scientific reports

OPEN

Demography, sanitation and previous disease prevalence associate with COVID-19 deaths across Indian States

Bithika Chatterjee[1] & Shekhar C. Mande[1,2]✉

The severity of COVID-19 has varied across regions, with a disproportionately higher case-fatality ratio in developed nations. In India, states with higher income have reported more COVID-19 related deaths compared to lower-income states. Understanding the underlying factors such as demographics, disease burden, urbanization, and sanitation can help in designing better public health policies to mitigate future pandemics. The objective of this study is to identify key predictors of COVID-19 mortality across Indian states by examining the role of disease prevalence, demographics, urbanization, and sanitation. We analysed data from the Global Burden of Diseases India 2019 and the National Health Profile 2019, correlating them with COVID-19 mortality during two peak periods of the pandemic. Spearman correlation analysis and multivariate regression models were employed to determine significant associations and build predictive models for COVID-19 deaths. Our analysis showed a positive correlation between COVID-19 mortality and demographic factors such as the percentage of the elderly population (ρ = 0.44, *p* < 0.05 for the first peak; ρ = 0.46, *p* < 0.05 for the second peak). Urbanization was also significantly associated with higher mortality (ρ = 0.71, *p* < 0.05 for the first peak; ρ = 0.57, *p* < 0.05 for the second peak). Additionally, the prevalence of autoimmune diseases and cancer correlated positively with deaths. An unexpected finding was the positive correlation between improved sanitation (e.g., closed drainage systems and indoor toilets) and COVID-19 mortality. The best-fit multivariate regression model, combining demographics, sanitation, autoimmune diseases, and cancer, achieved an adjusted R^2 of 0.71 for the first peak and 0.85 for the second peak. Our findings suggest that as states become wealthier, they undergo urbanization and infrastructural improvements, including better sanitation. However, these changes may also be associated with a rise in autoimmune diseases and cancer, potentially reducing immune resilience to emerging infections. This study provides novel insights into how improved living conditions and lifestyle changes may paradoxically contribute to increased COVID-19 mortality. By emphasizing the role of immune training in pandemic preparedness, our research offers a new perspective on public health strategies for mitigating future infectious disease outbreaks.

Keywords COVID-19, Indian States, Hygiene, Risk factors, Cancer, Autoimmune diseases

The COVID-19 pandemic has caused significant morbidity and mortality worldwide, profoundly impacting public health systems and economies. Since the emergence of SARS-CoV-2 in late 2019, the virus has spread rapidly across countries, demonstrating varying degrees of severity across different populations[1]. While many factors, including viral mutations, healthcare infrastructure, and public health responses, have influenced pandemic outcomes, a notable and often overlooked pattern has emerged, higher-income nations have experienced disproportionately high case-fatality rates compared to lower-income nations. The most probable factors indicate that demographics, such as the percentage of elderly[2,3] population, individuals with pre-existing health conditions and comorbidities, and frontline workers are acknowledged to be at a heightened risk of severe illness and mortality[4]. We believe that there might be more unique underlying factors responsible for exhibiting such paradoxical behavior to COVID-19 mortality beyond the conventional risk factors such as age, comorbidities, and healthcare access. It is therefore imperative to comprehend the multifaceted factors

[1]National Centre for Cell Science, NCCS Complex, Ganeshkhind, Pune 411007, India. [2]Bioinformatics Centre, Savitribai Phule Pune University, 411007 Ganeshkhind, Pune, India. ✉email: shekhar.mande@gmail.com

influencing COVID-19 outcomes to formulate effective public health policies and interventions for mitigating the pandemic's impact especially in the Indian context.

In our previous investigation, we endeavored to comprehend the determinants contributing to the divergent mortality rates observed across various countries. A discernible global pattern became evident during our analysis. Nations exhibiting superior access to sanitation, along with improved demographic indicators such as heightened life expectancy, elevated GDP, and an increased prevalence of autoimmune diseases, seemed to correlate with heightened COVID-19 mortality[5]. Building on these findings, our current study aims to explore which socio-economic, demographic, and health-related factors are associated with COVID-19 mortality across Indian states. Does improved sanitation, infrastructure correlate with higher COVID-19 mortality in Indian states, as observed at the global level? Within states does autoimmune diseases and other pre-existing health conditions influence variations in COVID-19 deaths? Moreover our objective was to build an accurate statistical model to predict the mortality of a state using the most influential variables contributing towards COVID-19 mortality.

Existing studies have established that individual risk factors allows the virus to manifests differently in populations, with specific comorbidities and underlying health conditions identified as significant risk factors for severe outcomes, including death[6]. Numerous studies have underscored the role of lifestyle-related conditions, such as obesity, hypertension, diabetes, and even certain cancers, as notable risk factors for COVID-19 mortality[7–9]. One research study suggests that improved sanitation could alter a population's exposure to pathogens, affecting immune training and susceptibility to severe infections[10]. This perspective aligns with the hygiene hypothesis, which proposes that reduced exposure to environmental microbes may impact immune responses. Despite these insights, most existing studies have focused on high-income countries, limiting insights into how these factors operate in diverse socio-economic settings. Furthermore, few studies have examined these relationships within a single country where regional disparities in wealth, healthcare access, and environmental exposures exist.

Our examination of data from various states in India has unveiled a notable correlation between specific sanitation parameters and COVID-19 mortality, particularly those associated with closed drainage and indoor toilets. We postulate that this correlation may be elucidated by evidence suggesting the airborne nature of COVID-19[11] which can potentially propagate within households through toilets, even when patients are in isolation[12–14]. Furthermore, our investigation has revealed that prior exposure to infections caused by viruses and bacteria may confer a protective immune response against COVID-19. Conversely, autoimmune diseases and certain lifestyle-related conditions may exacerbate an already heightened inflammatory response in the body, potentially elevating the risk of severe illness and death. This observation aligns with the hygiene hypothesis, suggesting that a person's immune system becomes more robust with increased exposure to pathogens throughout their lifetime[15,16]. This may elucidate the comparatively lower fatality rates observed in African and Asian countries, where populations have historically encountered numerous tropical diseases and parasitic infections[17]. This study contributes to the existing literature by introducing a novel perspective on the relationship between sanitation infrastructure, immune susceptibility, and COVID-19 mortality by providing statistical evidences. These insights can inform future epidemiological studies and public health policies aimed at balancing sanitation improvements with broader strategies for pandemic preparedness.

