BMJ Open

Representative estimates of covid-19 infection fatality rates from four locations in India: descriptive study

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2021-050920
Article Type:	Original research
Date Submitted by the Author:	04-Mar-2021
Complete List of Authors:	Cai, Rebecca; Development Data Lab Novosad, Paul; Dartmouth College, Economics Tandel, Vaidehi; IDFC Institute Asher, Sam; Johns Hopkins University School of Advanced International Studies, Economics Malani, Anup; University of Chicago Law School
Keywords:	COVID-19, Epidemiology < INFECTIOUS DISEASES, Public health < INFECTIOUS DISEASES

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

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

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

Title: Representative estimates of covid-19 infection fatality rates from four locations in India: descriptive study

Authors: R. Cai[1], P. Novosad[2], V. Tandel[3], S. Asher[4], A. Malani[5,*]

Affiliations:
[1] B.A. (0000-0003-2295-0667)
Research Associate
Development Data Lab
1807 California St. NW, #106
Washington D.C., USA.
cai@devdatalab.org

[2] Ph.D. (0000-0001-8331-6630)
Associate Professor of Economics
Department of Economics
Dartmouth College 311Rockefeller Hall
Hanover, NH 03755
paul.novosad@dartmouth.edu

[3] Ph.D (0000-0001-6322-0068)
Junior Fellow
IDFC Institute
301-306, 3rd floor Makhija Chambers
Mumbai, Maharashtra 400050, India
vaidehi.tandel@gmail.com

[4] Ph.D. (0000-0002-2584-8011)
Assistant Professor of Economics
Department of Economics
Johns Hopkins University School of Advanced International Studies
WymanPark Building 5th Floor, 3100 Wyman Park Drive
Baltimore, MD 21211
sasher2@jhu.edu

Corresponding author:
[5] Ph.D, J.D. (0000-0002-2594-5778)
Lee and Brena Freeman Professor of Law
University of Chicago Law School
University of Chicago Pritzker School of Medicine, Chicago, Illinois, USA.
1111 E. 60th Street
Chicago, IL 60637
*Correspondence to: amalani@uchicago.edu

Manuscript Word Count: 3,050

Structured Abstract

Objectives: To estimate age- and sex-specific mortality risk among all SARS-CoV-2 infections in four settings in India, a major lower-middle-income country, and to compare age trends in mortality with similar estimates in high-income countries.

Design: Descriptive study.

Setting: India, multiple regions representing combined population >150 million.

Participants: Aggregate infection counts were drawn from four large population-representative prevalence/seroprevalence surveys. Data on corresponding number of deaths were drawn from official government reports of confirmed SARS-CoV-2 deaths.

Primary and secondary outcome measures: The primary outcome was age- and sex-specific infection fatality rate (IFR), estimated as the number of confirmed deaths per infection. The secondary outcome was the slope of the IFR-by-age function, representing increased risk associated with age.

Results: Among males aged 50–89, measured IFR was 0.037% in Tamil Nadu (95% CI: 0.035%, 0.040%), 0.12% in Karnataka (0.09%, 0.15%), 0.53% in Mumbai (0.52%, 0.54%), and an imprecise 5.64% (0, 11.16%) among migrants in Bihar. Estimated IFR was approximately twice as high for males as for females, heterogeneous across contexts, and rose less dramatically at older ages compared to similar studies in high-income countries.

Conclusions: Estimated age-specific IFRs in India are generally lower than in high-income countries, but vary widely across the country. Low estimated IFR and geographic variation within India may partly be due to under-testing of suspected COVID-19 deaths, though reporting

errors are unlikely to explain cross-country disparities. The elderly in India are at an advantage relative to peers in high-income countries. Low elderly IFR may partly be due to survivorship bias in a country with lower life expectancy. These findings suggest that mortality research from high-income countries may not generalize well to low-income settings; further research in lower-income countries is needed for context-appropriate policy.

Article Summary

Strengths and limitations of this study

- This study provides representative estimates of the age-specific COVID-19 infection fatality rate (IFR) in four socio-economically diverse regions of India, a major lower-middle-income county
- Due to high measurement cost, there are very few age-specific IFR estimates in low- and middle-income countries (LMIC), despite concerns that LMIC are more vulnerable and plausibly have different mortality patterns.
- This study utilizes the primary method of estimating IFR, combining population-representative prevalence/seroprevalence surveys with official death reports, allowing direct methodological comparison with dozens of similar estimates from high-income countries.
- We provide population-representative estimates for over 150 million people using the largest sample to date in a low- or middle-income country, and the first documentation of IFR among the large, highly vulnerable population of migrant workers.
- The main limitation is our reliance on official reports of confirmed COVID-19 deaths, which, due to under-reporting and under-testing, likely under-count the true number of deaths.

Introduction

Measuring the infection fatality rate (IFR) for severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) has been a major research objective since the beginning of the global pandemic. Reliable IFR estimates are essential for policy decisions on non-pharmaceutical interventions and vaccine allocation.[1–3] IFR estimates almost universally rely upon large-scale seroprevalence samples drawn from the general population, matched to official death data. Because of these data requirements, the vast majority of age-specific IFR estimates are based on data from high-income countries (HICs);[2–6] meta-analyses estimating age-specific IFR in low- and middle-income countries (LMICs)[7,8] rely on untested assumptions that key epidemiological characteristics (*e.g.* transmission dynamics, age-specific death rate) in HIC are generalizable to low-income settings. Studies measuring IFR in LMICs mostly report age-aggregated IFR,[9–13] which are difficult to compare across contexts; the age pattern of infection may vary and aggregate IFRs skew higher where older people contract a larger share of infections. Estimates of age-specific IFR in LMICs have only been made from small or non-representative samples.[14,15]

Early modelers of lower-income settings warned that IFRs could be higher, due to worse baseline population health and under-resourced healthcare systems.[8,15,16] Other researchers observed low case fatality rates (CFRs) in Sub-Saharan Africa and proposed that vaccination, past infection history, and effective mitigation strategies might have reduced mortality.[17,18] The age pattern of deaths in lower-income countries has skewed younger than in high-income countries, more so than can be explained by age distribution alone.[19–21]

