Data in Brief 37 (2021) 107169

Contents lists available at ScienceDirect

Data in Brief

journal homepage: www.elsevier.com/locate/dib

Data Article

Heterogeneous patterns of COVID-19 transmission in an Urban set up – sero-epidemiological survey data from Ujjain, Madhya Pradesh (a central Indian city)

Ankur Joshi[a,1], Prem Shankar[b,1], Anirban Chatterjee[a,1], Jitendra Singh[b], Abhijit Pakhare[a], Kriti Yadav[a], Arti Shrivas[b], Anand Kumar Maurya[b], Raunaq Singh Nagi[a], Debasis Biswas[b], Arun M. Kokane[a,*], Sarman Singh[b]

[a] *Department of Community and Family Medicine, All India Institute of Medical Sciences (AIIMS), Bhopal, Madhya Pradesh, India*
[b] *Department of Microbiology, All India Institute of Medical Sciences (AIIMS), Bhopal, India*

ARTICLE INFO

Article history:
Received 9 March 2021
Revised 30 April 2021
Accepted 19 May 2021
Available online 25 May 2021

ABSTRACT

In the wake of rising number of SARS-CoV-2 cases, the Government of India had placed mass-quarantine measures, termed as "lockdown" measures from end-March 2020. The subsequent phase-wise relaxation from July 2020 led to a surge in the number of cases. This necessitated an understanding of the true burden of SARS-CoV-2 in the community. Consequently, a sero-epidemiological survey was carried out in the central Indian city of Ujjain, Madhya Pradesh. This article details the processes of data acquisition, compilation, handling, and information derivation from the survey.

Abbreviations: ANM, Auxiliary Nurse Midwife; COI, Cut-off Index; COVID-19, Coronavirus Disease-19; CT, Cycle Threshold; ECLIA, Electrochemiluminescence Assay; GCI, Galvanised Corrugated Iron; HB, High burden tertile; IB, Intermediate burden tertile; ICC, Intra-class Cluster Coefficient; LB, Low burden tertile; LBD, Laboratory-based data; LPG, Liquified Petroleum Gas; MHW, Multipurpose Health Worker; NR, Non-reactive; PSU, Primary Sampling Unit; QBD, Questionnaire-based data; R, Reactive; RBC, Reinforced Brick Concrete; RCC, Reinforced Cement Concrete; SSU, Secondary Sampling Unit; TSU, Tertiary Sampling Unit; UID No, Unique Identification Number.

[*] Corresponding author.
E-mail address: arun_Kokane.cfm@aiimsbhopal.edu.in (A.M. Kokane).

[1] Equal contribution.

https://doi.org/10.1016/j.dib.2021.107169

Keywords:
Coronavirus
COVID-19
Epidemiology
SARS-CoV-2
Disease outbreaks
Disease transmission
Infectious

Information on socio-demographic and serological variables were collected from 4,883 participants using a multi-stage stratified random sampling method. Appropriate weightage was calculated for each participant as sampling fraction derived from Primary Sampling Unit (PSU), Secondary Sampling Unit (SSU) and Tertiary Sampling Unit (TSU). The weightage was then applied to the data to adjust the findings at population level. The comprehensive and robust methodology employed here may act as a model for similar future endeavours. At the same time, the dataset can also be relevant for researchers in fields such as data science, epidemiology, virology and earth modelling.

Specifications Table

Subject	Health and medical sciences
Specific subject area	Epidemiology, Infectious diseases, Public health and health policy
Type of data	Table Graph Map
How data were acquired	Survey questionnaire (attached in supplementary material), venous blood sample, Electrochemiluminescence Immunoassay (ECLIA) technique
Data format	Raw analysed
Parameters for data collection	Data was collected from all the non-institutionalised residents of the city aged one year and above (i.e., excluding prisons, hospitals, old age homes, orphanages, etc.).
Description of data collection	Data was acquired through a multi-stage stratified cluster random sampling method from 24th August to 5th September 2020. There are 54 geo-administrative units (henceforth referred to as wards) in the study area. These 54 wards were organised according to decreasing COVID-19 positivity per thousand (information sourced from passive surveillance reports obtained from the city administration and National Health Mission, Ujjain) and then divided into three equal tertiles – 18 in High Burden tertile (HB), 18 in Intermediate Burden tertile (IB) and 18 in Low Burden tertile (LB). Primary Sampling Units (PSUs) were nested within each administrative unit. These tertiles cumulatively yielded 100 Primary Sampling Units – 33 from High burden, 33 from Intermediate burden, and 34 from Low burden tertile. Sampled households within each PSU were treated as Secondary Sampling Units (SSUs), and each participant comprised the Tertiary Sampling Unit (TSU). Data was collected in-person from each of the participant by a two-member team on mobile/tablet-based application. The data was subsequently uploaded onto a cloud-based data collection and compilation platform and downloaded in .csv format once data collection was over. A sample of venous blood was also collected, which was analysed for the presence of anti-SARS-CoV-2 antibody using the Electrochemiluminescence Immunoassay (ECLIA) technique. The data thus generated were merged together after being matched with the help of unique ID number generated for each participant.
Data source location	Online at: http://dx.doi.org/10.17632/s5c5ztwdvd.1; Data was acquired from Ujjain city, Madhya Pradesh, India
Data accessibility	Repository name: Mendeley Data repository: http://dx.doi.org/10.17632/s5c5ztwdvd.1

Value of the Data

- This dataset reveals clustered heterogenous transmission patterns of SARS-CoV-2 in the central Indian city of Ujjain, which potentially reflects transmission patterns of other similar cities.
- This dataset can be of benefit to researchers, urban health managers, civic administration and policy makers for designing control and containment strategies for SARS-CoV-2 and similar airborne diseases.
- Data might be used/reused to understand COVID-19 transmission trajectories and to explore the attributes of heterogeneity.

1. Data Description

With the spread of SARS-CoV-2 pandemic, the Indian government initiated travel /movement restrictions and containment strategies in the form of a nationwide lockdown in order to reduce interpersonal transmission [1]. The eventual relaxation of mass quarantine measures [2] coincided with an increase in the average daily positivity rate, i.e. the proportion of COVID-19 laboratory tests that are reported to be positive [3]. Since a major proportion of COVID-19 cases remains clinically asymptomatic while still being potentially infectious, there is a need to accurately quantify the magnitude of spread of infection in the population through a sero-epidemiological survey. This article presents data from one such sero-epidemiological survey, conducted in the city of Ujjain in Madhya Pradesh, India. The supplementary datasheet (in Microsoft Excel format) consists of 40 columns (variables) and 4883 rows (observations). The variable names and description of the variables is shown in Table 1.

Our data is divided into two parts – socio-demographic data, collected with the help of a questionnaire, and serological data on anti-SARS-CoV-2 antibody (tabulated in Table 1).

Table 2 presents initial unadjusted findings from the data. 2746 or 56.3% of the participants were women; 613 or 12.5% were aged <15 years, 1570 or 32.2% were aged 15 – 30 years, 1359 or 27.8% were aged 30 – 45 years, 882 or 18.1% were aged 45 – 60 years, and 459 or 9.4% were above 60 years in age. 1667 or 34.1% of the participants were from the High Burden tertile, 1451 or 29.7% from the Intermediate Burden tertile, and 1765 or 36.2% from the Low Burden tertile.

