epidemiological characteristics of 2019 novel coronavirus infections with right truncation: A statistical analysis of publicly available case data. Journal of Clinical Medicine 9(2), 538 (2020)
9. Spearman, C. "The Proof and Measurement of Association between Two Things." The American Journal of Psychology, vol. 15, no. 1, 1904, pp. 72–101. JSTOR, www.jstor.org/stable/1412159. Accessed 11 Oct. 2020.
10. Long, Q.-X., Liu, B.-Z., Deng, H.-J., Wu, G.-C., Deng, K., Chen, Y.-K., Liao, P., Qiu, J.-F., Lin, Y., Cai, X.-F., Wang, D.-Q., Hu, Y., Ren, J.-H., Tang, N., Xu, Y.-Y., Yu, L.-H., Mo, Z., Gong, F., Zhang, X.-L., Tian, W.-G., Hu, L., Zhang, X.-X., Xiang, J.-L., Du, H.-X., Liu, H.-W., Lang, C.-H., Luo, X.-H., Wu, S.-B., Cui, X.-P., Zhou, Z., Zhu, M.-M., Wang, J., Xue, C.-J., Li, X.-F., Wang, L., Li, Z.-J., Wang, K., Niu, C.-C., Yang, Q.-J., Tang, X.-J., Zhang, Y., Liu, X.-M., Li, J.-J., Zhang, D.-C., Zhang, F., Liu, P., Yuan, J., Li, Q., Hu, J.-L., Chen, J., Huang, A.-L.: Antibody responses to SARS-CoV-2 in patients with COVID-19. Nature Medicine, 2020, 1-4
11. Whitman, J.D., Hiatt, J., Mowery, C.T., Shy, B.R., Yu, R., Yamamoto, T.N., Rathore, U., Goldgof, G.M., Whitty, C., Woo, J.M., Gallman, A.E., Miller, T.E., Levine, A.G., Nguyen, D.N., Bapat, S.P., Balcerek, J., Bylsma, S.A., Lyons, A.M., Li, S., Wong, A.W.-Y., Gillis-Buck, E.M., Steinhart, Z.B., Lee, Y., Apathy, R., Lipke, M.J., Smith, J.A., Zheng, T., Boothby, I.C., Isaza, E.,

Chan, J., Acenas, n. Dante D, Lee, J., Macrae, T.A., Kyaw, T.S., Wu, D., Ng, D.L., Gu, W., York, V.A., Eskandarian, H.A., Callaway, P.C., Warrier, L., Moreno, M.E., Levan, J., Torres, L., Farrington, L.A., Loudermilk, R., Koshal, K., Zorn, K.C., Garcia-Beltran, W.F., Yang, D., Astudillo, M.G., Bernstein, B.E., Gelfand, J.A., Ryan, E.T., Charles, R.C., Iafrate, A.J., Lennerz, J.K., Miller, S., Chiu, C.Y., Stramer, S.L., Wilson, M.R., Manglik, A., Ye, C.J., Krogan, N.J., Anderson, M.S., Cyster, J.G., Ernst, J.D., Wu, A.H.B., Lynch, K.L., Bern, C., Hsu, P.D., Marson, A.: Test performance evaluation of SARS-CoV-2 serological assays. medRxiv, 2020

Figures

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.