We calculated age-specific IFRs from four samples in India representing a combined population exceeding 150 million. We first used population-representative seroprevalence

surveys in the city of Mumbai (N~=7000, population 18 million) and in the states of Karnataka (N~=1200, population 61 million) and Tamil Nadu (N~=26000, population 71 million). By matching these surveys to age-specific administrative death data, we calculated IFR without relying on non-representative testing data. The fourth data source is a survey of COVID-19 prevalence among randomly sampled short-term outmigrants (N ~=4000 infections, population 10 million), mostly working-age males, returning home to the state of Bihar, with mortality followup. Because these migrants were randomly sampled and tracked until recovery or death, the death rate among those who tested positive is interpretable as an IFR.

Our objective was to calculate age-specific IFRs in four locations and compare them to international estimates, which are based mostly on high-income countries. We further examined heterogeneity of IFR within India and by sex.

Methods

We studied three states and one mega-city with disparate demographic and health characteristics (Table 1). Qualitatively, Tamil Nadu and Karnataka are large, relatively wealthy, southern Indian states. Mumbai is India's most populous city, and the capital of the western state Maharashtra. Tamil Nadu, Karnataka, and Maharashtra have relatively robust health care infrastructure and vital registration.[22] In contrast, the northern state Bihar is one of the poorest in India, with the lowest stock of hospital beds per capita.[23] The Bihar sample is limited to a sub-population of returning migrants, primarily young male laborers who lost work opportunities during India's national lockdown. Short-term migrants were on average very poor even before the pandemic.[24] The sudden lockdown left them unemployed, and many experienced extreme physical and economic duress on the long journey home.[25,26]

India began its first nation-wide lockdown on March 24, 2020, and in February 2021 has the second highest number of country-wide confirmed cases in the world. The Indian government spends roughly 1.5% of GDP on healthcare, one of the world's lowest rates.[27] Discussion of India's COVID-19 preparedness has focused on under-resourced public hospitals, the largely unregulated private healthcare sector, and fear and stigma among the public surrounding infection.[27]

Data sources and study design

In Mumbai, Karnataka, and Tamil Nadu, we matched representative seroprevalence surveys to administrative reports of confirmed COVID-19 deaths.

In Mumbai, seroprevalence surveys were conducted for two weeks in July 2020 with representative sampling of three wards, one from each of the city's three zones, stratified by age, sex, and slum/non-slum dwellers.[10] The sample consisted of 6,904 participants (4,202 from slums and 2,702 from non-slums), tested for IgG antibodies to the SARS-CoV-2 N-protein using the Abbott Diagnostics Architect™ test. Data on cumulative deaths were collected from daily reports from the municipal governing body.

In Karnataka, seroprevalence surveys were conducted from June 15 to August 29, 2020, in representative samples of urban and rural areas in 20 out of 30 districts, stratified to generalize to five regions spanning all districts.[28] 1,196 participants were tested with an ELISA for antibodies to the receptor binding domain of the SARS-CoV-2 virus, developed by Translational Health Science and Technology Institute in India. We collected district-level death data from the Government of Karnataka Department of Health and Family Welfare bulletins.

In Tamil Nadu, a representative seroprevalence survey was conducted between October 19 to November 30, 2020, of adults aged 18 and older, covering the state's 37 districts.[29] Seropositivity was tested using either the iFlash-SARS-CoV-2 IgG or the Vitros anti-SARS-CoV-2 IgG CLIA kit. The analytical subsample was 26,107 antibody tests that could be conclusively determined as positive or negative. Case-level data on 12,019 recorded state-wide COVID-19 deaths, from March to December, 2020, was collected from daily government reports.

In Bihar, the state government began COVID-19 testing among returning out-of-state migrants soon after the first positive case was identified in a migrant on March 22, 2020. On May 4, Bihar began to randomly select migrants for testing. Random testing continued until July 21, though for a brief window (May 22–31) only migrants returning from seven major cities were sampled. We isolated the subsample of randomly selected migrants, yielding 4,362 individuals with positive tests.[25] Tests were conducted with TrueNat machines manufactured by MolBio Diagnostics in Goa, with positive tests confirmed by real-time PCR kits.[30] Bihar attempted to track all migrants who tested positive until they eventually recovered or died.

In all locations, population data came from the 2012 Socio-Economic and Caste Census.

Statistical analysis

In Mumbai, Karnataka, and Tamil Nadu, we estimated infection counts from representative seroprevalence surveys. Methods for estimating infection counts are described in detail below. We matched infection counts to deaths assuming that the infection-seroconversion delay is on average two days shorter than the infection-death delay.[31,32] To implement this, we calculated IFR as the cumulative number of deaths reported as of two days after the end of

seroprevalence testing, divided by the number of infections. Testing sensitivity to this assumption, we replicate results using deaths from one and two weeks after last day of seroprevalence testing, effectively generating upper bounds for the number of deaths (eFigures 1-3 in the Supplement). Where multiple evaluations of the antibody tests' sensitivity/specificity existed, we tested robustness to assuming minimum sensitivity (eFigures 4 and 5 in the Supplement).

In Mumbai, we first adjusted for test sensitivity and specificity using the Rogan-Gladen correction,[33] then calculated aggregate seroprevalence for each sampled ward and multiplied by ward population to estimate infection count. We estimated infection counts in non-sampled wards by assuming a constant rate of government under-reporting in wards in the same zone. This approach was supported by very similar case-to-seroprevalence ratios in the three wards with seroprevalence data (eTable 1). Age- and sex-specific infection shares were based on the seroprevalence survey (eFigure 6).