Overall unadjusted seroprevalence for SARS-CoV-2 was found to be 14.2% (95% CI: 13.2% - 15.2%). Unadjusted estimates of seroprevalence from high burden, intermediate burden, and low burden tertiles showed 326 participants from high burden tertile (19.6% of total participants from high burden tertile), 141 participants from intermediate burden tertile (9.7% of total participants from intermediate burden tertile) and 224 participants from the low burden tertile (12.7% of total participants from low burden tertile) were seropositive for anti-SARS-CoV-2 antibody. The unadjusted prevalence of seropositivity for anti-SARS-CoV-2 antibody was 16.7% in males, and 12.2% in females.

Table 3 depicts the distribution of the participants according to occupation, whether the participants work as essential services providers, type of family, and residence in containment areas. 23.9% of the seropositive participants were residents of containment areas, as compared to 72.8% who were not. 86.5% of participants were from nuclear families while 14.1% were from joint families.

Tables 4-6 depict the unadjusted ward-wise seropositivity in the three tertiles – high burden, intermediate burden and low burden.

Table 7 and Fig. 1 provides information on adjusted seroprevalence. The overall adjusted seroprevalence was found to be 13.9% (95% CI: 10.4% - 18%). Adjusted seroprevalence across the three tertiles was 18.6%, 10.5% and 13.6% in the HB, IB and LB tertiles respectively. Adjusted seroprevalence was found to be 16.5% in males, as compared to 11.7% in females. amongst age groups, the adjusted seroprevalence was highest in the 30 – 45 years age group (17.1%), followed by 45 – 60 years age group (16.7%), and was lowest in the youngest - <15 years (9.5%).

Table 1
Description of data variables.

Variable name	Type of data	Description
Ward	Socio-demographic	Geo-administrative unit number (ward number)
Tertile_cat	Socio-demographic	Ward categorised as per reported number of cases per 1000 population into three categories: (HB: High burden; IB: Intermediate burden; LB: Low burden)
Cluster	Socio-demographic	Unique cluster number assigned to each Primary Sampling Unit (PSU). The Primary Sampling Unit (PSU) was operationally defined by the geographical boundaries of nested colonies in each ward.
Hhn	Socio-demographic	Unique identification number assigned to each secondary sampling unit, i.e. Household in a particular cluster
Hhn_serial	Socio-demographic	Tertiary Sampling Unit (TSU) sequence in each Hhn. Each Hhn_serial corresponds to a participant selected from the Hhn.
UID_no	Socio-demographic	A concatenated unique ID number for each participant derived from sequential arrangement of Ward/PSU/SSU/TSU.
Age	Socio-demographic	Age (in completed years)
Sex	Socio-demographic	Sex of the participant
Education	Socio-demographic	Educational qualification (categorised into Illiterate, Primary school, Middle school, High school, Intermediate, Graduate and above)
Occupation	Socio-demographic	Specific occupation of the participant
Cat_occupation	Socio-demographic	The category of occupation as determined by interviewer (categorised into Unemployed, Unskilled, Semi-skilled, Skilled, Clerk/ Shop-keeper/ Farmer, Semi-professional, Professional)
Essential_service	Socio-demographic	Whether employed in the essential services as enlisted in section 2(1) in the Essential Services Maintenance Act, 1968, Republic of India.
Family_type	Socio-demographic	Type of family (nuclear or joint) where family is defined as biologically or legally related individuals sharing the same kitchen.
Male_family	Socio-demographic	Number of males in the family
Female_family	Socio-demographic	Number of females in the family
Children_family	Socio-demographic	Number of children (1–18 years) in the family
Adult_family	Socio-demographic	Number of adults (>18 years) in the family
BPL_card_holder	Socio-demographic	Whether family possesses a BPL card (BPL cards are a specific kind of ration cards which permit the families with annual incomes of less than Rs. 10,000 to utilise the benefits of Targeted Public Distribution System and procure essential commodities like rice, wheat, sugar, kerosene, fertilizers, LPG, etc. to its citizens at highly subsidized prices. Here it is used as a proxy indicator of poverty).
House_type	Socio-demographic	The type of house construction (categorised into Kaccha, Pucca, and mixed houses. Kaccha house is one where the wall and/or roof is made of temporary materials such as unburnt bricks, bamboos, mud, grass, reeds, thatch, loosely packed stones, etc. In pucca houses the walls are made of burnt bricks, stones (packed with lime or cement), cement concrete, timber, etc., while the roof is made of Tiles, GCI (Galvanised Corrugated Iron) sheets, asbestos cement sheet, RBC, (Reinforced Brick Concrete), RCC (Reinforced Cement Concrete) and timber etc. In mixed houses, walls are made up of pucca material but roof is made up of the material other than those used for pucca house.)
Housing_location	Socio-demographic	Location of the household (apartment building, independent house, housing society, others)

(*continued on next page*)

Table 1 (*continued*)

Variable name	Type of data	Description
Total_rooms	Socio-demographic	Total number of rooms in the household
Containment_area_resident	Socio-demographic	Whether resident of a containment zone (containments zones are zones of restricted mobility, and have a higher reported positivity rate)
Tested_previously_COVID_19	Socio-demographic	Whether tested previously for COVID-19
Test_results	Socio-demographic	Result of the COVID-19 test (negative, positive, not applicable, cannot tell or do not know) (condition on the participant being tested for COVID-19)
Reason_testing/Smptmtc	Socio-demographic	Reason for getting tested – participant was symptomatic (condition on the participant being tested for COVID-19)
Reason_testing/Hst_cntct	Socio-demographic	Reason for getting tested – history of contact (condition on the participant being tested for COVID-19)
Reason_testing/Migrant	Socio-demographic	Reason for getting tested – migration from other state within India (condition on the participant being tested for COVID-19)
Reason_testing/Hst_travel	Socio-demographic	Reason for getting tested – history of international travel (condition on the participant being tested for COVID-19)
Reason_testing/Other	Socio-demographic	Reason for getting tested – other (condition on the participant being tested for COVID-19)
Travel_history	Socio-demographic	History of travel outside the city in the past 4 months (in between April – July 2020)
Arogya_Setu_App_prsnt	Socio-demographic	Arogya Setu application installed in the participant's phone. This is a mobile-device based software application which aides in contact tracing and self-assessment and provides the user with exposure status with respect to SARS-CoV-2.
Current_Arogya_Setu_status	Socio-demographic	Current exposure status of the participant based on Arogya Setu information (Safe, Low, Moderate) (conditional variable)
Geolocate_GPS	Socio-demographic	Geospatial coordinates of household (SSU)
_Geolocate_GPS_latitude	Socio-demographic	Latitude of the household (SSU)
_Geolocate_GPS_longitude	Socio-demographic	Longitude of the household (SSU)
_Geolocate_GPS_altitude	Socio-demographic	Altitude of the household (in metres) above the sea surface level (SSU)
_Geolocate_GPS_precision	Socio-demographic	Accuracy of the geolocation (in metres)
titre_value	Serological	ECLIA derived Cycle Threshold (CT) titre values of anti-SARS-CoV-2 antibody
Result	Serological	SARS-CoV-2 antibody test results (R: reactive, NR: non-reactive). 200μL serum was run in the automated instrument using the pre-defined "ECOV2" program. Following sample initialization, test values were obtained in numerical format at intervals of one minute. The analyser automatically determined a cut-off value based on the measurement of signals generated from the 2 calibrators provided by the manufacturer. Laboratory results were interpreted as "Reactive" and "Non-reactive" from the Cut-off Index (COI), defined as the ratio between the signal intensity of the unknown sample and the cut-off value. The COI value of $\geq$ 1.0 was taken as indicative of reactivity, as specified by the manufacturer.
Weight	Derived	Weightage applied to each participant

Fig. 2 depicts the adjusted tertile-wise antibody titre value for anti-SARS-CoV-2 antibody. Adjusted titre values were found to be 10.4 COI (SE = 3.38 COI) for the HB tertile; 4.8 (SE = 1.34 COI) for the IB tertile; and 6.1 COI (SE = 2.19 COI) for the LB tertile.