Methodology

The state-wise data on COVID-19 mortality was gathered from https://www.covid19india.org, approximately corresponding to the peak COVID-19 cases in India, as indicated by the World Health Organization[18]. While the cases surged at three distinct time points, our analysis focused solely on deaths during the first and second peaks. By the time the third peak occurred, a substantial portion of the Indian population had been vaccinated[18]. The first peak was observed around September 17, 2020, and the second peak occurred on May 7, 2021. These reported deaths are the cumulative deaths reported in that state and do not report the patient's residence. However much of the population remained at their local residence due to strict lockdown measures implemented by the government. To standardize the cumulative deaths recorded on both days, we normalized the data by the state's and Union Territory's (UT) total population, sourced from the Census of India 2011. The 2011 Census was used as no new Census was carried out in the country after this due to COVID-19. Moreover the long term population characteristics of each state has remained relatively stable over the years so it makes minimal difference to the present study[19]. The population data for the states for comparison of the trend is available at https://censusindia.gov.in/nada/index.php/catalog/42611. This normalization yielded deaths per million (DPM) as the chosen parameter, a more reliable metric compared to cases or case-fatality ratio.

It's worth noting that Telangana was excluded from the analysis due to the unavailability of census data for this state in 2011. Additionally, many parameters were missing for this newly formed state. Mizoram was also omitted from the analysis as it reported no deaths during one of the peaks of COVID-19 cases. In total, data from 26 Indian states and 2 Union territories were collected and analyzed to provide a comprehensive understanding of the state-wise variations in COVID-19 mortality. Delhi is treated as a union territory rather than a state.

The demographic and socio-economic parameters for our analysis were sourced from various reputable databases. Population density, gender ratio, and urban population percentage were obtained from the 2011 Census of India[20]. Literacy rates for each state and Union Territory were sourced from the Office of the Registrar General and Census Commissioner[21]. The Gross State Domestic Product (GSDP) per capita for the year 2018-19 was acquired from https://statisticstimes.com/economy/india/indian-states-gdp-per-capita.php. The Good Governance Index for the year 2019 was collected from http://data-analytics.github.io/Good_Governance/. Prevalence data for communicable and non-communicable diseases were extracted from the Global Burden of Diseases India 2019 database (https://vizhub.healthdata.org/gbd-compare/india#0). These prevalence percentages were standardized by age and obtained for both genders. Additionally, data on infant vaccination parameters, slum households, below poverty line population, sanitation, and expenditure on health hospital beds were sourced from the National Health Profile 2019, 14th issue, available at https://www.cbhidghs.nic.in/. For the urban percentage of population in newly formed state Chhattisgarh, we used the same value from Bihar, as these states have comparable state characteristics such as age, sex ratio as seen in supplementary S1 dataset. This comprehensive dataset allows for a robust and nuanced analysis of the various factors influencing COVID-19 mortality across different states and Union Territories in India.

Statistical analysis

We conducted Spearman correlation coefficient analyses to assess the relationships between various variables and Deaths Per Million (DPM) on two significant dates, namely September 17, 2020 (DPM1), and May 7, 2021 (DPM2). The chosen variables were those demonstrating correlation coefficients above 0.40 or below − 0.40, indicative of positive or negative associations with COVID-19 mortality. While there is no universal threshold for meaningful correlations, several studies consider |ρ| ≥ 0.40 as a moderate-to-strong correlation[22,23]. Choosing a moderate threshold helps filter out weak associations while simultaneously capturing variables that hold promise of significant association. Subsequently, we employed a multivariate linear regression analysis to explore the selected parameters' combined impact on DPM.

To enhance the interpretability of the results, we amalgamated similar variables and calculated adjusted R-squared values for DPM. These adjusted R-squared values elucidate the percentage of variability in DPM that can be explained by each of the broader category variables. Furthermore, we extended our analysis by combining the broader variables exhibiting adjusted R-squared values exceeding 0.40 for both DPM1 and DPM2.

All statistical analyses were conducted using R (version 4.3.1). Linear regression was performed using the lm() function, while Spearman's correlation was computed using the cor() function, both were utilised from the stats package of R base 2023. Data plots and visualization were carried out using the ggplot2[24] .

Choosing the two peak cases dates

As of March 16, 2023, the World Health Organization (WHO) has reported a total of 4,46,91,956 confirmed cases of COVID-19 in India, with 5,30,789 recorded deaths since the onset of the pandemic[18]. Over the course of the three years of the pandemic in India, the country experienced three distinct surges in COVID-19 cases. The first peak surged in the middle of September 2020 with 6,46,263 cases and 8166 deaths. This initial surge was characterized by the prevalence of the milder Alpha variant. Subsequently, the second peak emerged in early May 2021, with a significant increase to 27,38,957 cases and 28,982 deaths[18]. Notably, the second wave was dominated by the more severe Delta variant, leading to a higher casualty rate[25]. The third peak took place in January 2022, resulting in 7,888 deaths attributed to the milder Omicron variant. It is worth noting that by this time, substantial vaccination coverage had been achieved across the states, contributing to the mitigation of the impact of the Omicron variant[18]. Hence, we proceeded with choosing only two peak dates for our analysis.

Results

Demography

In our analysis, a noteworthy observation emerges as we identify higher COVID-19 mortality rates in wealthier states, exemplified by Delhi leading the statistics, followed by Maharashtra (Fig. 1a, b). In contrast, states with lower income levels, such as Bihar and Jharkhand, reported comparatively lower death rates. This pattern prompts further exploration to elucidate the factors contributing to the significant variation in COVID-19 mortality across different Indian states.

Figure 2a, b reveals a discernible linear relationship between Deaths Per Million (DPM) and Gross State Domestic Product (GSDP) per capita for all states, with a Spearman correlation coefficient of 0.66 with DPM2 and 0.54 with DPM1 (Table 1). This statistical correlation underscores the association between economic affluence, as measured by GSDP per capita, and COVID-19 mortality.

The percentage of the elderly population (over 60) exhibited a positive correlation with both DPM1 and DPM2, indicating a potential influence of age demographics on COVID-19 mortality. Similarly, positive correlations were observed with the literacy percentage and the Good Governance Index, emphasizing the significance of education and effective governance in pandemic outcomes. The most robust correlation was observed with Urbanization percentage, demonstrating a Spearman correlation coefficient (rho) of 0.71 for DPM1 and 0.57 for DPM2 (Table 1). This highlights the pivotal role of urbanization in shaping COVID-19 mortality patterns, suggesting that higher levels of urbanization may contribute to increased mortality rates. Conversely, the percentage of the population below the poverty line emerged as the only parameter with a negative correlation with both DPM1 and DPM2 (Table 1). This implies that regions with higher poverty percentages may experience comparatively lower COVID-19 mortality rates. In a comprehensive linear regression analysis, combining all considered parameters except Gross State Domestic Product (GSDP) per capita, we obtained adjusted R-squared values of 0.67 and 0.51 for DPM1 and DPM2, respectively (Fig. 3a).