In Karnataka we adjusted for test inaccuracies,[33] then used census population counts to aggregate from regional to state-level infection counts, reweighting to match regional age-sex distributions. Methods for matching dates and deaths to infections is described in detail in the Supplement (eFigure 7). Because the seroprevalence survey period in Bangalore spanned two months (compared with less than three weeks in the other regions), we show results excluding Bangalore, where deaths may have been overestimated due to the longer survey period (eFigure 8).

In Tamil Nadu, we first calculated the population-representative seropositivity rate by district-age-sex group and type of test kit, then adjusted for test inaccuracies. We estimated the

number of state-wide infections per district-age-sex group by combining kit-specific seroprevalence estimates and multiplying by population, then summing across districts. In sensitivity checks, we re-estimated IFR limiting samples to districts where seroprevalence surveillance lasted less than three weeks (eTable 2, eFigure 9).

In Bihar, although enumerators attempted to track outcomes for all migrants, 1,530 (35%) infected individuals could not be tracked. In main estimates, we assumed that their fatality rates were the same as successfully tracked individuals; in sensitivity checks, we considered the possibility that all survived (Figure 2). High attrition is common in studies of migrant workers,[25] with followup in this case complicated by the ongoing crisis. We limited our analytic sample to 3,921 randomly-sampled male migrants, for whom 2,536 outcomes are known.

Matching representative seroprevalence surveys to administrative death data is the primary method of IFR measurement everywhere in the world.[2,4,5] The surveys in Mumbai, Karnataka, and Tamil Nadu thus provide credible, comparable estimates of age-specific fatality rates in those regions. In Bihar, because migrants were randomly sampled, there was no selection on symptomatic or severe cases, and mortality rates among positive cases can be interpreted as IFRs. As noted above, short-term migrants from Bihar are economically marginalized; their IFRs can be understood as representative for migrants, but not necessarily the general population.

We calculated IFRs in 10-year age bins, plus bins 10—49 and 50—89, in all locations. We used two large-scale meta-analyses[1,7]l of age-specific SARS-CoV-2 IFRs as reference groups. Both Levin et al.[1] and O'Driscoll et al.[7] draw almost exclusively from seroprevalence samples from Europe and the United States. The application of these samples to

mortality in LMICs (as in O'Driscoll et al.[7]) requires the as-yet untested assumption that multiple epidemiological factors (*e.g.*, transmission dynamics) are uniform between HIC and LMIC. Levin et al.[1] do not report IFR by sex; we estimated sex-specific IFRs in Levin et al. by assuming the same sex ratio in IFR as reported in O'Driscoll et al. For the larger age bins, we weighted age-specific IFR estimates from sample populations and meta-analyses by the Indian national population distribution, to ensure differences across contexts were driven by differences in age-specific IFRs, rather than population age distribution.

We calculated the slope of the natural log of IFR as a function of age by fitting a linear function to the most granular age-specific IFR data that could be obtained in each location. Additional details on the underlying samples and the methodology are in the Supplementary Materials. All analyses were conducted in Stata 16.0.

Patient and public involvement

No patients were directly involved in this study, all data used for statistical analysis was secondary, and anonymized and/or aggregated. Patients would not be able to identify themselves in the data. Because there were no patients or human subjects, the study was exempt from ethics committee approval, and conforms to the principles of the Declaration of Helsinki.

Seroprevalence studies were designed and implemented in partnership with local city and state governments. Details of patient involvement and protocols have been published in separate papers, and in reports from the respective city or state governments[10,25,28,29].

Results

We plotted age-specific IFR for each location on a log scale, to enable comparison at all ages despite exponential increases at higher ages found in all countries (Figures 1a and 1b). For

both males and females, there is substantial variation in IFR across the four locations in India. In Tamil Nadu and Karnataka, age-specific IFRs are an order of magnitude lower than those reported in the meta-analyses, especially over age 70. In Mumbai, estimates were close to the lower of the two meta-analyses at younger ages,[7] but were considerably lower than meta-analyses after age 60. For 60–69-year-old men, for example, we measured an IFR of 0.04% (95% CI: 0.038 to 0.048) in Tamil Nadu, 0.17% (0.092 to 0.240) in Karnataka and 0.62% (0.591 to 0.647) in Mumbai (Table 2); the two meta-analyses reported male IFR of 1.02%[7] and 1.86%[1] in this age group.

In contrast, mortality among male migrants returning to Bihar was an order of magnitude higher. Mortality among males aged 60–69 was extremely high but measured imprecisely due to the small sample of older males (4.26% [95% CI: 0.0 to 10.0%]). The larger age bins allowed a more precise measure of IFR in Bihar (Table 3). In both the 10–49 and 50–89 age bins, mortality in Bihar was an order of magnitude higher than in the other Indian locations and at least twice as high as rates in meta-analyses, after weighting to the Indian age distribution to ensure cross-context comparability. For the 50–89 age group, estimates were not precise enough to rule out equality between Bihar and the other locations. For the 10–49 age group, we can rule out equality ($p < 0.01$).

To the extent that an IFR advantage exists in India, it appears more strongly among the elderly. In all four regions, the overall increase in IFR with age was considerably less steep than in the reference meta-analyses (Figure 1), particularly at older ages. The meta-analyses suggest that an 80-year-old has about 100x the IFR of a 40-year-old; in Mumbai, the increase in risk factor is 40x and in Bihar it is only 10x. Specifically, male IFR increased on average by 4.7%, 9.6%, 8.9%, and 10.3% with each year of age in Bihar, Mumbai, Tamil Nadu, and Karnataka

respectively. We calculated comparable figures in the meta-analyses as 11.4%[7] and 12.3%.[1] The differences between the Indian and the reference groups were similar among females.

The main estimates are replicated in the Supplementary Materials under a range of different scenarios and assumptions; the ordering of IFRs across regions and with respect to the reference groups is highly robust (Figures 2a-d).

Discussion

Using best-practice methods applied in many high-income countries, we found substantial heterogeneity in age-specific COVID-19 infection fatality rate in India. In all four locations, we found a weaker increase in IFR over age than seen in other countries.