Fig. 3 depicts findings from density analysis of anti-SARS-CoV-2 antibody titres amongst those found seropositive in the three tertiles.

Fig. 4 depicts the ward-wise adjusted seropositivity in a choropleth map.

Table 2
Distribution of the participants according to age, sex, and burden tertile and seropositivity for SARS-CoV-2.

Characteristic	Overall, N (%)	Non-Reactive, N (%)	Reactive, N (%)
Overall	4883 (100)	4192 (85.8)	691 (14.2)
Tertile category			
H	1667 (100)	1341 (80.4)	326 (19.6)
I	1451 (100)	1310 (90.3)	141 (9.7)
L	1765 (100)	1541 (87.3)	224 (12.7)
Gender			
Female	2746 (100)	2412 (87.8)	334 (12.2)
Male	2137 (100)	1780 (83.3)	357 (16.7)
Age groups			
<15years	613 (100)	555 (90.5)	58 (9.5)
>60years	459 (100)	413 (90.0)	46 (10.0)
15–30years	1570 (100)	1349 (85.9)	221 (14.1)
30–45years	1359 (100)	1134 (83.4)	225 (16.6)
45–60years	882 (100)	741 (84.0)	141 (16.0)

Table 3
Distribution of the participants according to occupation, provision of essential services, type of family, and residence in containment areas.

Characteristic	Overall (%)	Seronegative (%)	Seropositive (%)
Occupational Categories	4883 (100)	4192 (100)	691 (100)
Unemployed	2678 (54.8)	2375 (56.7)	303 (43.8)
Unskilled	648 (13.3)	498 (11.9)	150 (21.7)
Semi-skilled	266 (5.4)	231 (5.5)	35 (5.1)
Skilled	587 (12.0)	487 (11.6)	100 (14.5)
Clerk/ shopkeeper/ farmer	361(7.4)	295 (7.0)	66 (9.6)
Semi-professional	179 (3.7)	160 (3.8)	19 (2.7)
Professional	164 (3.4)	146 (3.5)	18 (2.6)
Essential services providers	4883 (100)	4192 (100)	691 (100)
Yes	86 (1.8)	54 (1.3)	32 (4.6)
No	4635 (94.9)	3996 (95.3)	639 (92.5)
Cannot tell	162 (3.3)	142 (3.4)	20 (2.9)
Type of family	4883 (100)	4192 (100)	691 (100)
Joint	658 (13.5)	593 (14.1)	65 (9.4)
Nuclear	4225 (86.5)	3599 (85.9)	626 (90.6)
Resident of containment areas	4883 (100)	4192 (100)	691 (100)
Yes	427 (8.7)	262 (6.3)	165 (23.9)
No	4226 (86.6)	3723 (88.8)	503 (72.8)
Cannot tell	230 (4.7)	207 (4.9)	23 (3.3)

2. Experimental Design, Materials and Methods

2.1. Sample size and cluster number calculation

2.1.1. COVID-19 positivity estimates of administrative units

The COVID-19 positivity estimates of the geo-administrative units (henceforth referred to as wards) were obtained from the city administration and National Health Mission, Ujjain. The administrative units were then divided according to the gradient of reported positivity per 1000 population as shown in Fig. 5.

$$Positivity\ Rate\ for\ Ward = \frac{Number\ of\ SARS-CoV-2\ positive\ cases\ in\ the\ ward}{Total\ population\ of\ the\ ward} \times 1000$$

Table 8 and Fig. 6 represents the clubbing of the wards into three tertiles based on seropositivity rate per 1000 population. The wards were arranged in a descending order based on their positivity rate, and were arbitrarily divided into High Burden (HB), Intermediate Burden (IB) and Low Burden (LB) tertiles, such that each tertile had 18 wards.

Table 4
Seropositivity in wards of high burden tertile.

Ward no	Total participants (%)	Non-positive (%)	Seropositive (%)
Total	1667 (100)	1341 (80.0)	326 (20)
7	95 (100)	89 (94.0)	6 (6.3)
8	106 (100)	85 (80.0)	21 (20)
9	49 (100)	35 (71.0)	14 (29)
11	163 (100)	53 (33.0)	110 (67)
14	84 (100)	75 (89.0)	9 (11)
15	108 (100)	96 (89.0)	12 (11)
20	105 (100)	83 (79.0)	22 (21)
23	53 (100)	44 (83.0)	9 (17)
25	55 (100)	53 (96.4)	2 (3.6)
26	107 (100)	63 (59.0)	44 (41)
27	26 (100)	11 (42.0)	15 (58)
28	38 (100)	25 (66.0)	13 (34)
29	95 (100)	78 (82.0)	17 (18)
33	109 (100)	93 (85.0)	16 (15)
37	56 (100)	53 (94.6)	3 (5.4)
38	111 (100)	108 (97.3)	3 (2.7)
48	152 (100)	149 (98.0)	3 (2.0)
51	155 (100)	148 (95.5)	7 (4.5)

Table 5
Seropositivity in wards of intermediate burden tertile.

Ward no	Total participants (%)	Non-positive (%)	Seropositive (%)
Total	1451 (100)	1310 (90.0)	141 (10.0)
1	99 (100)	89 (90.0)	10 (10.0)
2	102 (100)	91 (89.0)	11 (11.0)
16	98 (100)	85 (87.0)	13 (13.0)
17	1 (100)	1 (100.0)	0 (0.0)
18	93 (100)	81 (87.0)	12 (13.0)
21	44 (100)	36 (82.0)	8 (18.0)
24	91 (100)	79 (87.0)	12 (13.0)
34	86 (100)	68 (79.0)	18 (21.0)
35	124 (100)	102 (82.0)	22 (18.0)
39	40 (100)	39 (97.5)	1 (2.5)
42	82 (100)	76 (92.7)	6 (7.3)
43	51 (100)	41 (80.0)	10 (20.0)
45	108 (100)	103 (95.4)	5 (4.6)
46	112 (100)	110 (98.2)	2 (1.8)
52	106 (100)	102 (96.2)	4 (3.8)
53	107 (100)	103 (96.3)	4 (3.7)
54	107 (100)	104 (97.2)	3 (2.8)

2.1.2. *Design effect calculation, sample size calculation and cluster size and number estimation*

Since we were sampling via multi-stage stratified cluster random sampling, we needed to calculate design effect to account for the increased variance expected with cluster random sampling as opposed to simple random sampling. We presumed a mean unadjusted prevalence of sero-positivity for anti-SARS-CoV-2 antibody to be 5%, with a standard deviation of 1%. The inter-cluster variation (also called intra-class cluster coefficient, ICC or ρ) was determined to be 0.20. We decided to sample 50 participants per cluster. Accordingly, we calculated the design effect based on the formula:

$$\mathrm{DEFF} = 1 + \rho(\mathrm{ppC} - 1)$$

Where:

DEFF = Design Effect
ppC = Persons per cluster (here 50)

Table 6
Seropositivity in wards of low burden tertile.