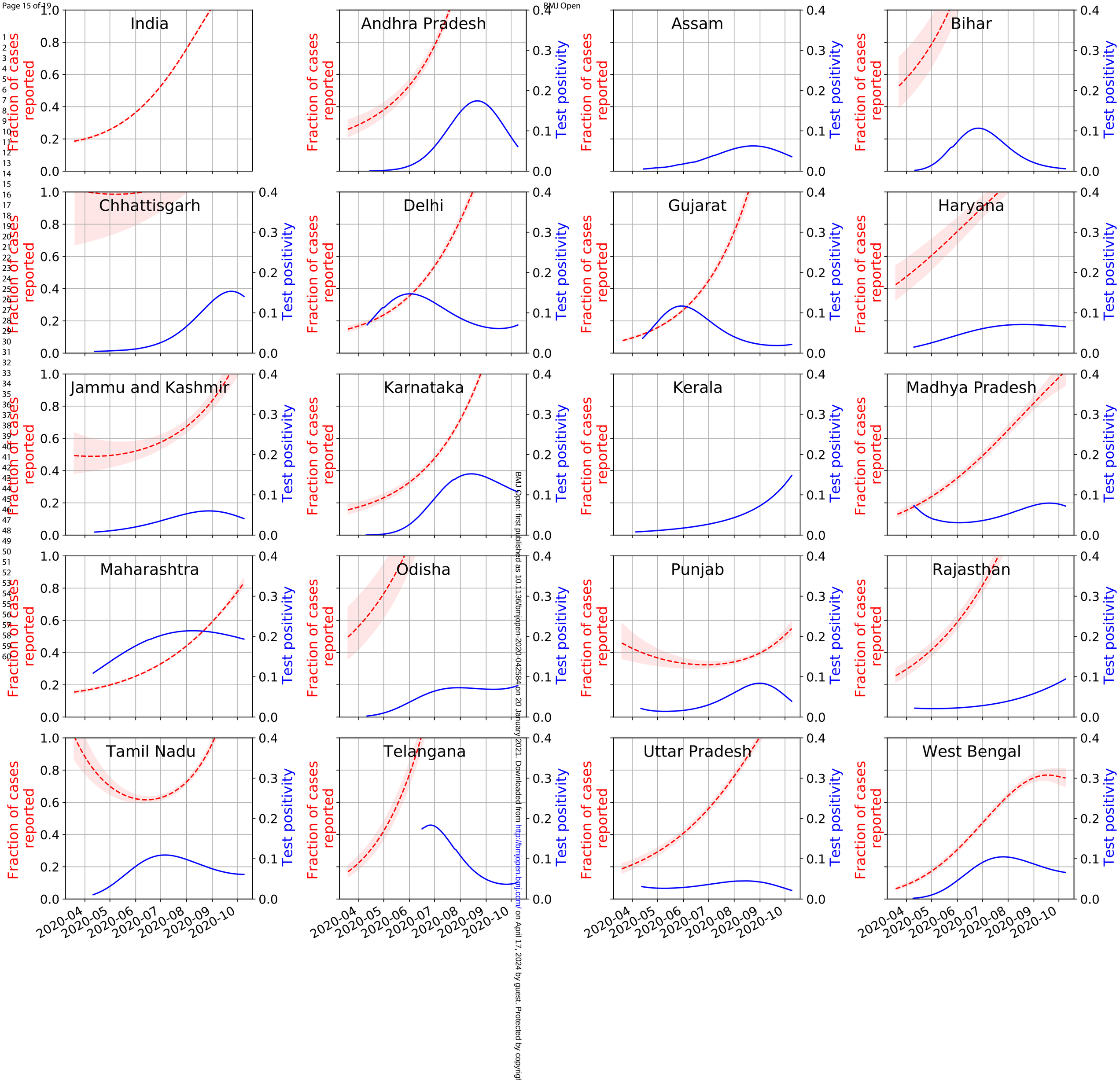
Figure 2. Curves in red show the estimates of the fraction of cases reported in various regions as a function of time, assuming a baseline CFR of 0.1%. The first subplot shows the results for India and the other subplots show results for the top 19 states with most reported deaths. Curves in blue show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