In Mumbai, estimated IFRs were lower than those measured in richer countries, particularly at ages where most deaths occur. In Tamil Nadu and Karnataka, IFRs were much lower at all ages. In a tracked sample of male migrants returning to Bihar, IFR estimates were an order of magnitude higher than the other two locations and twice as high as the international reference groups.

The strength of this study was the use of seroprevalence data representing over 150 million people, with a sufficiently large sample to calculate age-disaggregated IFR in a lower middle-income country. The main weakness of the study is that, like all COVID-19 population estimates, our estimates depend on the quality of underlying mortality data. The largest potential source of bias was our use of official reports of COVID-19 deaths, which likely undercount the true number of deaths.[22,34]

Misreporting of deaths would have to be substantial, however, to conclude that IFR in India is worse than in high-income countries. Focusing on the 50--89 age group, in Mumbai, a doubling of COVID-19 deaths is required to put estimated IFR in the range of the meta-analyses. It is at least plausible that deaths in Mumbai were undercounted by a factor of 2; between March and July, Mumbai recorded 6,600 excess deaths in addition to the 6,400 COVID-19 deaths used in this study.[34] In Karnataka and Tamil Nadu, however, COVID-19 deaths would have to be under-reported by factors of 10 and 30 respectively to bring IFR in line with international estimates. We cannot rule out extreme misreporting of deaths, but it would imply a scale of mortality in these states far higher than suggested by any other evidence. Further, we calculated IFR with standard methodology used in many cross-national settings, many of which are also characterized by under-reporting of COVID-19 deaths. As described in the Supplement, wherever possible we made conservative choices that would bias our IFR estimates upward rather than downward. In particular, antibodies may fade over time, so seroprevalence tests provide a lower bound on the cumulative infection rate.[35]

Official misreporting of COVID-19 deaths would not bias our IFR estimates in Bihar, due to the mortality followup methodology underlying these estimates. For our Bihar estimates to match the range of meta-analyses, deaths would need to have been *overcounted* by a factor of 2 for ages 50–89, and by 10 for ages 10–49. However, we do not know the base rate of migrant death. If migrant deaths would be high in absence of COVID-19, due to migrants' arduous return journeys, we may overstate the mortality attributable to COVID-19 in this group.

Few other studies have utilized sufficiently large seroprevalence samples to estimate age-specific IFR for a large lower-income population. Seroprevalence-based IFR estimates for older individuals in a Brazilian city[14] were slightly lower than our estimate for Bihari migrants, and

much higher than our seroprevalence-based estimates. However, seroprevalence samples of non-representative groups in Sub-Saharan Africa suggested high infection rates which are difficult to square with low overall mortality, consistent with our findings in India.[11]

It has been noted that, if age-specific IFR was uniform across countries, the pattern of mortality in low- and middle-income countries skews younger than would be predicted from age distributions alone.[19,21] Our study suggests that a flatter age profile in mortality could be a major factor driving this difference.

In large samples representing India's higher-income South, we found infection fatality rates that were broadly lower than those reported in richer countries. Among a sample of economically distressed migrants, we found IFRs that were twice as high. While incomplete administrative death data can explain some of the gap, the scale of the difference suggests that measurement error is unlikely to account for all of it. The relatively low IFR found in the South Indian settings parallels speculation about low fatality rates in Sub-Saharan Africa.[11,17,36] The drivers of these relatively low IFR numbers remain unknown, and could include survivorship bias, pre-existing immunity, or advanced public health preparation due to later pandemic onset in lower-income countries.

Bihari migrants may have had higher IFRs due to severe economic and physical distress. Migrant workers have worse health than the general population at baseline;[37] the circumstances at the beginning of the pandemic may have made this group exceptionally vulnerable to adverse health events following viral infection.

At the time of writing, these estimates are among the best available in a lower-income setting. Improved surveillance and accounting of SARS-CoV-2 are critical investments that would improve our understanding of the fatality risk of the virus in lower-income settings.

Table 1: Health and demographic context of sample locations

	Median age	GDP/capita	Cumulative infections on July 31	Cumulative COVID-19 deaths on July 31	Hospital beds per 100,000 population.
	Population Census 2011	NSDP nominal (2018-19 INT$)	JHU CSSE Covid-19 data[38]	JHU CSSE Covid-19 data[38]	Kapoor et al.[23]
Bihar	19.9	640	51,233	296	25.55
Maharashtra*	28.2	2,802	411,798	14,994	172.94
Karnataka	27.4	3,082	124,115	2,314	391.62
Tamil Nadu	29.9	2,831	245,859	3,935	174.83
India	24.0	1,964	1,695,988	36,511	137.62

Row 2 indicates the data source. *Mumbai is the capital city of Maharashtra.

Table 2: Age-specific infection fatality rates (%) from four locations in India

	Mumbai	Mumbai	Karnataka	Karnataka	Tamil Nadu	Tamil Nadu	Bihar migrants
Age	**Male**	**Female**	**Male**	**Female**	**Male**	**Female**	**Male**
10-19	0.004 (0.004,0.004)	0.001 (0.001,0.001)	0.000 (0.000,0.000)	0.001 (0.000,0.001)	NA	NA	0.000 (0.000,0.000)
20-29	0.013 (0.012,0.013)	0.005 (0.005,0.005)	0.004 (0.002,0.005)	0.002 (0.001,0.002)	0.001 (0.001,0.001)	0.001 (0.001,0.001)	0.649 (0.131,1.166)
30-39	0.041 (0.039,0.042)	0.019 (0.018,0.020)	0.013 (0.009,0.016)	0.006 (0.005,0.008)	0.003 (0.003,0.003)	0.001 (0.001,0.001)	1.810 (0.795,2.825)
40-49	0.112 (0.107,0.116)	0.058 (0.056,0.061)	0.027 (0.022,0.032)	0.012 (0.010,0.015)	0.010 (0.009,0.011)	0.004 (0.004,0.004)	1.529 (0.199,2.859)
50-59	0.355 (0.341,0.370)	0.172 (0.164,0.179)	0.073 (0.058,0.088)	0.040 (0.032,0.048)	0.026 (0.023,0.028)	0.009 (0.009,0.010)	2.381 (0.000,5.043)
60-69	0.619 (0.592,0.645)	0.317 (0.303,0.330)	0.166 (0.092,0.240)	0.050 (0.030,0.070)	0.043 (0.038,0.048)	0.016 (0.015,0.018)	4.255 (0.000,10.026)
70-89	0.837 (0.802,0.873)	0.511 (0.489,0.533)	0.163 (0.074,0.252)	0.106 (0.051,0.160)	0.058 (0.051,0.066)	0.015 (0.013,0.017)	12.500 (0.000,35.418)

Infection fatality rates as percentages. 95% confidence intervals in square brackets.