Ward no	Total participants (%)	Non-positive (%)	Seropositive (%)
Total	1765 (100)	1541 (87.0)	224 (13.0)
3	108 (100)	97 (90.0)	11 (10.0)
4	108 (100)	106 (98.1)	2 (1.9)
5	120 (100)	109 (90.8)	11 (9.2)
6	95 (100)	85 (89.0)	10 (11.0)
10	55 (100)	48 (87.0)	7 (13.0)
12	110 (100)	105 (95.5)	5 (4.5)
13	107 (100)	68 (64.0)	39 (36.0)
19	102 (100)	95 (93.1)	7 (6.9)
22	54 (100)	46 (85)	8 (15)
30	82 (100)	45 (55.0)	37 (45.0)
31	55 (100)	29 (53.0)	26 (47.0)
32	55 (100)	32 (58.0)	23 (42.0)
36	59 (100)	58 (98.3)	1 (1.7)
40	161 (100)	158 (98.1)	3 (1.9)
41	108 (100)	99 (91.7)	9 (8.3)
47	171 (100)	160 (93.6)	11 (6.4)
49	107 (100)	103 (96.3)	4 (3.7)
50	108 (100)	98 (91.7)	10 (9.3)

Table 7
Adjusted seroprevalence rates according to tertile, gender and age group.

Characteristic	Population	Projected sero-prevalence (%)
Tertile category		
High	158,172	29,370 (18.6)
Intermediate	196,467	20,564 (10.5)
Low	2,019,417	28,580 (13.6)
Gender		
Male	255,506	42,252 (16.5)
Female	308,550	36,262 (11.7)
Age Category		
<15 years	71,171	6783 (9.5)
15 – 30 years	183,093	22,375 (12.2)
30 – 45 years	156,806	26,808 (17.1)
45 – 60 years	101,516	16,965 (16.7)
>60 years	51,472	5583 (10.8)

ρ = Intra-class cluster coefficient (here 0.2)

As seen in Table 9, for multi-stage stratified cluster sampling with size of each cluster taken to be fifty (50), the DEFF was derived to be 10.8.

Base sample size for the study was estimated using the formula

$$n = \frac{\left(z_{1-\alpha/2}\right)^2 * p * (1-p)}{d^2}$$

Where:

$Z_{1\text{-}\alpha/2}$ = is the standard normal variate; at a 5% standard error (i.e. p-value of 0.05), it was estimated to be 1.96
p = prevalence of the health condition, here positive test for SARS-CoV-2, assumed at 5%
d = absolute precision, here taken to be 2%

Using this formula, we arrived at a base sample size of 457.

Since the design effect was calculated to be 10.8, the corrected sample size for the purpose of our study was calculated by multiplying the base sample size with the design effect.

Table 8
Ward populations and cases per 1000 population.

Ward number	Ward population	Total number of cases	Cases per 1000 population
High burden tertile (cut-off = 2.04 per 1000 population)			
Ward 28	7861	81	10.30
Ward 27	8966	60	6.69
Ward 29	9853	52	5.28
Ward 20	8024	41	5.11
Ward 7	11,460	50	4.36
Ward 8	8356	33	3.95
Ward 33	10,456	38	3.63
Ward 25	6953	25	3.60
Ward 26	9012	32	3.55
Ward 23	8921	28	3.14
Ward 38	6145	19	3.09
Ward 14	10,124	31	3.06
Ward 9	7921	24	3.03
Ward 48	17,265	51	2.95
Ward 15	9945	28	2.82
Ward 11	12,350	34	2.75
Ward 51	14,021	33	2.35
Ward 37	7345	15	2.04
Intermediate burden tertile (cut-off = 0.89 per 1000 population)			
Ward 21	8860	18	2.03
Ward 44	7825	13	1.66
Ward 53	16,512	27	1.64
Ward 1	12,240	19	1.55
Ward 16	9125	14	1.53
Ward 43	11,648	17	1.46
Ward 18	8912	13	1.46
Ward 39	9903	13	1.31
Ward 35	15,874	19	1.20
Ward 24	7621	8	1.05
Ward 46	10,523	11	1.05
Ward 17	16,250	16	0.98
Ward 2	12,356	12	0.97
Ward 52	9645	9	0.93
Ward 34	16,245	15	0.92
Ward 42	9904	9	0.91
Ward 45	10,145	9	0.89
Ward 54	15,791	14	0.89
Low burden tertile (rest)			
Ward 49	10,190	9	0.88
Ward 4	16,245	14	0.86
Ward 19	11,618	10	0.86
Ward 36	9680	7	0.72
Ward 47	15,489	11	0.71
Ward 22	7246	5	0.69
Ward 50	12,450	8	0.64
Ward 3	14,021	9	0.64
Ward 30	11,201	7	0.62
Ward 31	10,532	6	0.57
Ward 6	12,160	6	0.49
Ward 13	11,045	5	0.45
Ward 32	9251	4	0.43
Ward 5	9680	4	0.41
Ward 10	10,411	4	0.38
Ward 40	16,489	3	0.18
Ward 41	12,125	2	0.16
Ward 12	13,459	2	0.15

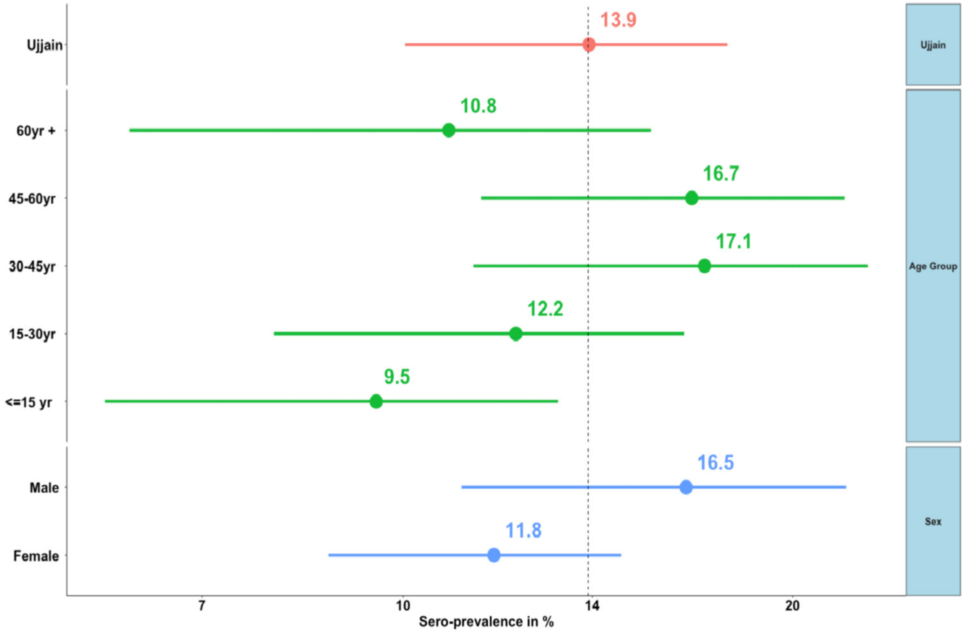


Fig. 1. Adjusted seroprevalence (in percentages) according to gender and age group.