Vaccination

The Universal Childhood Immunization Program in India, encompassing vaccines such as BCG, POLIO, DPT, Measles, and Hepatitis, has been a critical public health initiative. Our analysis reveals intriguing associations between vaccine coverage percentages and COVID-19 mortality. Positive correlations were observed between COVID-19 deaths (DPM1 and DPM2) and the vaccine coverage percentages of BCG, POLIO, DPT, Measles, and Hepatitis (Table 1). Conversely, vaccine percentages for Acute Hepatitis B and Acute Hepatitis E demonstrated negative correlations with both DPM1 and DPM2 (Table 1). The dosage of Vitamin A administered during infancy also exhibited a positive correlation with COVID-19 deaths (Table 1). Combining the vaccination

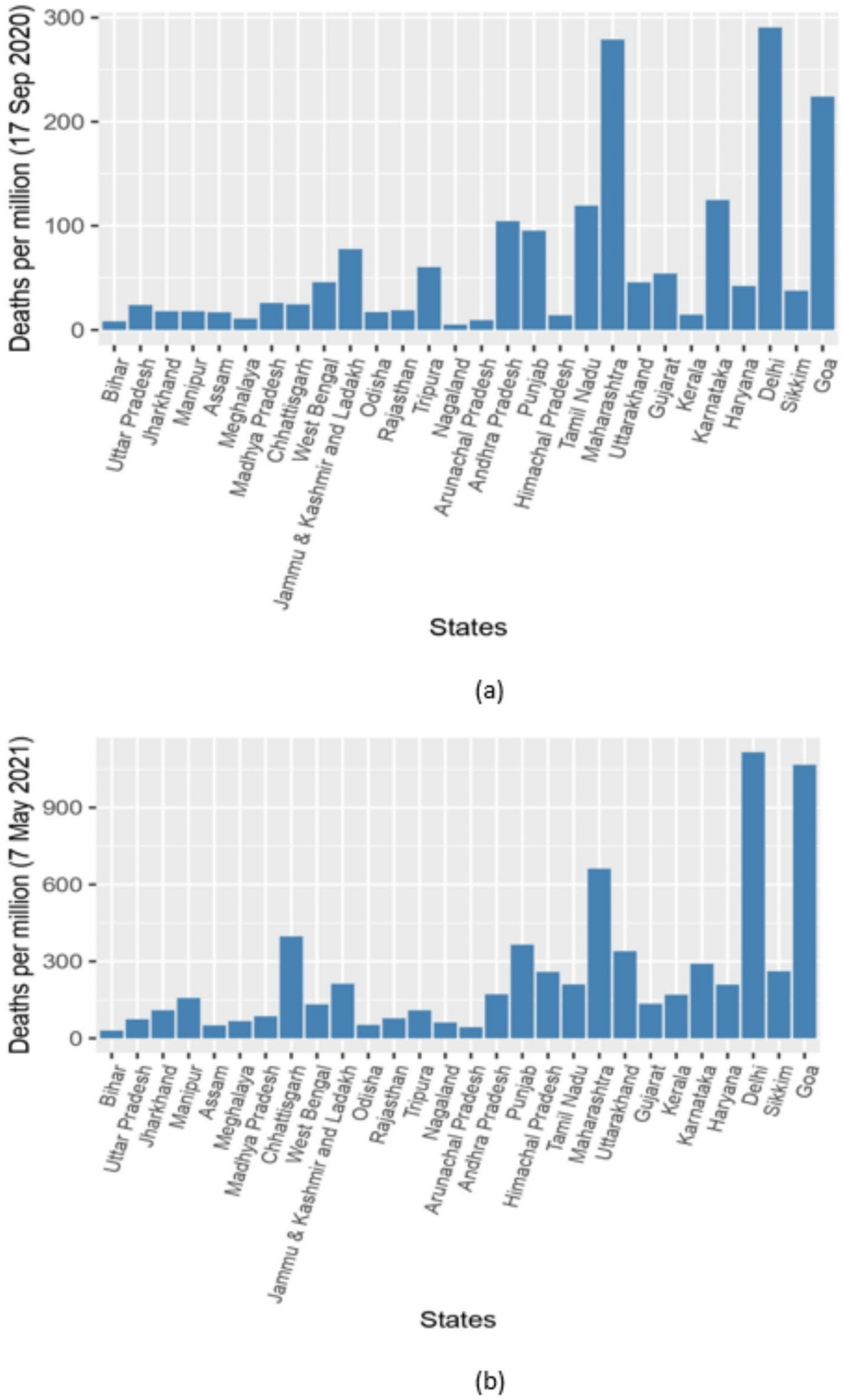


Fig. 1. (**a**) Number of deaths per million inhabitants in each state due to COVID-19 as on 17 September 2020. The states are arranged in ascending GSDP per capita for year 2018 to 2019. (**b**) Number of deaths per million inhabitants in each state due to COVID-19 as on 7 May 2021. The states are arranged in ascending GSDP per capita for year 2018 to 2019.