Table 3: Age-specific fatality rates in India ages 10—49 and 50—89

	Mumbai	Mumbai	Karnataka	Karnataka	Tamil Nadu	Tamil Nadu	Bihar migrants
Age	**Male**	**Female**	**Male**	**Female**	**Male**	**Female**	**Male**
10-49	0.033 (0.032,0.034)	0.016 (0.016,0.017)	0.009 0.007,0.010)	0.004 (0.004,0.005)	0.003* (0.003,0.003)	0.001* (0.001,0.001)	0.851 (0.467,1.235)
50-89	0.530 (0.517,0.543)	0.285 (0.278,0.292)	0.120 (0.090,0.150)	0.056 (0.043,0.069)	0.037 (0.035,0.040)	0.013 (0.012,0.014)	5.393 (0.000,11.156)

Infection fatality rates as percentages. 95% confidence intervals in square brackets. *In Tamil Nadu, seroprevalence collection and deaths were restricted to
Infection fatality rates as percentages. 95% confidence intervals in square brackets. *In Tamil Nadu, seroprevalence collection and deaths were restricted to
adults aged 18+. The 10-49 group assumes IFR in the 10—17 and 18—29 groups were equal, for weighting purposes.

Funding: This paper was partially supported by Emergent Ventures grant #466, awarded to Malani, Asher, and Novosad. The funder of the study had no role in the following: study design; collection, analysis, management, or interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit for publication. The corresponding author had full access to all of the data, and takes responsibility for the integrity of the data and accuracy of data analysis.

Author contributions: All authors participated in idea generation and development, empirical strategy design, and manuscript development. Malani and Tandel provided data on seroprevalence and mortality, and contextual knowledge regarding government sampling schemes and mortality registration. Cai and Novosad conducted the data analysis. All authors saw and approved the final version of the manuscript. The corresponding author attests that all listed authors meet authorship criteria and that no others meeting the criteria have been omitted.

Competing interests: Authors declare no competing interests.

Data and materials sharing: Replication code, data dictionary, and data will be posted in a public repository on Github. The repository will include all data on demographics and COVID-19 deaths by location, seroprevalence aggregates for Mumbai, Karnataka, and Tamil Nadu, and mortality rates by age and gender for migrants from Bihar. We do not have permission to share seroprevalence microdata. Replication code will be provided to reconstruct all results in the paper from these data.

References:

[1] Levin AT, Hanage WP, Owusu-Boaitey N, Cochran KB, Walsh SP, Meyerowitz-Katz G. Assessing the age specificity of infection fatality rates for COVID-19: systematic review, meta-analysis, and public policy implications. Eur J Epidemiol 2020. https://doi.org/10.1007/s10654-020-00698-1.

[2] Pastor-Barriuso R, Pérez-Gómez B, Hernán MA, Pérez-Olmeda M, Yotti R, Oteo-Iglesias J, et al. Infection fatality risk for SARS-CoV-2 in community dwelling population of Spain: nationwide seroepidemiological study. BMJ 2020;371:m4509. https://doi.org/10.1136/bmj.m4509.

[3] Yang W, Kandula S, Huynh M, Greene SK, Wye GV, Li W, et al. Estimating the infection-fatality risk of SARS-CoV-2 in New York City during the spring 2020 pandemic wave: a model-based analysis. Lancet Infect Dis 2020;0. https://doi.org/10.1016/S1473-3099(20)30769-6.

[4] Perez-Saez J, Lauer SA, Kaiser L, Regard S, Delaporte E, Guessous I, et al. Serology-informed estimates of SARS-CoV-2 infection fatality risk in Geneva, Switzerland. Lancet Infect Dis 2020. https://doi.org/10.1016/S1473-3099(20)30584-3.

[5] Salje H, Kiem CT, Lefrancq N, Courtejoie N, Bosetti P, Paireau J, et al. Estimating the burden of SARS-CoV-2 in France. Science 2020;369:208–11. https://doi.org/10.1126/science.abc3517.

[6] Ward H, Atchison CJ, Whitaker M, Ainslie KEC, Elliott J, Okell LC, et al. Antibody prevalence for SARS-CoV-2 in England following first peak of the pandemic: REACT2 study in 100,000 adults. MedRxiv 2020. https://doi.org/10.1101/2020.08.12.20173690.

[7] O'Driscoll M, Ribeiro Dos Santos G, Wang L, Cummings DAT, Azman AS, Paireau J, et al. Age-specific mortality and immunity patterns of SARS-CoV-2. Nature 2020:1–6. https://doi.org/10.1038/s41586-020-2918-0.

[8] Ghisolfi S, Almås I, Sandefur JC, Carnap T von, Heitner J, Bold T. Predicted COVID-19 fatality rates based on age, sex, comorbidities and health system capacity. BMJ Glob Health 2020;5:e003094. https://doi.org/10.1136/bmjgh-2020-003094.

[9] Hallal PC, Hartwig FP, Horta BL, Silveira MF, Struchiner CJ, Vidaletti LP, et al. SARS-CoV-2 antibody prevalence in Brazil: results from two successive nationwide serological household surveys. Lancet Glob Health 2020;8:e1390–8. https://doi.org/10.1016/S2214-109X(20)30387-9.

