Supplementary material for

Demography, sanitation and previous disease prevalence associate with Covid-19 deaths across Indian states

Bithika Chatterjee[1] and Shekhar C. Mande[1,2*]

[1]National Centre for Cell Science, NCCS Complex, Ganeshkhind, Pune- 411007

[2]Bioinformatics Centre, Savitribai Phule Pune University, Ganeshkhind, Pune- 411007

* Correspondence:
Shekhar C. Mande

Bioinformatics Centre, Savitribai Phule Pune University, Pune and

National Centre for Cell Science, NCCS Complex, Ganeshkhind, Pune- 411007

Email: shekhar.mande@gmail.com

S.1 Data

Covid-19 Mortality data

We used the https://www.covid19india.org API to fetch the COVID-19 mortality numbers for each state. The MoHFW was the official site to host the COVID deaths and was the original source for this data, it also gives the sub national and national level real time updated data. This data was accumulated data up to the two peak points of total cases in India. As of today, the datasets do not give information about the patient's details such as residence and do not account for the deaths being declared based on symptoms or a positive test.

Disease Prevalence data

The India State-Level Disease burden consortium in India was launched to measure the State level health metrices such as prevalence, incidences and disability-adjusted life year rate in India. Indian Council of Medical Research (ICMR), the Institute for Health Metrics and Evaluation (IHME) and the Public Health Foundation of India (PHFI) have collaborated for the analysis of the state level data. Both communicable and non-communicable disease matrices are available for comparison in different metrices. For interactive data https://vizhub.healthdata.org/gbd-compare/india API is available to compare and fetch the prevalence percentage. Our datasets were adjusted for age and sex. The year 2019 was chosen to fetch all the diseases prevalence percentages in states.

Sanitation and Lifestyle data

The National health profile report 14th issue is available at https://www.cbhidghs.nic.in/. It is a robust analysis to present all health related parameters of states, including allocation of funds, equipment, sanitation, BMI, infant vaccination coverage of states. All the variables are mentioned in terms of the percentage of household reporting that variable analysed in the state. The lifestyle related parameters such as child and maternal nutrition, fasting blood glucose percentages were also captured from this health report.

Demographic parameters

Population, density, gender ratio, age profile and urban population percentage were obtained from the 2011 Census of India. The 2011 Census was used as no new Census was carried out in the country after 2011 due to Covid-19. The census was planned for 2021, however, due to Covid-19 pandemic, this census could not be carried out as planned. Moreover, the long term population characteristics of each state has remained relatively stable over the years, and we therefore believe that it makes minimal difference to the present study. The national Office of the Registrar also maintains the literacy rates of state along with the income of the states such as the GSDP. The Good Governance Index for the year 2019 was collected from http://data-analytics.github.io/Good_Governance/. The index compositely ranks each state in terms of 10 sectors and 58 indicators based on economic, social and health parameters. Further all description source and values of parameters is available in datasets.

S.2 Study Design

The deaths occurred in each state was normalised by the state's population to achieve deaths per million. The variables were checked for their completeness and covariance with other variables. We checked the standard deviation and distribution of each variable to see what statistical test would be appropriate for determining which variable has the maximum impact in explaining the variation in deaths across the states.

We fit a multivariate linear regression, using COVID-19 deaths per million as the outcome variable and the demographic, health, diseases, vaccination and sanitation parameters as explanatory variable to account for the variability in the deaths. We had two different outcome variables corresponding to the two peak dates of the COVID cases and fit the explanatory variables separately with the two outcome variables. We kept adjusting the model for potential confounders by dropping each variable one by one to see if they had significant role in improving the regression score. Since we had a large number of parameters we chose the parameters that showed a significant correlation of greater than 0.4 with our outcome variables.

Coefficient of determination, is a parameter of linear regression that determines the variance proportion in our dependent variable which is influenced by the explanatory variables. We used the Adjusted R^2 value to choose our best model. The adjusted R^2 adjusts the R^2 value based on the number of predictors and the sample size.

S.3 Graphs comparing quality of healthcare and parameters of sanitation.