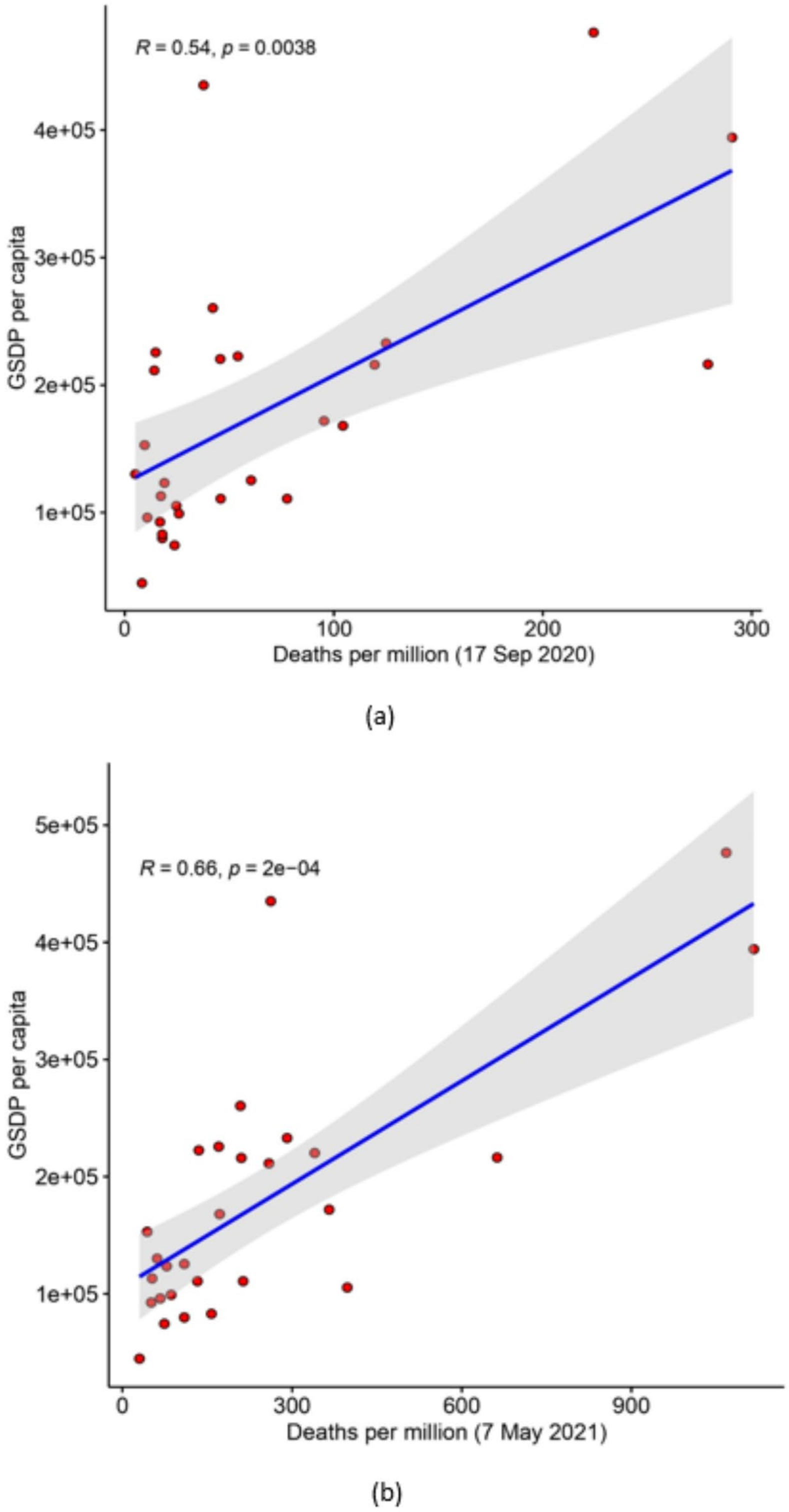


Fig. 2. (**a**) Plot showing positive correlation between Deaths due to COVID-19 as on 17 September 2020 vs. the state's GSDP per capita for year 2018 to 2019. R denotes Spearman's correlation coefficient of 0.54. Each dot denotes an Indian state($N = 28$). (**b**) Plot showing positive correlation between Deaths due to COVID-19 as on 7 May 2021 vs. the state's GSDP per capita for year 2018 to 2019. R denotes Spearman's correlation coefficient of 0.66. Each dot denotes an Indian state($N = 28$).

	Deaths per million as on 17-09-2020			Deaths per million as on 07-05-2021		
	rho	S	p-value	rho	S	p-value
Explanatory variables						
Demography						
GSDP per capita(2018–2019)	0.54	1696	0.004	0.66	1248	0.0002
Age over 60(%)	0.44	2060.47	0.02034	0.46	1979.4	0.01418
Literacy(%)	0.41	2138	0.02906	0.59	1500	0.00119
Urbanization(%)	0.71	1074.65	0.00003	0.57	1567.71	0.00151
Good Governance Index(2019)	0.46	1963.77	0.01319	0.47	1922.76	0.01087
Population below poverty(%)	-0.4	5113.2	0.03527	-0.53	5599.27	0.00354
Slum household(%)	0.44	2053.56	0.01974	0.11	3263.89	0.5887
Sanitation						
Drinking water within premises(%)	0.39	2236	0.04216	0.47	1938	0.01249
Bathroom available(%)	0.47	1926.53	0.01107	0.69	1126.31	0.00005
Bathing in enclosure without roof(%)	-0.13	4128	0.50899	-0.4	5122	0.03498
Closed drainage(%)	0.67	1212	0.00015	0.71	1070	0.00004
No drainage(%)	-0.49	5444	0.00885	-0.63	5950	0.00046
Unsafe water sanitation and handwash deaths(%)	-0.38	5028	0.04941	-0.62	5934	0.00051
Vaccination						
BCG(%)	0.41	2160.59	0.03082	0.65	1283.35	0.00019
POLIO(%)	0.29	2584	0.13041	0.52	1756	0.00517
DPT(%)	0.35	2392.83	0.07205	0.58	1529.71	0.00118
Measles(%)	0.47	1926.76	0.01108	0.69	1122.65	0.00004
Hepatitis(%)	0.22	2868	0.2704	0.47	1942	0.01271
Varicella and herpes zoster(%)	0.36	2356	0.06427	0.59	1484	0.00108
Acute hepatitis B(%)	-0.49	5460	0.00819	-0.49	5440	0.00902
Acute hepatitis E(%)	-0.30	4746	0.12241	-0.50	5474	0.00765
Vit A(%)	0.51	1778	0.00579	0.51	1794	0.00627
Tropical diseases						
Trichuriasis(%)	-0.4	5114	0.03516	-0.43	5232.33	0.02171
Leprosy(%)	-0.44	5252	0.02085	-0.59	5812	0.00116
Hookworm(%)	-0.44	5274.47	0.01809	-0.56	5703.77	0.0019
Cystic echinococcus(%)	0.19	2962.49	0.33479	0.45	2014.94	0.01666
Ascaris(%)	-0.32	4826	0.0964	-0.47	5382	0.01182
Autoimmune diseases						
Gout(%)	0.4	2178	0.03393	0.56	1600	0.00219
Psoriasis(%)	-0.5	5476	0.00758	-0.63	5956	0.00044
Diabetes mellitus type 2(%)	0.4	2192	0.03579	0.53	1704	0.00393
Asthma(%)	-0.51	5530	0.00579	-0.64	5998	0.00032
IBD(%)	0.45	2002	0.01658	0.64	1316	0.00034
Viral diseases						
Measles(%)	-0.57	5733.28	0.00158	-0.73	6309.36	0.00001
Rabies(%)	-0.48	5414.24	0.00944	-0.45	5306.23	0.0157
Meningitis(%)	-0.46	5322	0.01546	-0.59	5818	0.00112
Acute hep(%)	-0.6	5848	0.00092	-0.63	5968	0.0004
Dengue(%)	0.6	1470	0.00098	0.4	2208	0.03801
Bacterial diseases						
Whooping cough(%)	-0.51	5500	0.00673	-0.65	6032	0.00025
Diphtheria(%)	-0.44	5252	0.02085	-0.64	5996	0.00033
UTI(%)	0.41	2174	0.03342	0.59	1504	0.00122
Chlamydia(%)	-0.6	5858	0.00086	-0.6	5860	0.00085
Respiratory diseases						
Upper respiratory infections(%)	-0.5	5482	0.00736	-0.54	5640	0.00322
Interstitial lung disease and pulmonary sarcoidosis(%)	0.42	2108	0.0258	0.35	2378	0.06917
Lifestyle diseases						
Child maternal nutrition(%)	-0.54	5638	0.00326	-0.7	6226	0.00005
Continued						

	Deaths per million as on 17-09-2020			Deaths per million as on 07-05-2021		
	rho	S	p-value	rho	S	p-value
High bmi deaths(%)	0.65	1278	0.00025	0.73	994	0.00002
High fasting plasma glucose(%)	0.69	1144	0.00008	0.72	1016	0.00002
High systolic blood pressure(%)	0.53	1704	0.00393	0.55	1636	0.00269
Cancer						
Breast cancer(%)	0.46	1966	0.01416	0.44	2028	0.01854
Brain and central nervous system cancer(%)	0.5	1834	0.00765	0.59	1512	0.00129
Prostate cancer(%)	0.35	2390	0.07196	0.61	1436	0.00078
Kidney cancer(%)	0.5	1816.5	0.00638	0.73	984.27	0.00001
Hodgkin lymphoma(%)	0.56	1610.72	0.00198	0.56	1591.72	0.00176
Non Hodgkin lymphoma(%)	0.49	1874.54	0.00858	0.43	2067.74	0.02099

Table 1. Shows the spearman correlation coefficient(rho), S-statistic(S) and P value (p-value) of explanatory variables with COVID-19 deaths/million in Indian States as on 17-09-2020 and 07-05-2021 respectively; $N=28$. Refer to the supplementary dataset for details of the variables.

parameters in linear regression yielded adjusted R-squared values of 0.37 and 0.62 for DPM1 and DPM2, respectively (Fig. 3a).