[10] Malani A, Shah D, Kang G, Lobo GN, Shastri J, Mohanan M, et al. Seroprevalence of SARS-CoV-2 in slums versus non-slums in Mumbai, India. Lancet Glob Health 2020;0. https://doi.org/10.1016/S2214-109X(20)30467-8.

[11] Uyoga S, Adetifa IMO, Karanja HK, Nyagwange J, Tuju J, Wanjiku P, et al. Seroprevalence of anti–SARS-CoV-2 IgG antibodies in Kenyan blood donors. Science 2020. https://doi.org/10.1126/science.abe1916.

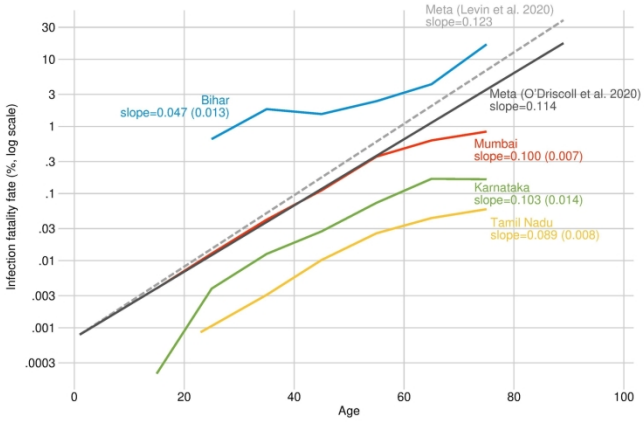
[12] Ioannidis J. The infection fatality rate of COVID-19 inferred from seroprevalence data. Bull World Health Organ 2020;BLT.20.265892.

[13] Xu X, Sun J, Nie S, Li H, Kong Y, Liang M, et al. Seroprevalence of immunoglobulin M and G antibodies against SARS-CoV-2 in China. Nat Med 2020;26:1193–5. https://doi.org/10.1038/s41591-020-0949-6.

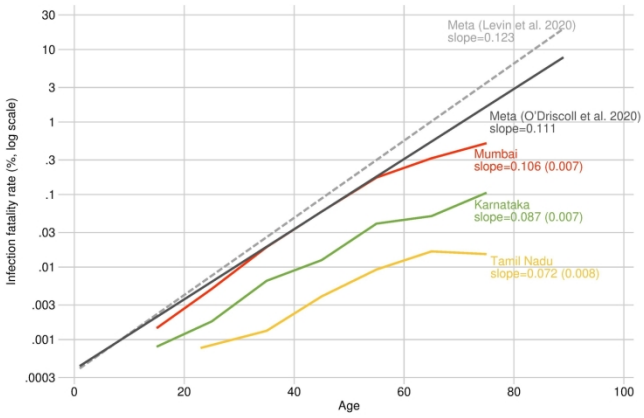
[14] Picon RV, Carreno I, da Silva AA, Mossmann M, Laste G, Domingues G de C, et al. Coronavirus disease 2019 population-based prevalence, risk factors, hospitalization, and

fatality rates in southern Brazil. Int J Infect Dis 2020;100:402–10. https://doi.org/10.1016/j.ijid.2020.09.028.
[15] Verity R, Okell LC, Dorigatti I, Winskill P, Whittaker C, Imai N, et al. Estimates of the severity of coronavirus disease 2019: a model-based analysis. Lancet Infect Dis 2020;20:669–77. https://doi.org/10.1016/S1473-3099(20)30243-7.
[16] Walker PGT, Whittaker C, Watson OJ, Baguelin M, Winskill P, Hamlet A, et al. The impact of COVID-19 and strategies for mitigation and suppression in low- and middle-income countries. Science 2020;369:413–22. https://doi.org/10.1126/science.abc0035.
[17] Nordling L. The pandemic appears to have spared Africa so far. Scientists are struggling to explain why. Sci AAAS 2020.
[18] Njenga MK, Dawa J, Nanyingi M, Gachohi J, Ngere I, Letko M, et al. Why is There Low Morbidity and Mortality of COVID-19 in Africa? Am J Trop Med Hyg 2020;103:564–9. https://doi.org/10.4269/ajtmh.20-0474.
[19] Demombynes G. COVID-19 Age-Mortality Curves Are Flatter in Developing Countries. The World Bank; 2020.
[20] Philip M, Ray D, Subramanian S. Decoding India's Low Covid-19 Case Fatality rate. National Bureau of Economic Research; 2020. https://doi.org/10.3386/w27696.
[21] Novosad P, Jain R, Campion A, Asher S. COVID-19 mortality effects of underlying health conditions in India: a modelling study. BMJ Open 2020;10. https://doi.org/10.1136/bmjopen-2020-043165.
[22] Chatterjee P. Is India missing COVID-19 deaths? The Lancet 2020;396:657. https://doi.org/10.1016/S0140-6736(20)31857-2.
[23] Kapoor G, Hauck S, Sriram A, Joshi J, Schueller E, Frost I, et al. State-wise estimates of current hospital beds, intensive care unit (ICU) beds and ventilators in India: Are we prepared for a surge in COVID-19 hospitalizations? MedRxiv 2020:2020.06.16.20132787. https://doi.org/10.1101/2020.06.16.20132787.
[24] Deshingkar P, Khandelwal R, Farrington J. Support for migrant workers: The missing link in India's development. Overseas Development Institute; 2008.
[25] Malani A, Mohanan M, Kumar C, Kramer J, Tandel V. Prevalence of SARS-CoV-2 among workers returning to Bihar gives snapshot of COVID across India. MedRxiv 2020. https://doi.org/10.1101/2020.06.26.20138545.
[26] Abi-Habib M, Yasir S. India's Coronavirus Lockdown Leaves Vast Numbers Stranded and Hungry. N Y Times 2020.
[27] Chetterje P. Gaps in India's preparedness for COVID-19 control. Lancet Infect Dis 2020;20:544. https://doi.org/10.1016/S1473-3099(20)30300-5.
[28] Mohanan M, Malani A, Krishnan K, Acharya A. Prevalence of SARS-CoV-2 in Karnataka, India. JAMA 2021. https://doi.org/10.1001/jama.2021.0332.
[29] Malani A, Tandel V, Parasa R, Ramachandran S, Sudharshini S, Prakash V, et al. Seroprevalence in Tamil Nadu in October-November 2020. MedRxiv n.d.
[30] CMO-PRC. Pre-release cm-243: Directions for random testing 2020.
[31] Linton NM, Kobayashi T, Yang Y, Hayashi K, Akhmetzhanov AR, Jung S-M, et al. Incubation Period and Other Epidemiological Characteristics of 2019 Novel Coronavirus Infections with Right Truncation: A Statistical Analysis of Publicly Available Case Data. J Clin Med 2020;9. https://doi.org/10.3390/jcm9020538.
[32] Stringhini S, Wisniak A, Piumatti G, Azman AS, Lauer SA, Baysson H, et al. Seroprevalence of anti-SARS-CoV-2 IgG antibodies in Geneva, Switzerland (SEROCoV-