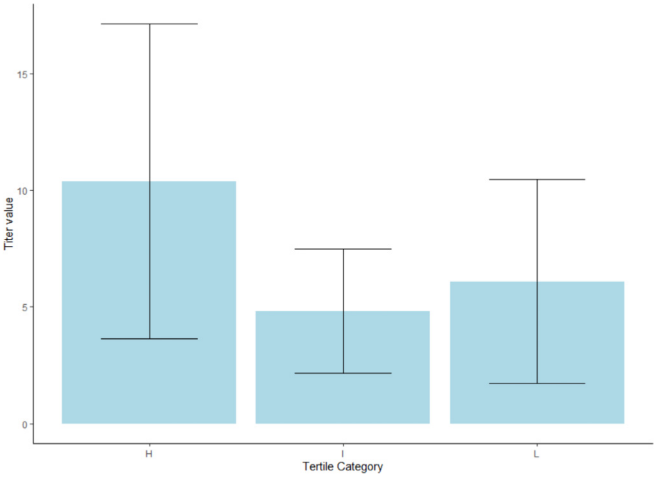


Fig. 2. Adjusted titre values for of anti-SARS-CoV-2 antibody according to burden tertiles.

The corrected sample size (adjusted for clustering) was calculated using the formula:

$$N = n * DEFF$$

Where:

N = Corrected sample size
n = Base sample size (here 457)
DEFF = Design Effect (here 10.8)

The corrected sample size therefore was calculated to be 4936, rounded off to 5000.

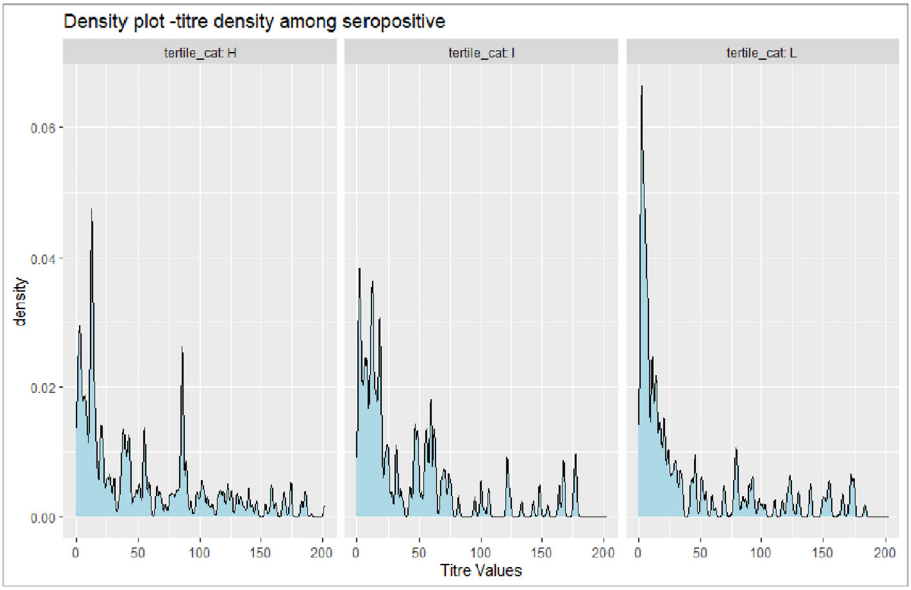


Fig. 3. Density plot of antibody tires amongst seropositive individuals according to tertile categories (tertile_cat:H = High burden tertile; tertile_cat:I = Intermediate burden tertile; tertile_cat:L = Low burden tertile).

Table 9
Calculation of design effect.

ρ (ICC)	Person per Cluster (ppC)	Deff =1+(ppC − 1) *ρ(when ρ is constant)
0.2	10	2.8
0.2	20	4.8
0.2	30	6.8
0.2	40	8.8
0.2	50	10.8
0.2	60	12.8
0.2	70	14.8
0.2	80	16.8
0.2	90	18.8
0.2	110	22.8
0.2	120	24.8

Since we had decided to collect a sample of 50 participants per cluster, it thus was surmised that we would need to collect data from 5000/50, i.e. 100 clusters.

2.1.3. Calculation of number of clusters needed per administrative unit

The administrative units (wards) were arranged in a descending fashion according to the calculated positivity rate. The clusters were nested within each ward, and were called colonies. For obtaining the number of colonies to be sampled from each ward, we cumulated the populations of each ward and then divided the result with the serial interval. The outcomes of the division were taken as the number of clusters to be collected per ward (Figs 7, 8, and 9).

2.1.4. Selection of colonies from each ward

A list of all the colonies from the wards was obtained from the Collector's Office. For each ward, random numbers were generated based on the total number of colonies in the ward, and the colonies corresponding to the generated numbers were chosen. So, for a ward with seven

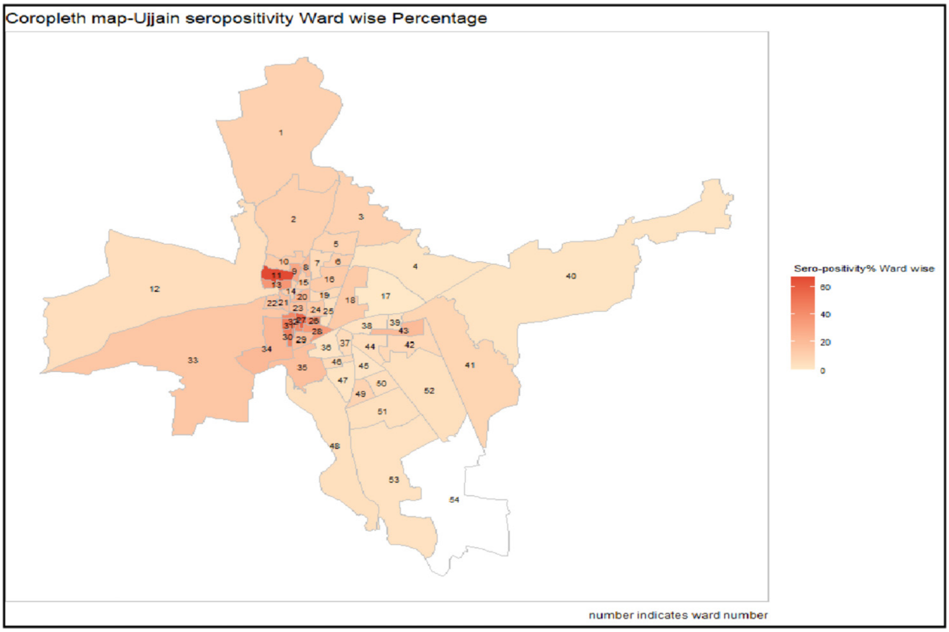

Coropleth map-Ujjain seropositivity Ward wise Percentage
Sero-positivity% Ward wise
60
40
20
0
number indicates ward number
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54

Fig. 4. Choropleth maps of Ujjain representing ward-wise adjusted seropositivity.

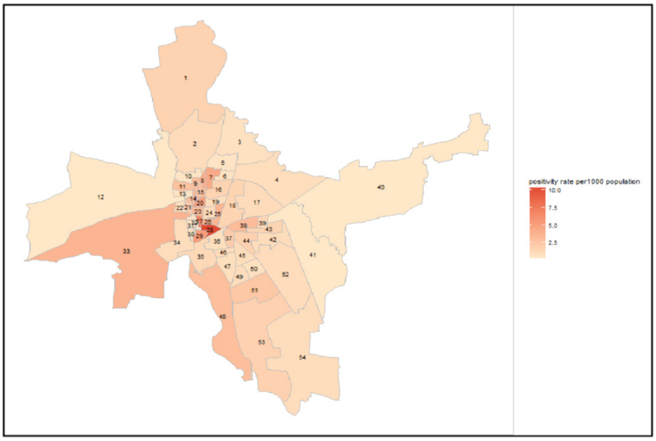


Fig. 5. Reported test positivity per 1000 population for each ward (geo-administrative unit).

colonies if three colonies were to be chosen, three random numbers out of a possible seven were generated. Subsequently, the corresponding colonies were chosen for our study.