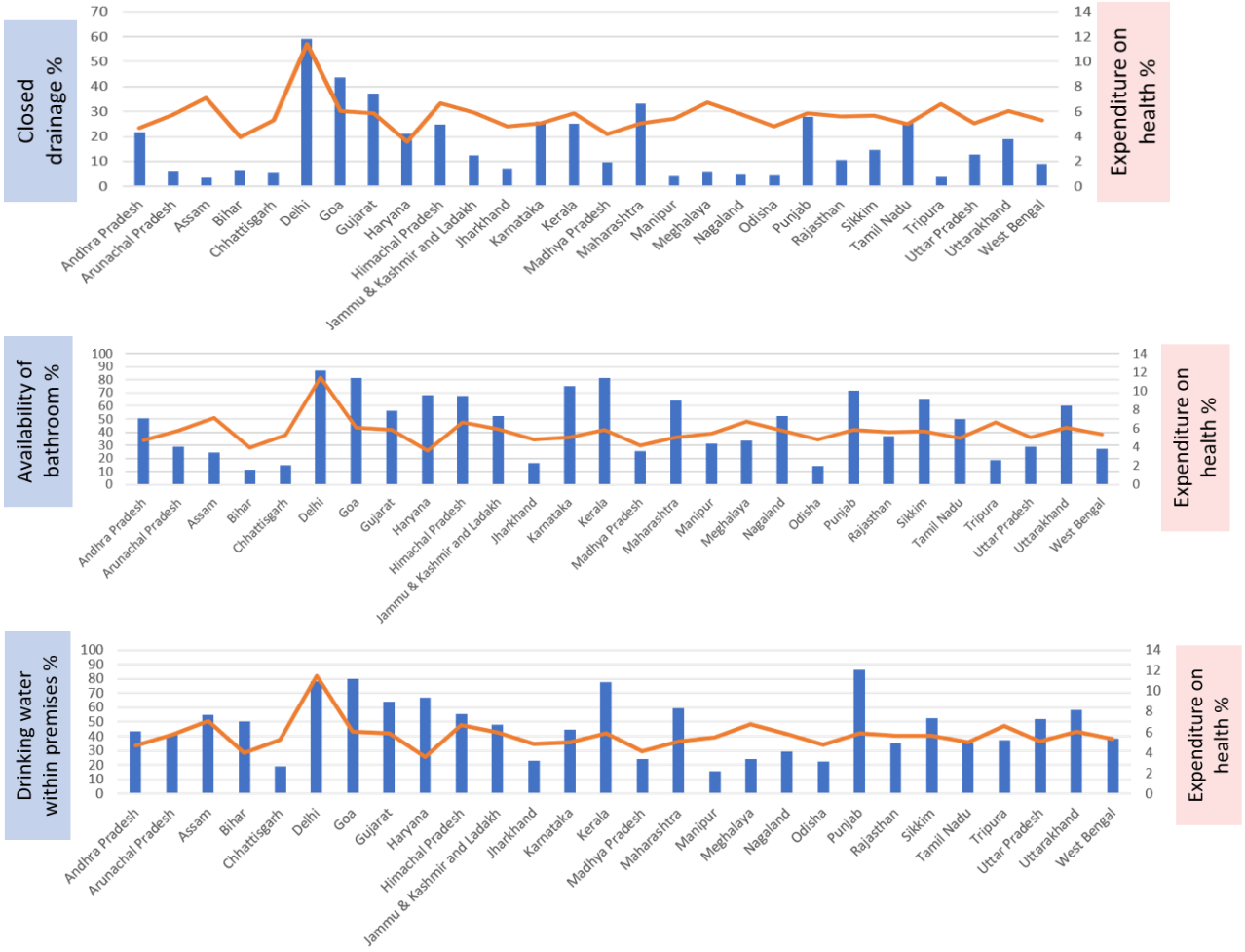

Figure S1 showing the comparison of states with respect to their quality of health care represented by the states percentage of GSDP spent on health expenditure shown as a line in red with axis on the right and the percentage coverage of sanitation parameters namely Closed drainage, Availability of bathrooms and Availability of drinking water within the household premises. The sanitation parameters are shown in blue with axis on left.