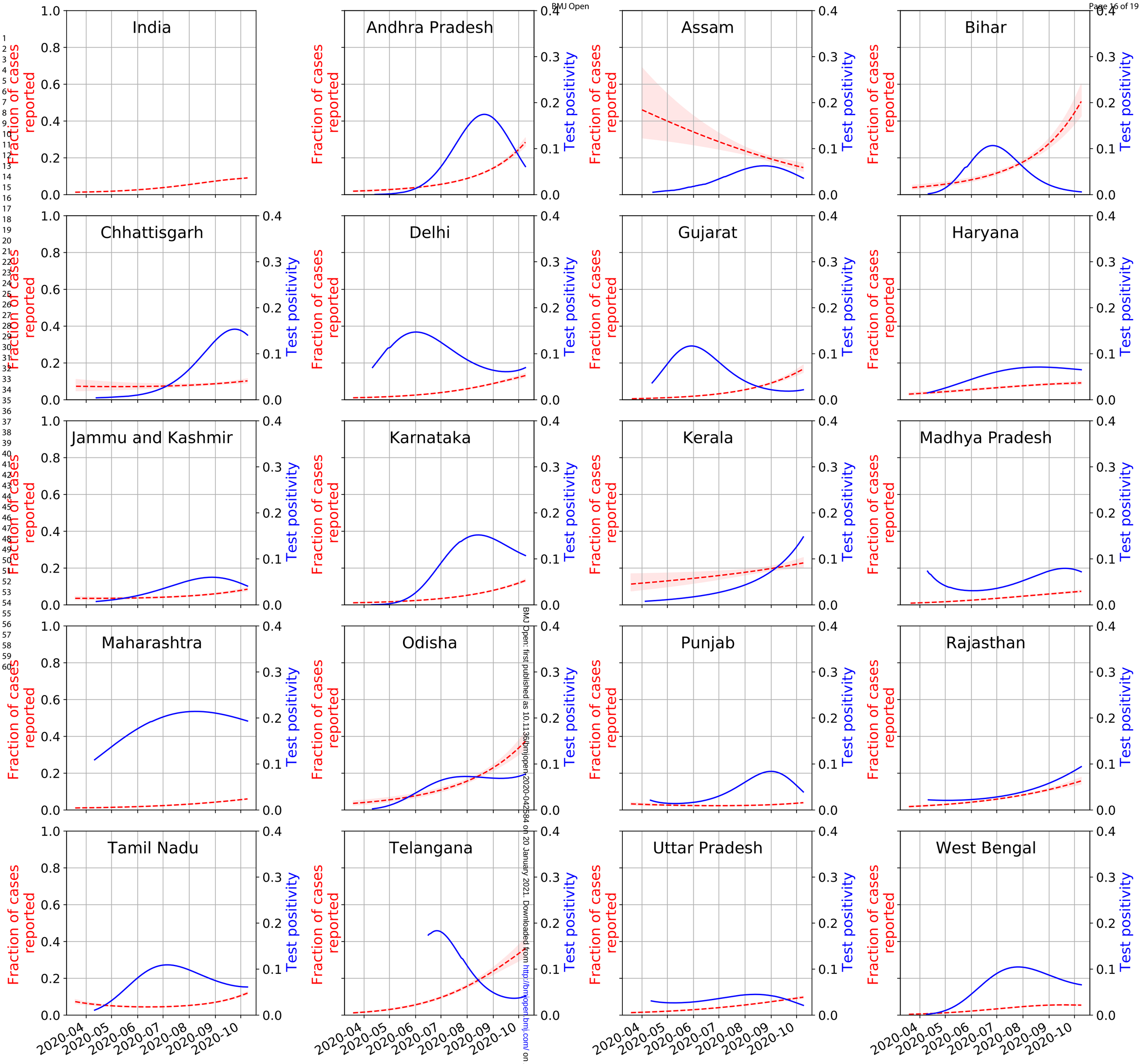
Figure 3. Curves in blue shows the test positivity rate estimated via the Poisson regression method. Curves in green show the ratio of cumulative positive cases to cumulative tests performed.