POP): a population-based study. The Lancet 2020;396:313–9. https://doi.org/10.1016/S0140-6736(20)31304-0.
[33] Rogan WJ, Gladen B. Estimating prevalence from the results of a screening test. Am J Epidemiol 1978;107:71–6. https://doi.org/10.1093/oxfordjournals.aje.a112510.
[34] Banaji M. Mumbai Had 13k "Excess" Deaths in March-July. How Does Its COVID Story Change? Wire Sci 2020. https://science.thewire.in/the-sciences/covid-19-mumbai-all-cause-mortality-data-ifr-bmc-seroprevalence-survey/ (accessed December 21, 2020).
[35] Ripperger TJ, Uhrlaub JL, Watanabe M, Wong R, Castaneda Y, Pizzato HA, et al. Orthogonal SARS-CoV-2 Serological Assays Enable Surveillance of Low-Prevalence Communities and Reveal Durable Humoral Immunity. Immunity 2020;53:925-933.e4. https://doi.org/10.1016/j.immuni.2020.10.004.
[36] Mukherjee S. Why Does the Pandemic Seem to Be Hitting Some Countries Harder Than Others? New Yorker n.d.
[37] Choudhari R. COVID 19 pandemic: Mental health challenges of internal migrant workers of India. Asian J Psychiatry 2020;54:102254. https://doi.org/10.1016/j.ajp.2020.102254.
[38] Dong E, Du H, Gardner L. An interactive web-based dashboard to track COVID-19 in real time. Lancet Infect Dis 2020;20:533–4. https://doi.org/10.1016/S1473-3099(20)30120-1.



(a) Male



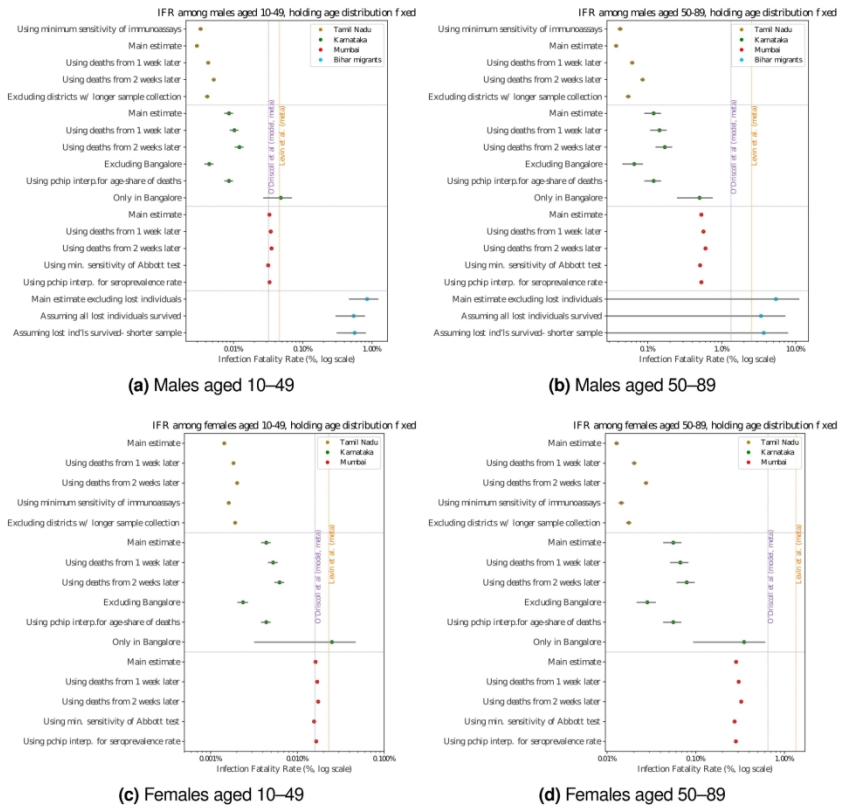
(b) Female

Figure 1
Age-specific infection fatality rate,
comparing four locations in India with international estimates

1

Age-specific infection fatality rate, comparing four locations in India with international estimates

132x239mm (300 x 300 DPI)



(a) Males aged 10–49 **(b)** Males aged 50–89

(c) Females aged 10–49 **(d)** Females aged 50–89

Figure 2
Age-specific infection fatality rates in India:
sensitivity tests

Markers indicate population-weighted pooled IFRs with India as the reference population. 95% confidence intervals are in grey.