2.1.5. Selection of households

Each colony was assigned a random number, which corresponded to a fixed cardinal direction (North, South, East or West).

On the day of data collection, the data collection team reached the centre of the allotted colony and approached the first household in the pre-assigned randomly allotted direction. After data from the first enroled household was collected, all the subsequent households were chosen in the same direction till data from fifty participants was collected.

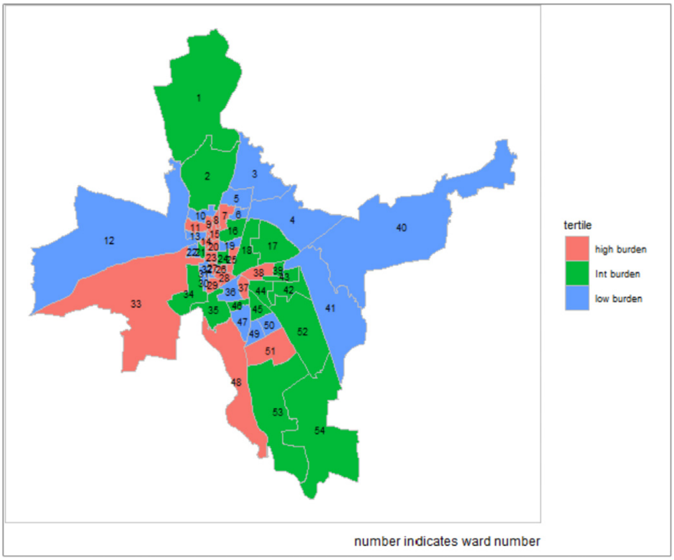
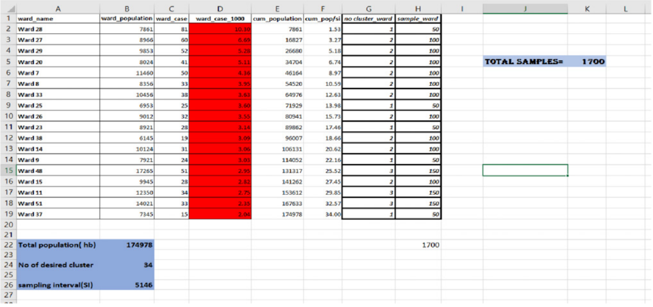
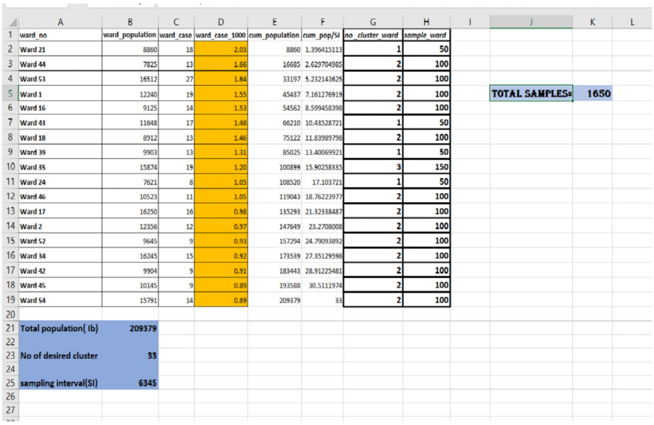


Fig. 6. Categorisation of wards according to reported test positivity for SARS-CoV-2 per 1000 population into three groups – High burden, Intermediate burden and Low burden.



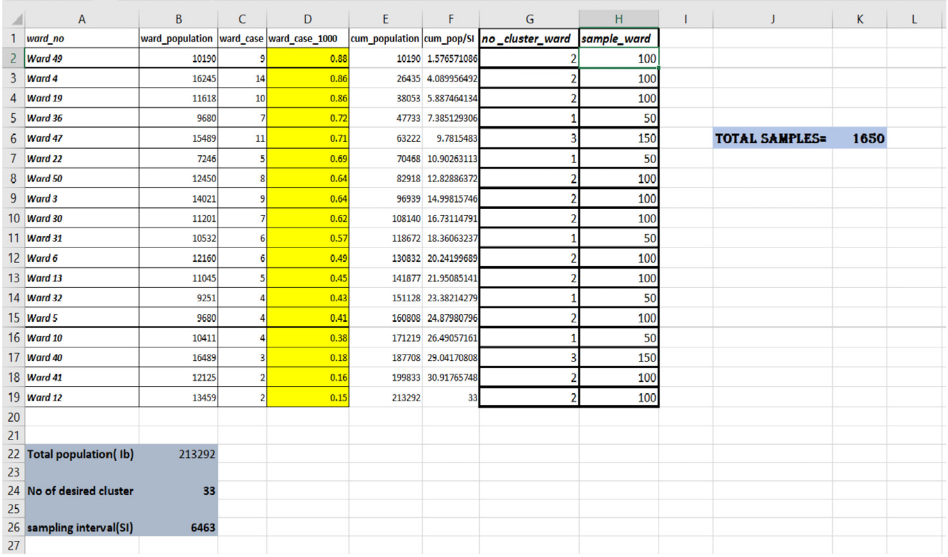
	A	B	C	D	E	F	G	H	I	J	K	L
1	ward_name	ward_population	ward_case	ward_case_1000	cum_population	cum_pop/si	no cluster_ward	sample_ward				
2	Ward 28	7861	81	10.30	7861	1.53	1	50				
3	Ward 27	8966	60	6.69	16827	3.27	2	100				
4	Ward 29	9853	52	5.28	26680	5.18	2	100				
5	Ward 20	8024	41	5.11	34704	6.74	2	100		TOTAL SAMPLES=	1700	
6	Ward 7	11460	50	4.36	46164	8.97	2	100				
7	Ward 8	8356	33	3.95	54520	10.59	2	100				
8	Ward 33	10456	38	3.63	64976	12.63	2	100				
9	Ward 25	6953	25	3.60	71929	13.98	1	50				
10	Ward 26	9012	32	3.55	80941	15.73	2	100				
11	Ward 23	8921	28	3.14	89862	17.46	1	50				
12	Ward 38	6145	19	3.09	96007	18.66	2	100				
13	Ward 14	10124	31	3.06	106131	20.62	2	100				
14	Ward 9	7921	24	3.03	114052	22.16	1	50				
15	Ward 48	17265	51	2.95	131317	25.52	3	150				
16	Ward 15	9945	28	2.82	141262	27.45	2	100				
17	Ward 11	12350	34	2.75	153612	29.85	3	150				
18	Ward 51	14021	33	2.35	167633	32.57	3	150				
19	Ward 37	7345	15	2.04	174978	34.00	1	50				
20												
21												
22	Total population(hb)	174978						1700				
23												
24	No of desired cluster	34										
25												
26	sampling interval(SI)	5146										
27												

Fig. 7. Estimation of number of clusters needed from high burden tertile.