Tropical diseases

The prevalence of tropical diseases, including Trichuriasis, Leprosy, Hookworm, Ascaris, and Cystic Echinococcus, introduces a nuanced perspective on their potential relationship with COVID-19 mortality across states. Most tropical diseases, such as Trichuriasis, Leprosy, Hookworm, and Ascaris, displayed negative correlations with COVID-19 deaths (DPM1 and DPM2), indicating lower mortality rates in states with higher prevalence of these diseases (Table 1). Notably, Cystic Echinococcus exhibited a positive correlation with DPM, suggesting a potential association with increased COVID-19 mortality (Table 1). Combining the tropical diseases parameters in linear regression yielded adjusted R-squared values of 0.15 and 0.33 for DPM1 and DPM2, respectively (Fig. 3a). Given the low prevalence of these diseases in most states, further research is warranted to establish conclusive associations.

Autoimmune diseases

Positive correlations were observed between the prevalence of Gout, Diabetes Mellitus Type 2, and Inflammatory Bowel Disease (IBD) with COVID-19 deaths (DPM1 and DPM2), indicating potential higher mortality rates in states with a higher prevalence of these conditions (Table 1). Conversely, negative correlations were noted for Asthma and Psoriasis, suggesting potential lower mortality rates in states with a higher prevalence of these conditions (Table 1). Regression analysis with DPM1 and DPM2 yielded adjusted R-squared values of 0.42 and 0.46, respectively, suggesting that the considered chronic diseases collectively contribute to the observed variability in COVID-19 mortality (Fig. 3a).

Viral diseases

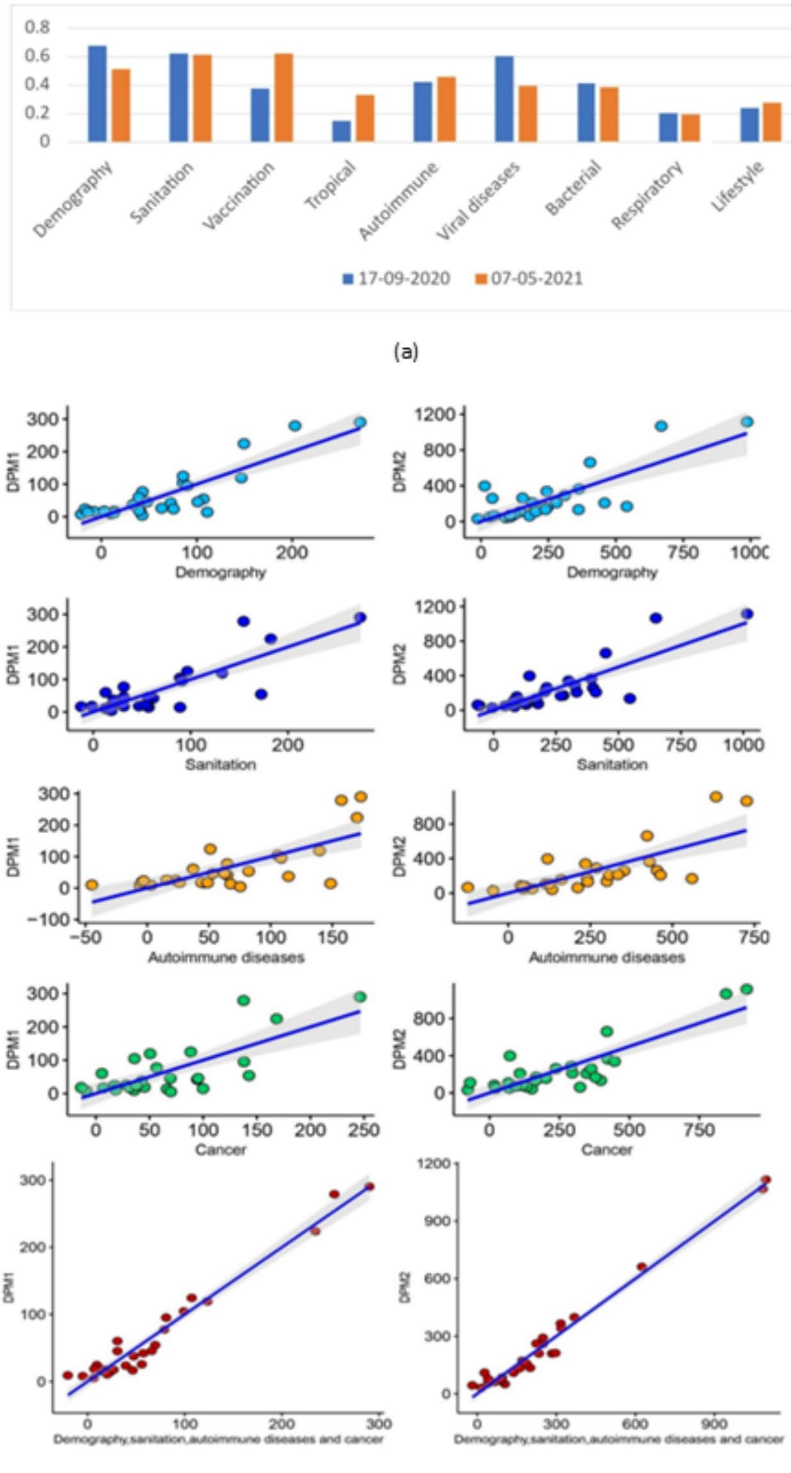
Negative correlations were observed between the prevalence of Rabies, Meningitis, and Acute Hepatitis with COVID-19 deaths (DPM1 and DPM2), indicating potential lower mortality rates in states with a higher prevalence of these infectious diseases (Table 1). Measles showed the highest negative correlation, with coefficients of -0.57 and −0.73 for DPM1 and DPM2, respectively, suggesting a pronounced protective association (Table 1). In contrast, Dengue prevalence exhibited a positive correlation with COVID-19 deaths, indicating potential higher mortality rates in states with a higher prevalence of Dengue (Table 1). Regression analysis with DPM1 and DPM2 yielded adjusted R-squared values of 0.61 and 0.39, respectively, highlighting the potential collective impact of these infectious diseases on COVID-19 mortality (Fig. 3a).