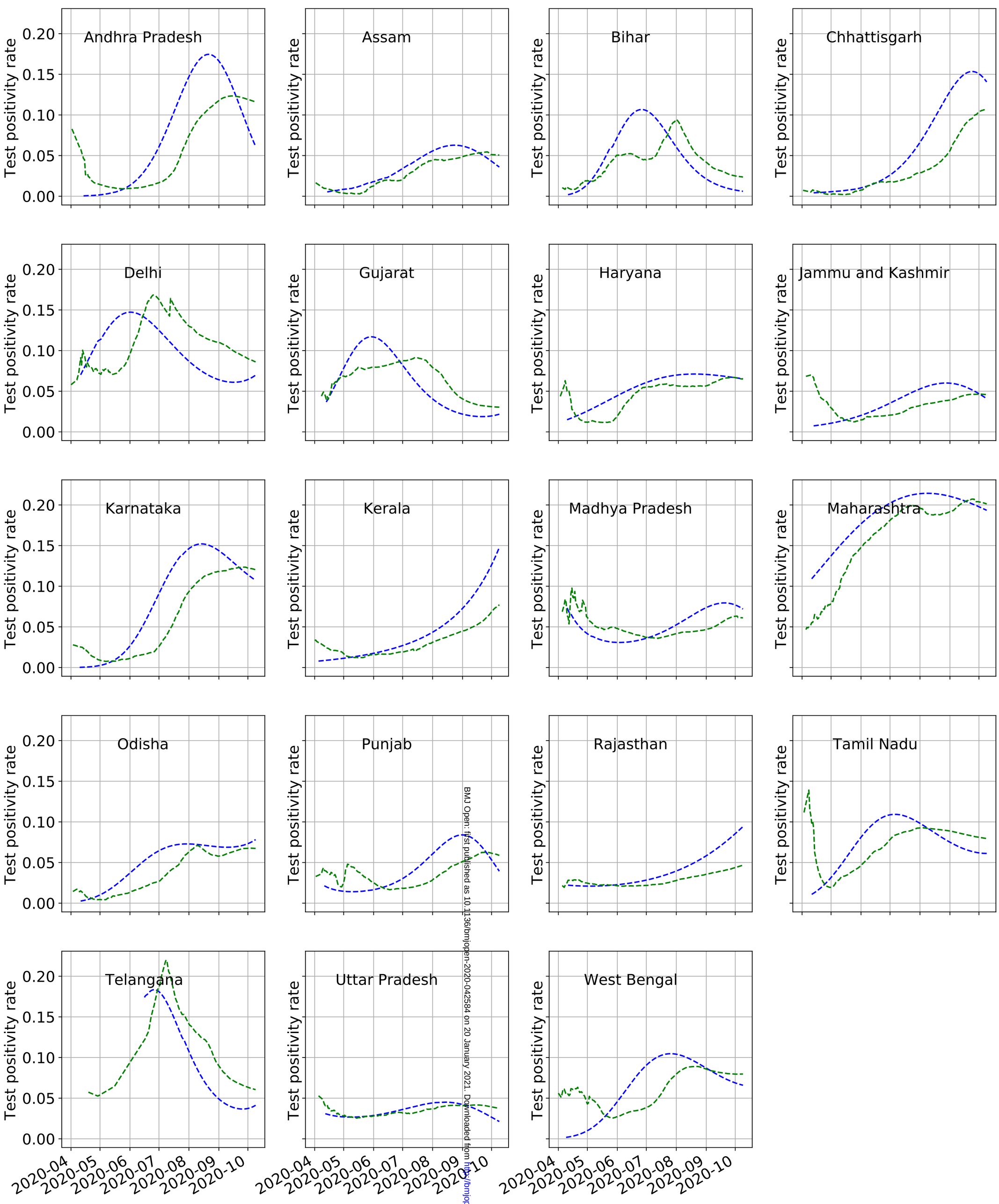
Figure 4. Scatter plot of the estimate of the fraction f_t of cases reported from different states evaluated on the last date considered, against the corresponding test positivity rate