	A	B	C	D	E	F	G	H	I	J	K	L
1	ward_no	ward_population	ward_case	ward_case_1000	cum_population	cum_pop/SI	no_cluster_ward	sample_ward				
2	Ward 21	8860	18	2.03	8860	1.396415113	1	50				
3	Ward 44	7825	13	1.66	16685	2.629704985	2	100				
4	Ward 53	16512	27	1.64	33197	5.232143625	2	100				
5	Ward 1	12240	19	1.55	45437	7.161276919	2	100		TOTAL SAMPLES=	1650	
6	Ward 16	9125	14	1.53	54562	8.599458398	2	100				
7	Ward 43	11648	17	1.46	66210	10.43528721	1	50				
8	Ward 18	8912	13	1.46	75122	11.83989798	2	100				
9	Ward 39	9903	13	1.31	85025	13.40069921	1	50				
10	Ward 35	15874	19	1.20	100899	15.90258335	3	150				
11	Ward 24	7621	8	1.05	108520	17.103721	1	50				
12	Ward 46	10523	11	1.05	119043	18.76223977	2	100				
13	Ward 17	16250	16	0.98	135293	21.32338487	2	100				
14	Ward 2	12356	12	0.97	147649	23.2708008	2	100				
15	Ward 52	9645	9	0.93	157294	24.79093892	2	100				
16	Ward 34	16245	15	0.92	173539	27.35129598	2	100				
17	Ward 42	9904	9	0.91	183443	28.91225481	2	100				
18	Ward 45	10145	9	0.89	193588	30.5111974	2	100				
19	Ward 54	15791	14	0.89	209379	33	2	100				
20												
21	Total population(lb)	209379										
22												
23	No of desired cluster	33										
24												
25	sampling interval(SI)	6345										
26												
27												

Fig. 8. Estimation of number of clusters needed from intermediate burden tertile.



	A	B	C	D	E	F	G	H	I	J	K	L
1	*ward_no*	ward_population	ward_case	ward_case_1000	cum_population	cum_pop/SI	***no _cluster_ward***	***sample_ward***				
2	***Ward 49***	10190	9	0.88	10190	1.576571086	2	100				
3	***Ward 4***	16245	14	0.86	26435	4.089956492	2	100				
4	***Ward 19***	11618	10	0.86	38053	5.887464134	2	100				
5	***Ward 36***	9680	7	0.72	47733	7.385129306	1	50				
6	***Ward 47***	15489	11	0.71	63222	9.7815483	3	150		**TOTAL SAMPLES=**	**1650**	
7	***Ward 22***	7246	5	0.69	70468	10.90263113	1	50				
8	***Ward 50***	12450	8	0.64	82918	12.82886372	2	100				
9	***Ward 3***	14021	9	0.64	96939	14.99815746	2	100				
10	***Ward 30***	11201	7	0.62	108140	16.73114791	2	100				
11	***Ward 31***	10532	6	0.57	118672	18.36063237	1	50				
12	***Ward 6***	12160	6	0.49	130832	20.24199689	2	100				
13	***Ward 13***	11045	5	0.45	141877	21.95085141	2	100				
14	***Ward 32***	9251	4	0.43	151128	23.38214279	1	50				
15	***Ward 5***	9680	4	0.41	160808	24.87980796	2	100				
16	***Ward 10***	10411	4	0.38	171219	26.49057161	1	50				
17	***Ward 40***	16489	3	0.18	187708	29.04170808	3	150				
18	***Ward 41***	12125	2	0.16	199833	30.91765748	2	100				
19	***Ward 12***	13459	2	0.15	213292	33	2	100				
20												
21												
22	**Total population(Ib)**	213292										
23												
24	**No of desired cluster**	**33**										
25												
26	**sampling interval(SI)**	**6463**										
27												

Fig. 9. Estimation of number of clusters needed from low burden tertile.

2.1.6. Selection of participants

Every individual in the selected household was invited to participate in the study, given that they met the inclusion criteria mentioned earlier. Prior to enrolment in the study, the participants were requested to sign an informed consent form, written in the vernacular language (here: Hindi). In case the participant was below 18 years of age, their assent in addition to consent from their parents/legal guardians was sought. In case the participant was less than 7 years in age, consent was sought from their parents/legal guardians.

2.1.7. Data collection from participants

Two types of data were collected from the selected and consenting participants – socio-demographic and clinical profile, and venous blood sample.

The questionnaire used for data collection was divided into parts – general information, socio-demographic details, medical information, and record of geolocation.

Data collection commenced on the 24th of August 2020, and continued till the 5th of September 2020. For the purpose of data collection, 30 teams were devised, and each team was given roughly 3 colonies for data collection.

2.1.8. Composition of data collection team

The data collection team was composed of two members – an Auxiliary Nurse Midwife (ANM), and a Multi-Purpose Health Worker (MHW) or a doctor. The team members were extensively trained on the methodology of the study, and also had a day-long hands-on training session about data collection procedure using mobile-based application. The team was provided with the name of the colonies they were supposed to collect data from, with the total number of participants they needed to enrol from each colony, and the direction they are supposed to go in each colony.

2.2. Quality assurance mechanisms

Multiple steps were employed in order to ensure that the data collected is robust and representative of Ujjain.

2.2.1. Pilot testing data collection platform

Prior to deployment of the data collection platform in the field, multiple iterations of the same were pilot tested by an in-house team for trouble shooting and fool-proofing. Multiple measures and counter-measures were put in place in order to minimise mistakes and glitches in the field. For example, most questions were devised to be multiple choice to minimise the chances of error while typing in responses.

2.2.2. Pre-data collection training of data collection team members

Since data collection using tablet/mobile phone-based platforms is relatively new in India, care was taken to ensure that the entire process was accomplished smoothly. Prior to commencement of the data collection phase, the data collection teams were trained through a day-long hands-on session on how to use the platform for data collection and compilation.

2.2.3. Automatic generation of a unique ID based on demographic details

The data collection platform was engineered to generate a unique ID for each participant on the basis of collected demographic details. The team was instructed to label the blood sample vial with the same number. This ensured data rigour and minimised mix-ups and duplications.

2.2.4. Quality control (QC) for laboratory analysis

The instrument was calibrated while initiating every new reagent lot. We adopted a two-pronged strategy for maintaining the stringency of quality control throughout the laboratory analysis. Firstly, manufacturer-provided control materials (ACOV2 Cal1 and ACOV2 Cal2) were run at pre-determined intervals and the obtained values were compared with the acceptable range provided with the kit. Secondly, pooled negative and positive control material were generated in the laboratory following the protocol provided in the kit. For negative control, five non-reactive serum samples with COI values ≤150% of ACOV2 Cal1 were pooled. For positive control, three reactive samples with COI values more than ACOV2 Cal2 were pooled and diluted with Diluent MultiAssay (Roche®) to obtain COI value between 3 and 15. Aliquots of these in-house control materials were used for monitoring of analytical precision.

2.2.5. Refusal record

The team also maintained a record of the number of refusals from each colony. They did so by recording the age, sex and reason for refusal from every individual in the colony who refused to participate in the study.

3. Data and Sample Analysis

3.1. Questionnaire-based data (henceforth referred to as QBD)

Compilation of data

The data collection team members were instructed to synchronize their mobile phone/tablet-based applications every day after collection of data. This allowed for regular cloud compilation of data, which could then be downloaded for further analysis. It also ensured that the data collection process remained closely monitored throughout.