Bacterial diseases

The prevalence of infectious diseases, including Whooping Cough, Chlamydia, and Diphtheria, demonstrated negative correlations with COVID-19 deaths (DPM1 and DPM2), suggesting potential lower mortality rates in states with a higher prevalence of these diseases. In contrast, Urinary Tract Infection (UTI) exhibited a positive correlation with COVID-19 deaths, indicating potential higher mortality rates in states with a higher prevalence of UTI (Table 1). Combining these infectious disease parameters in linear regression analysis produced statistically significant adjusted R-squared values of 0.4 (p-value = 0.0024) and 0.39 (p-value = 0.0035) for DPM1 and DPM2, respectively (Fig. 3a).

Respiratory diseases

Upper Respiratory Infection exhibited a negative correlation with COVID-19 deaths (DPM1 and DPM2), indicating a potential lower mortality rate in states with a higher prevalence of this respiratory condition. Conversely, Interstitial Lung Disease and Pulmonary Sarcoidosis showed positive correlations with COVID-19



(a)

(b)

Fig. 3. (**a**) Regression analysis was done with COVID-19 deaths/million in Indian states as on 17-09-2020 and 07-05-2021 as the dependent variable and combinations of different variables; $N = 28$. Graph shows the Adjusted R- squared values depicting how much percentage of variability in the mortality could be explained by the variables in each category. (**b**) Actual values of deaths per million as on 17 September 2020 (DPM1) and 7 May 2021(DPM2) plotted against their predicted values by combining variables of demographics, sanitation, autoimmune diseases and cancer separately and then combining all of them together. Regression analysis was done with DPM1 and DPM2 as the dependent variable and combinations of different variables as explanatory variables; $N = 28$. Each dot represents an Indian state.

deaths, suggesting potential higher mortality rates in states with a higher prevalence of these respiratory diseases (Table 1).

Linear regression analysis combining these respiratory disease parameters produced adjusted R-squared values of 0.20 and 0.19 for DPM1 and DPM2, respectively (Fig. 3a). These results indicate that, in our studies, no significant association of respiratory diseases with deaths due to COVID-19 could be identified.

Lifestyle diseases

A high percentage of BMI (Body Mass Index) deaths, elevated fasting plasma glucose, and high systolic blood pressure exhibited positive correlations with COVID-19 deaths (DPM1 and DPM2), implying potential higher mortality rates in states with a higher prevalence of these health indicators. Conversely, a good percentage of child and maternal nutrition showed negative correlations, suggesting potential lower mortality rates in states with better nutritional indicators (Table 1).

The linear regression analysis combining these health indicators yielded adjusted R-squared values of 0.24 and 0.28 for DPM1 and DPM2, respectively (Fig. 3a). However, these values did not reach statistical significance in our studies, indicating that the observed associations were not robust enough to consider these health indicators as significant predictors of COVID-19 mortality.

Cancer

All cancers except prostate cancer displayed a high positive correlation with COVID-19 deaths (DPM), indicating a potential association between overall cancer prevalence and higher mortality rates across states (Table 1). Since prostate cancer still displays a high positive correlation with DPM2 we include it in our analysis. Combining all cancer parameters for linear regression analysis with DPM1 and DPM2 produced adjusted R-squared values of 0.45 and 0.64, respectively (Fig. 3a). These results highlight a substantial explanatory power of cancer prevalence in understanding the observed variability in COVID-19 mortality.

The robust associations observed between overall cancer prevalence and COVID-19 mortality underscore the importance of considering cancer as a significant factor in the broader context of health outcomes during the pandemic.

Sanitation

Improved sanitation parameters across states, including the percentage of drinking water within premises and bathroom availability, displayed positive correlations with COVID-19 deaths (DPM1), indicating potential higher mortality rates in states with better sanitation infrastructure. Notably, Closed drainage showed a very high correlation, with rho values of 0.67 and 0.71 for DPM1 and DPM2, respectively (Table 1).

Conversely, parameters indicative of poor sanitation, such as the percentage of bathing in an enclosure without a roof, lack of drainage, and percentage of deaths due to unsafe handwashing and sanitation, exhibited negative correlations with COVID-19 deaths, suggesting potential lower mortality rates in states with inadequate sanitation facilities (Table 1).

The linear regression analysis, combining these sanitation parameters, yielded adjusted R-squared values of 0.62 and 0.61 for DPM1 and DPM2, respectively (Fig. 3a). These values indicate a substantial explanatory power of sanitation parameters in understanding the observed variability in COVID-19 mortality. The strong associations observed underscore the critical role of sanitation infrastructure in influencing COVID-19 outcomes, emphasizing the importance of public health measures related to sanitation.

Multivariate linear regression

The combination of broader variables encompassing demography, sanitation, autoimmune diseases, and cancer collectively produced adjusted R-squared values of 0.71 and 0.85 for DPM1 and DPM2, respectively (Fig. 3b). These high values indicate a substantial explanatory power of these combined variables in understanding the observed variability in COVID-19 mortality. Although the GSDP variable individually was found to be highly correlated with COVID-19 mortality, it becomes insignificant when other variables are accounted for. When GSDP was included among the explanatory variables it further reduced the R-squared value to 0.62 and 0.82 with DPM1 and DPM2 respectively. The correlation between the GSDP and the residuals obtained from the two models using DPM1 and DPM2 was insignificant (-0.040 and 0.12 respectively) which denotes that GSDP contributes very less to whatever remains unpredicted by the two models, thus we removed the variable as we found it a confounding factor.

States with a higher percentage of older population, improved sanitation parameters, elevated prevalence of autoimmune disorders, and increased cancer rates were more likely to experience higher mortality rates due to COVID-19. The intricate interplay of these factors underscores the complexity of the determinants influencing the impact of the pandemic across states.

Discussion

A discernible pattern of COVID-19 mortality has previously been seen on a global scale, where nations with higher economic affluence have exhibited an elevated mortality rate compared to their lower middle and middle-income counterparts[1]. The major focus of our studies was to examine if the states in India exhibited a similar trend and what factors influence this trend. We emphasized on socioeconomic, demographic, sanitation, and health-related variables. Our examination of various Indian states aligns with this trend, as illustrated in Fig. 1a, b. It is widely acknowledged that economic prosperity exerts substantial influence on determinants like life expectancy, literacy, and enhanced sanitation practices and lower exposure to parasitic infections within a nation. Our findings corroborate this established association, revealing that Indian states characterized by heightened GSDP and improved demographic indicators, including elevated literacy rates and increased life expectancy, have reported a higher incidence of Covid-related deaths, as illustrated in Fig. 3a. We discuss if our findings align with previous knowledge from literature.