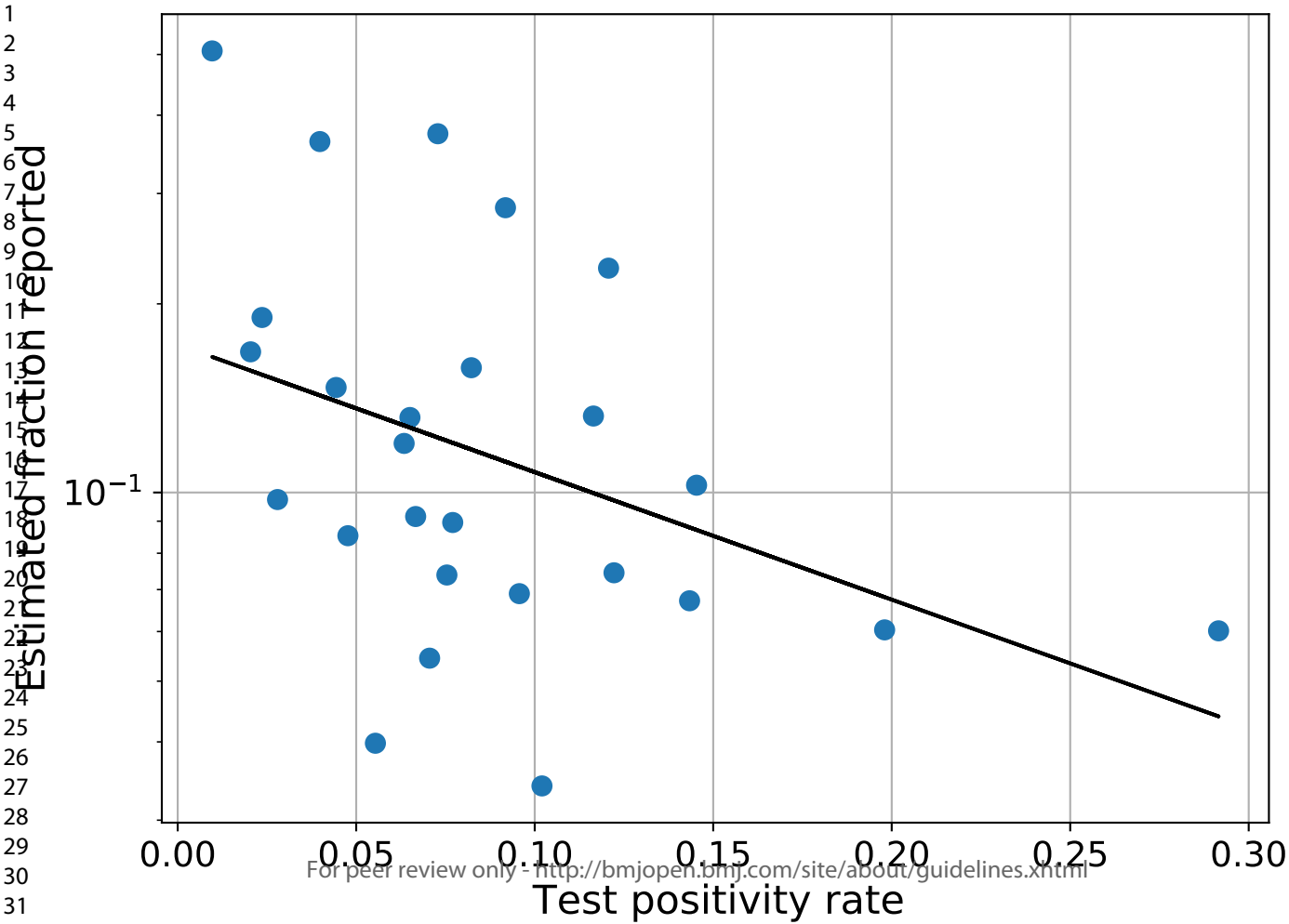
Table 1. Estimates of fraction of cases reported in different states

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	106863	6976461	-	1.53	1.78	77.62	37.12	5.62
Andhra Pradesh	6159	744864	6.1	0.83	0.93	100.00	71.07	10.77
Assam	807	192314	3.6	0.42	0.47	100.00	100.00	21.11
Bihar	934	193826	0.6	0.48	0.53	100.00	100.00	18.92
Chhattisgarh	1196	137570	14.1	0.87	1.14	100.00	57.86	8.77
Delhi	5692	303693	7.0	1.87	2.13	64.85	31.01	4.70
Gujarat	3549	149193	2.2	2.38	2.68	51.59	24.67	3.74
Haryana	1562	139932	6.5	1.12	1.29	100.00	51.13	7.75
Jammu and Kashmir	1306	82429	4.1	1.58	1.84	74.84	35.79	5.42
Karnataka	9200	690269	10.7	1.33	1.60	86.35	41.30	6.26
Kerala	956	268101	14.8	0.36	0.51	100.00	100.00	19.53
Madhya Pradesh	2575	143629	7.2	1.79	2.14	64.57	30.88	4.68
Maharashtra	39731	1506018	19.3	2.64	3.02	45.67	21.84	3.31
Odisha	1044	246839	7.8	0.42	0.51	100.00	100.00	19.70
Punjab	3774	122462	3.9	3.08	3.55	38.88	18.59	2.82
Rajasthan	1621	154785	9.4	1.05	1.25	100.00	52.81	8.00
Tamil Nadu	10120	646128	6.1	1.57	1.75	78.80	37.69	5.71
Telangana	1208	208025	4.1	0.58	0.66	100.00	100.00	15.18
Uttar Pradesh	6293	430666	2.1	1.46	1.66	83.16	39.77	6.03
West Bengal	5501	287603	6.6	1.91	2.23	61.89	29.60	4.49