Cleaning of data

The data thus compiled was downloaded at the end of the data collection session and cleaned to remove any discrepancies. The master-chart was anonymised prior to analysis, and coded in order to make it easier to analyse.

Calculation of weights

Application of weights is critical for the purpose of calculating adjusted outcome variables. Weighing allows for adjustment of the observed data for inherent biases, and also makes it more representative of the population being sampled. It does so by more applying more weightage to data from more populous clusters (clusters which are more representative of the surveyed population), and less weightage to data from less populous clusters (clusters which are less representative of the surveyed population).

We calculated weights for every colony surveyed by taking into account:

1. The population of the ward
2. The number of colonies chosen from each ward for sampling purposes
3. The population of each colony
4. The median household size – thereby ascertaining the number of households in each colony (by dividing colony population by the number of households enroled in each colony)
5. Number of households selected (arrived at by dividing the number of samples needed per colony, viz. 50 by the median household size)

The calculation of weights was a step-wise process, and included:

1. Probability of household selection (by dividing the number of households selected in the study by the total number of households in the colony)
2. Colony based weightage, i.e. the weightage applicable to each participant vis-à-vis the total number of residents in the colony (the inverse of household selection probability), and finally,
3. Ward based weightage, i.e. the weightage assigned to each participant vis-à-vis the total number of residents in the colony (calculated by multiplying the proportion of colony population to ward population by colony-based weightage).

3.2. Laboratory-based data (henceforth referred to as LBD)

Collection of samples

Venous blood sample was collected from all willing participants after they had consented/assented for participating in the study and had responded to the questionnaire. Blood samples were collected from the study participants using proper aseptic precautions in sterile yellow-capped vacutainers (BD® India Pvt. Ltd.).

Storage and transportation of samples

At the end of each data collection round every day, all the vacutainers were collected, labelled with the participant's name, age, sex and the unique ID allotted to each participant, packed into separate temperature controlled and leak-proof boxes, and transported within 24 h to the reference laboratory (Department of Microbiology, AIIMS, Bhopal) for further analysis.

Statistical analysis

QBD was imported from a web-based data collection tool to R-global environment [18]. The data was duly checked for duplication, redundancies, missing values and outliers. The QBD dataset was further merged with LBD using common identifiers. The predetermined survey weight (described above) was assigned to each row (participant) according to their originating PSU. Data was again checked for possible missing data and discrepancies. Some of the interval variables like age were categorized into categories and appropriate class to variables as per R-environment were assigned.

The key socio-demographic characteristics of the participants were summarized by measures of central tendencies and dispersion as per the nature of variable.

Unadjusted and adjusted seropositivity at different geo-administrative strata (ward and city level), different socio-demographic strata and as per characteristics of interest (occupation in essential services, and residence in COVID-19 containment areas) were calculated through cross tabulations. All the variables of interest were estimated as point estimates and 95% confidence interval. The type 1 error was set at 0.5% for the analysis purpose. Bivariate analyses as found appropriate was conducted using chi square test. Choropleth maps were created by combining *.shp and *.dbf files. This file was further melted into a data frame in order to perform spatial manipulation in R environment using the tidyverse, jsonlite, lubridate, survey, ggplot2, and dplyr packages [19–22].

Adjusted analysis was performed with the help of 'survey' package and base R-software which is in public domain. Suitable visualizations were drawn with the help of ggplot 2 and base R.

Laboratory Analysis

Samples were transported to the laboratory within 24 h of sample collection, maintaining cold chain. Serum was separated from the blood samples by centrifugation at 3500 rpm for 10 min.

Serum was separated from the blood samples by centrifugation at 3500 rpm for 10 min and processed in COBAS e411 (Roche®) by using Elecsys Anti-SARS-CoV-2 Kit, as per manufacturer's instructions. 200µL serum was run in the automated instrument using the pre-defined "ECOV2" program. Following sample initialization, test values were obtained in numerical format at intervals of one minute. The analyser automatically determined a cut-off value based on the measurement of signals generated from the 2 calibrators provided by the manufacturer. Laboratory results were interpreted as "Reactive" and "Non-reactive" from the Cut-off Index (COI), defined as the ratio between the signal intensity of the unknown sample and the cut-off value. The COI value of $\geq$ 1.0 was taken as indicative of reactivity, as specified by the manufacturer.

Ethics Statement

The survey was conducted as per directive by the Government of Madhya Pradesh (Vide D.O. No. 219, Government of Madhya Pradesh, Medical Education Department, dated – 22nd July 2020) issued to understand the accurate magnitude of spread of SARS-CoV-2 in Ujjain city.

The directive mandated conducting this survey to aid policy-making related to setting up of containment zones in response to clustered spread of COVID-19. Consequently, this activity helped in assessment of seroprevalence of anti-SARS-CoV-2 antibodies by the Public Health Department. Since this activity was not conducted in the research mode, no Ethics approval was required for this study.

Informed consent from the participants was nonetheless obtained in English or vernacular (samples uploaded), whichever is applicable, wherein the participants were informed about the aim and purpose of the survey. In case the participant was below 18 years of age, their assent in addition to consent from their parents/legal guardians was sought. In case the participant was less than 7 years in age, consent was sought from their parents/legal guardians. Only those participants who were willing to take part and voluntarily provided informed consent/assent were included in the survey.

The participants' sero-status was shared with them while maintaining strict privacy. All the data generated was anonymized and all the individual identifiers were delinked before processing for analysis.

CRediT Author Statement

Ankur Joshi, Prem Shankar and Anirban Chatterjee: conception / design of the protocol; overall data management which includes development of data collection tool, coordinating real time data capture from various sites and aggregating data; data analysis / interpretation; drafting / critically reviewing the paper; giving approval for the final version to be published; **Prem Shankar, Jitendra Singh, Arti Shrivas and Anand Kumar Maury:** Laboratory sample processing, data interpretation; critically reviewing the paper; giving approval for the final version to be published; **Arun M. Kokane and Abhijit Pakhare:** Conception, supervision of data collection and data management, critically reviewing the paper; giving approval for the final version to be published; **Kriti Yadav and Raunaq Singh Nagi:** supervision of data collection and data management, drafting and critically reviewing the paper; giving approval for the final version to be published; **Debasis Biswas:** Overall data management, laboratory sample processing, data interpretation, critically reviewing the paper; giving approval for the final version to be published;

Sarman Singh: Overall supervision of the study; coordination, conception, critically reviewing the paper; giving approval for the final version to be published.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships which have or could be perceived to have influenced the work reported in this article.

Acknowledgments

Authors would like to thank to Madhya Pradesh Administration of Ujjain for providing support to collection and transportation of the samples to AIIMS, Bhopal (MP).

Funding statement and sponsorship: Authors would like to express sincere gratitude to the Government of Madhya Pradesh for providing the financial assistance to conduct this study (D.O No. 219 ACS, Govt. of MP dated 22.07.2020).

Supplementary Materials

Supplementary material associated with this article can be found in the online version at doi:10.1016/j.dib.2021.107169.

References

[1] The Lancet, India under COVID-19 Lockdown, Lancet 395 (2020) 1315.
[2] Ministry of Home Affairs, New Guidelines to fight COVID-19 to be effective from 1st June 2020, Press Inf. Bur., 3, 2020.
[3] S Thevar, Month After Unlock: Spike in Covid Cases, Deaths an Inevitability Pune Unable to Escape - Pune News - Hindustan Times, The Hindustan Times, 2020.