Among demographic factors, the positive correlation between elderly population percentages and COVID-19 deaths was an expected factor since several patient's data reveal age as a critical factor[2,3]. Additionally, the observed positive correlation between higher literacy rates in the states and increased deaths can be attributed to enhanced employment opportunities and greater mobility, potentially leading to elevated exposure to the virus. One such study in Italy implicated migration habits of inhabits to increased exposure to COVID[26]. Level of urbanization within a state and higher mortality rate can be attributed by heightened probability of viral transmission in densely populated areas. Contrary to expectations, our analysis suggests that Gross State Domestic Product (GSDP) may act as a confounding variable influenced by other developmental factors. By excluding GSDP per capita from the analysis, we observe that the regression score of remaining developmental factors with DPM1 and DPM2 still remains significant (Fig. 3b), emphasizing their more significant contribution to COVID-19 mortality.

In our earlier analysis[5], we noted, that despite improved sanitation parameters, including access to safe drinking water, hygienic handwashing practices, closed toilets, and efficient drainage systems, affluent nations have faced a disproportionately higher burden of COVID-19 mortality per million individuals. Our current analysis of Indian states reveals a parallel trend, where indicators associated with enhanced sanitation, such as access to safe drinking water, in-house toilets, and closed drainage systems, exhibit a positive correlation with COVID-19 deaths. Among the sanitation parameters the most perplexing observations in our analysis is the robust correlation between deaths per million and the presence of closed drainage and/or the availability of indoor toilets. Intuitively, the availability of closed or open toilets should not inherently influence the spread of the virus or subsequent mortality, although fragments of SARS-CoV-2 RNA have been detected in drainage systems globally[27,28]. A plausible explanation for this unexpected correlation could be rooted in the airborne nature of COVID-19, enabling its transmission throughout an entire household via aerosols, including through toilets, even when patients are in isolation[11,12]. Emerging studies suggest that flushing toilets can facilitate the suspension of the virus in larger droplets, consequently increasing airborne transmission and the infectivity of the virus[13,14]. Furthermore, it is noteworthy that states with open sanitation systems lacking bathrooms, drainage, or roofs tend to exhibit lower COVID-19 death rates. This phenomenon may be attributed to a higher probability of the virus dissipating in the air, leading to reduced transmission to individuals. While we do not advocate practices of lowering sanitation in populations but this observation underscores the need for a comprehensive understanding of the various modes of virus transmission within household settings.

A defining feature of developed states, attributed to superior sanitation conditions, is the lower incidence of communicable diseases, encompassing helminths, parasitic, and viral infections[17]. These diseases are categorized into tropical, viral, and bacterial types. Notably, individual tropical diseases exhibit a negative but low correlation with COVID-19 deaths. This observation may be justified by the likelihood that many states in India, particularly urbanized areas, report minimal rates of these parasitic diseases. Conversely, the incidence of viral and bacterial diseases individually correlates negatively with deaths, exhibiting higher regression scores with DPM1 and DPM2. Existing studies propose that prior infections with viruses, bacteria, and parasites confer protective humoral and cell-mediated immunity that may endure a lifetime[29]. It is suggested that exposure to infections aids dendritic cells in inducing T cell regulation, which contributes to an anti-inflammatory network, thereby suppressing cytokines[30,31]. Given that the majority of Covid deaths and severity result from acute respiratory distress syndrome (ARDS) leading to respiratory failure[32] it may be hypothesized that prior infections assist in modulating an elevated immune response compared to encountering lower infections.

In context of autoimmune disorders, impact on immune reaction, whether heightened or reduced, remains challenging to ascertain, particularly considering that many patients may be under immunosuppressive medication. Notably, diseases like Asthma and Psoriasis, which necessitate constant medication, may not elicit a pronounced reaction to COVID-19 antigens.

Various types of cancers associated with a very high regression score with COVID – 19 mortality. This seems to be an expected factor as numerous studies have identified cancer as a significant comorbidity factor associated with heightened risk in various populations[8]. Moreover, cancer prevalence is known to be elevated in populations with lifestyles associated with urban living[33,34]. Cancer patients may also have their immune cells already overexpressed and hence may succumb by a heightened immune response when encountering an unseen pathogen.

It is imperative to acknowledge that while certain risk factors may exhibit associations with varying rates of Covid-related deaths, these correlations do not inherently imply causation. Additionally, findings derived from population-level studies may not comprehensively capture individual-level distinctions, which are subject to genetic variability. Nonetheless, some correlations can be elucidated through existing domain knowledge and global trends observed in Covid patients. Furthermore, we advocate for exploring the feasibility of microbiome therapy as a preventive measure against future pandemics. Our research endeavors not only contribute to understanding the multifaceted factors influencing COVID-19 outcomes but also provide valuable insights for identifying states necessitating urgent governmental interventions to mitigate COVID-19 mortality.

Limitations

One of the prominent limitations inherent in our regression studies is the inability to capture individual-level variations across all parameters considered. While our analyses elucidate associations at the area level, they may not fully reflect differences in individuals with diverse physiologies and genetic profiles. Additionally, the absence of individual patient's data, such as existing comorbidities or the location of the patient's death, poses a challenge. Acquiring such information is hindered by ethical and data privacy considerations, presenting an avenue for future research within the scope of our project.

Furthermore, we acknowledge the potential for misreporting, miscalculating, or underreporting of deaths in certain states, which could introduce bias into our studies. Access to national registry data with validated causes of death could mitigate this bias, representing an area for future investigation. Some of the states having poor sanitation and unsafe drinking water also seem to have low expenditure on health care as can be seen in figure S1 of supplementary file, it is quite possible that these areas may report lower number of deaths due to COVID.

Conclusions

The findings of this study provide crucial insights into the complex interplay between demography, sanitation, disease prevalence, and COVID-19 mortality thus deviating from conventional assumptions that improved hygiene uniformly leads to better health outcomes. Longitudinal studies examining how changes in sanitation practices over time impact immune system development and disease resistance could provide deeper insights into the interplay between hygiene and immunity. Additionally, extending this analysis to other infectious diseases beyond COVID-19 such as influenza, tuberculosis, or gastrointestinal infections could help determine whether similar patterns exist across different pathogens. These studies could lend more support to our hypothesis and remains future scope of our work. By demonstrating a positive correlation between certain sanitation parameters such as closed drainage and indoor toilets and COVID-19 deaths, this research underscores the need for a more nuanced understanding of how environmental factors influence pandemic outcomes.

Furthermore, by integrating disease prevalence, urbanization, and demographic factors into a multivariate model, our work contributes to the broader discourse on social determinants of health. The identification of autoimmune diseases and cancer as significant predictors of COVID-19 mortality highlights the need for targeted interventions in regions with higher burdens of these conditions. These findings have broader implications for public health policies, particularly in rapidly urbanizing regions where changing sanitation practices may inadvertently alter immune system conditioning and disease susceptibility.

Data availability

The data used for analysis is publicly available. We have also enclosed our dataset used for analysis. The link to the dataset and supplementary file is https://github.com/chattyfoot/Covid-indian-states.git.