STROBE Statement—Checklist of items that should be included in reports of ***cross-sectional studies***

	Item No	Page No.	Recommendation
Title and abstract	1	1	(*a*) Indicate the study's design with a commonly used term in the title or the abstract
		1	(*b*) Provide in the abstract an informative and balanced summary of what was done and what was found
			Introduction
Background/rationale	2	2	Explain the scientific background and rationale for the investigation being reported
Objectives	3	2	State specific objectives, including any prespecified hypotheses
			Methods
Study design	4	2,3	Present key elements of study design early in the paper
Setting	5	2,3	Describe the setting, locations, and relevant dates, including periods of recruitment, exposure, follow-up, and data collection
Participants	6	3	(*a*) Give the eligibility criteria, and the sources and methods of selection of participants
Variables	7	3	Clearly define all outcomes, exposures, predictors, potential confounders, and effect modifiers. Give diagnostic criteria, if applicable
Data sources/ measurement	8*	*3*	For each variable of interest, give sources of data and details of methods of assessment (measurement). Describe comparability of assessment methods if there is more than one group
Bias	9	3,4	Describe any efforts to address potential sources of bias
Study size	10	3,4	Explain how the study size was arrived at
Quantitative variables	11	3.4	Explain how quantitative variables were handled in the analyses. If applicable, describe which groupings were chosen and why
Statistical methods	12	4	(*a*) Describe all statistical methods, including those used to control for confounding
		4	(*b*) Describe any methods used to examine subgroups and interactions
		NA	(*c*) Explain how missing data were addressed
		NA	(*d*) If applicable, describe analytical methods taking account of sampling strategy
		4,5	(*e*) Describe any sensitivity analyses
			Results
Participants	13*	6,7	(a) Report numbers of individuals at each stage of study—eg numbers potentially eligible, examined for eligibility, confirmed eligible, included in the study, completing follow-up, and analysed
		NA	(b) Give reasons for non-participation at each stage
		NA	(c) Consider use of a flow diagram
Descriptive data	14*	6,7	(a) Give characteristics of study participants (eg demographic, clinical, social) and information on exposures and potential confounders
		NA	(b) Indicate number of participants with missing data for each variable of interest

Outcome data	15*	NA	Report numbers of outcome events or summary measures
Main results	16	6,7	(*a*) Give unadjusted estimates and, if applicable, confounder-adjusted estimates and their precision (eg, 95% confidence interval). Make clear which confounders were adjusted for and why they were included
			(*b*) Report category boundaries when continuous variables were categorized
			(*c*) If relevant, consider translating estimates of relative risk into absolute risk for a meaningful time period
Other analyses	17	7	Report other analyses done—eg analyses of subgroups and interactions, and sensitivity analyses
Discussion			
Key results	18	7,8	Summarise key results with reference to study objectives
Limitations	19	8	Discuss limitations of the study, taking into account sources of potential bias or imprecision. Discuss both direction and magnitude of any potential bias
Interpretation	20	8	Give a cautious overall interpretation of results considering objectives, limitations, multiplicity of analyses, results from similar studies, and other relevant evidence
Generalisability	21	8,9	Discuss the generalisability (external validity) of the study results
Other information			
Funding	22	10	Give the source of funding and the role of the funders for the present study and, if applicable, for the original study on which the present article is based

*Give information separately for exposed and unexposed groups.

Note: An Explanation and Elaboration article discusses each checklist item and gives methodological background and published examples of transparent reporting. The STROBE checklist is best used in conjunction with this article (freely available on the Web sites of PLoS Medicine at http://www.plosmedicine.org/, Annals of Internal Medicine at http://www.annals.org/, and Epidemiology at http://www.epidem.com/). Information on the STROBE Initiative is available at www.strobe-statement.org.