Received: 7 June 2024; Accepted: 7 March 2025
Published online: 25 March 2025

References

1. WHO Coronavirus (COVID. -19) Dashboard. https://covid19.who.int
2. Yanez, N. D., Weiss, N. S., Romand, J. A. & Treggiari, M. M. COVID-19 mortality risk for older men and women. *BMC Public. Health*. **20**, 1742 (2020).
3. Mallapaty, S. The coronavirus is most deadly if you are older and male—new data reveal the risks. *Nature* **585**, 16–17 (2020).
4. Esai Selvan, M. Risk factors for death from COVID-19. *Nat. Rev. Immunol.* **20**, 407–407 (2020).
5. Chatterjee, B., Karandikar, R. L. & Mande, S. C. Mortality due to COVID-19 in different countries is associated with their demographic character and prevalence of autoimmunity. *Curr. Sci.* **120**, 00113891 (2021).
6. Gao, Y. et al. Risk factors for severe and critically ill COVID-19 patients: A review. *Allergy* **76**, 428–455 (2021).
7. Gao, F. et al. Obesity is a risk factor for greater COVID-19 severity. *Diabetes Care*. **43**, e72–e74 (2020).
8. Liang, W. et al. Cancer patients in SARS-CoV-2 infection: A nationwide analysis in China. *Lancet Oncol.* **21**, 335–337 (2020).
9. Guo, L. et al. Comorbid diabetes and the risk of disease severity or death among 8807 COVID-19 patients in China: A meta-analysis. *Diabetes Res. Clin. Pract.* **166**, 108346 (2020).
10. Parker, W., Patel, E., Jirků-Pomajbíková, K. & Laman, J. D. COVID-19 morbidity in lower versus higher income populations underscores the need to restore lost biodiversity of eukaryotic symbionts. *iScience* **26**, 106167 (2023).
11. Zhang, R., Li, Y., Zhang, A. L., Wang, Y. & Molina, M. J. Identifying airborne transmission as the dominant route for the spread of COVID-19. *Proc. Natl. Acad. Sci.* **117**, 14857–14863 (2020).
12. Moharir, S. C. et al. Detection of SARS-CoV-2 in the air in Indian hospitals and houses of COVID-19 patients. *J. Aerosol. Sci.* **164**, 106002 (2022).
13. Johnson, D. L., Mead, K. R., Lynch, R. A. & Hirst, D. V. L. Lifting the lid on toilet plume aerosol: A literature review with suggestions for future research. *Am. J. Infect. Control.* **41**, 254–258 (2013).
14. Wang, J. X. et al. Ventilation reconstruction in bathrooms for restraining hazardous plume: Mitigate COVID-19 and beyond. *J. Hazard. Mater.* **439**, 129697 (2022).
15. Bach, J. F. The effect of infections on susceptibility to autoimmune and allergic diseases. *N. Engl. J. Med.* **347**, 911–920 (2002).
16. Okada, H., Kuhn, C., Feillet, H. & Bach, J. F. The 'hygiene hypothesis' for autoimmune and allergic diseases: An update. *Clin. Exp. Immunol.* **160**, 1–9 (2010).
17. Roser, M., Ritchie, H. & Spooner, F. Burden of disease. *Our World Data* [cited 2023 Apr 18]; (2021). https://ourworldindata.org/burden-of-disease
18. India WHO Coronavirus Disease (COVID-19) Dashboard With Vaccination Data. https://covid19.who.int
19. 2011 Census of India. Wikipedia. (2023). https://en.wikipedia.org/w/index.php?title=2011_Census_of_India&oldid=1148488377
20. List of states and union territories of India by population. Wikipedia. (2023). Available from:https://en.wikipedia.org/w/index.php?title=List_of_states_and_union_territories_of
21. List of Indian states and union territories by literacy rate. Wikipedia. (2023). https://en.wikipedia.org/w/index.php?title=List_of_Indian_states_and_u
22. Cohen, J. *Statistical Power Analysis for the Behavioral Sciences* (Routledge, 2013). https://doi.org/10.4324/9780203771587
23. Evans, J. D. *Straightforward Statistics for the Behavioral Sciences* (Brooks/Cole Pub. Co., 1996).
24. Kassambara, A. ggplot2 Based Publication Ready Plots. https://rpkgs.datanovia.com/ggpubr/
25. Tareq, A. M., Emran, T. B., Dhama, K., Dhawan, M. & Tallei, T. E. Impact of SARS-CoV-2 delta variant (B.1.617.2) in surging second wave of COVID-19 and efficacy of vaccines in tackling the ongoing pandemic. *Hum. Vaccines Immunother*. **17**, 4126–4127 (2021).
26. Cartenì, A., Di Francesco, L. & Martino, M. How mobility habits influenced the spread of the COVID-19 pandemic: Results from the Italian case study. *Sci. Total Environ.* **741**, 140489 (2020).
27. Sharif, S. et al. Detection of SARs-CoV-2 in wastewater using the existing environmental surveillance network: A potential supplementary system for monitoring COVID-19 transmission. *PLoS ONE*. **16**, e0249568 (2021).

28. Mousazadeh, M. et al. Wastewater based epidemiology perspective as a faster protocol for detecting coronavirus RNA in human populations: A review with specific reference to SARS-CoV-2 virus. *Pathogens* **10**, 1008 (2021).
29. Blok, B. A., Arts, R. J. W., van Crevel, R., Benn, C. S. & Netea, M. G. Trained innate immunity as underlying mechanism for the long-term, nonspecific effects of vaccines. *J. Leukoc. Biol.* **98**, 347–356 (2015).
30. Cecere, T. E., Todd, S. M. & LeRoith, T. Regulatory T cells in arterivirus and coronavirus infections: Do they protect against disease or enhance it? *Viruses* **4**, 833–846 (2012).
31. Belkaid, Y. Regulatory T cells and infection: A dangerous necessity. *Nat. Rev. Immunol.* **7**, 875–888 (2007).
32. Tang, Y. et al. Cytokine storm in COVID-19: The current evidence and treatment strategies. *Front. Immunol.* **11**, 1708 (2020).
33. Thakur, J. S. et al. Urban–rural differences in cancer incidence and pattern in Punjab and Chandigarh: Findings from four new population-based cancer registries in North India. *Int. J. Noncommunicable Dis.* **2**, 49 (2017).
34. Enayatrad, M. et al. Urbanization levels and its association with prevalence of risk factors and colorectal Cancer incidence. *Iran. J. Public. Health.* **50**, 2317–2325 (2021).

Acknowledgements

We thank Dr. Rajeeva Karandikar for several useful discussions. BC gratefully acknowledges financial support by the National Centre for Cell Science, Pune. Authors declare no competing financial interests. SCM acknowledges Anand Deshpande for philanthropic funding for Distinguished Professorship at the Savitribai Phule Pune University.

Author contributions

SCM and BC conceptualized the study, BC collected the data, BC and SCM analyzed the data, both the authors wrote the manuscript.

Funding

This work was supported by National Centre for Cell Science funding.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-93622-0.

Correspondence and requests for materials should be addressed to S.C.M.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.