Age-specific infection fatality rates in India: sensitivity tests

165x235mm (300 x 300 DPI)

Supplementary materials for:

Representative estimates of covid-19 infection fatality rates from four locations in India: descriptive study

Rebecca Cai, Paul Novosad, Vaidehi Tandel, Sam Asher, Anup Malani*

* Correspondence to: amalani@uchicago.edu

This PDF file includes:

Detailed Materials and Methods

eTables 1 to 2

eFigures 1 to 9

Materials and Methods

Bihar

Data

We made use of data on all positive cases in the state of Bihar found during random testing of incoming migrants during an early phase of the pandemic. The data was provided by the Health Department of the Government of Bihar. The data contained a sample of 4,954 active infections and their outcomes, reported between March 22 (the date on which the first positive case in Bihar was detected) and July 21, 2020. The vast majority of the sample (over 99%) consisted of migrants travelling from within India into Bihar, most on designated trains. Migrants were more likely to be sampled if they presented symptoms between March 22 and May 3. State policy beginning May 4 during the sample collection period mandated that travellers from within or outside India (mainly migrant workers returning home due to travel restrictions) be randomly sampled and tested for COVID-19 infection from March 20 to May 22, and after May 31. Between May 22–31, only migrants from seven high-infection cities (National Capital Region, Mumbai, Ahmedabad, Pune, Surat, Kolkata, and Bangalore) in India were randomly sampled. We isolated the subsample of migrants who were randomly selected for testing, yielding 4,362 cases.

During the sample period, migrants were tested with TrueNat machines manufactured by MolBio Diagnostics in Goa (India), and positive tests were confirmed with real-time reverse polymerase chain reaction (RT-PCR) kits Importantly, all infected migrants were tracked by the monitoring team, to determine whether they eventually recovered or died. Among randomly sampled male migrants, 1,385 infected individuals (35.%), whom we call "lost", could not be tracked and thus their final outcome is uncertain. The high level of attrition is common in studies of migrant workers, whose frequent movement complicates administrative registration and tracking, particularly during a crisis(**?**). We considered several approaches to adjusting for attrition, described below. The migrant sample, reflecting typical labor migration patterns in India, was overwhelmingly male (90%). Thus we limited our final analytical sample to 3,921 randomly sampled male migrants, for 2,536 of whom outcomes (recovery or death) are known.

Estimating infection fatality rate

Because everyone in the sample had tested positive for SARS-CoV-2, IFRs were estimated as the share of deaths among non-lost individuals in each age group. To account for potential biases due to attrition and delays between infection and recovery/death/reporting, we estimated IFRs using three separate methods, and report estimates from all three.

In age group a, denote the number of lost cases as $n_{a,lost}$, the number of recovered cases as $n_{a,recovered}$, and the number of cases ending in death as $n_{a,died}$.

Method 1 (main estimation): In our main estimation, we assumed that lost cases had the same IFR as successfully tracked cases, within each age group. This assumption was implemented by excluding lost individuals from the IFR calculation. Method 1 provided a midline IFR estimate:

$$IFR1_a = \frac{n_{a,died}}{n_{a,died} + n_{a,recovered}}$$

Method 2:In this estimation, we assumed that all lost cases eventually recovered. Thus Method 2 provided a lower bound IFR estimate:

$$IFR2_a = \frac{n_{a,died}}{n_{a,died} + n_{a,recovered} + n_{a,lost}}$$

Method 3: The share of cases with successful followup declined in late July as the volume of migrants increased. To account for potential right-censoring of reported outcome rate due to delays between report of initial infection and report of recovery/death, in the third method, we dropped all cases reported within two weeks of the last report date (July 21st):

$$IFR3_a = (\frac{(n_{a,died})}{n_{a,died} + n_{a,recovered} + n_{a,lost}} | \text{ infection reported on or before July 7})$$

Standard errors were estimated with the normal approximation for a proportion from multiple draws from a binomial distribution.

Mumbai

Data

Data on seroprevalence were obtained from a representative, stratified, random sample of slum and non-slum populations in three of twenty-four wards of Mumbai (see (**?**) for full survey design). Sample collection lasted two weeks and ended on July 14th in slums and July 19th in non-slums. The three wards were selected to represent the city's three broad zones (city, eastern suburbs, western suburbs); choice of sampled ward within each zone was by convenience. The sample consists of 6,904 participants (4,202 from slums and 2,702 from non-slums), who were tested for IgG antibodies to the SARS-CoV-2 N-protein using the Abbott Diagnostics Architect^{TM} N-protein based test. The samples were stratified by four age groups, sex, ward, and slum/non-slum residence.

Data on reported infections and deaths by ward and age distribution of deaths were provided in reports released by the municipal governing body (Brihanmumbai Municipal Corporation, hereafter BMC). Data on ward population in slums and non-slums came from the 2011 Population Census. Data on shares of population by age and sex in each ward-slum came from the 2012 Socio-Economic and Caste Census.

Estimating IFR

Estimating number of infections. The seroprevalence survey reported seropositivity in four age groups (12–24, 25–39, 40–60, 61+), called "coarse bins". To generate infection counts that could be compared with city death statistics (which are reported in 10-year age bins), seropositivity by 10-year age bin was interpolated by fitting a non-linear function over seropositivity in the coarse bins. For the main estimation, we interpolated seropositivity in 10-year bins, using the inverse distance-weighted mean of non-missing values (using the Stata package mipolate), weighting with the squared inverse of distance. In each coarse bin, the median age of residents in Mumbai City was used as the non-missing value for age. As a sensitivity analysis, we report IFR estimates using a piecewise cubic Hermite ("pchip") interpolation for seropositivity Interpolation predicted seroprevalence for the midoint of each 10-year age bin, separately by sex, ward, and slum status.

The estimated sensitivity of the chemiluminiscence immunoassay ranges from 90% (95% CI: 74. to 96.)(**?**) to 96.% (89. to 99.) (**?**) while specificity in those studies was 100% (95% CI: 95. to 100) and 99.0%, respectively. We estimated seroprevalence from seropositivity using the Rogan-Gladen correction(**?**) to account for imperfect accuracy of tests. In the main results, we used the midpoint of mean sensitivity estimates (93.5%) and the midpoint of corresponding specificities (99.%). As a sensitivity analysis, we replicated results with an upper bound for seroprevalence based on the Abbott test's lower bound of sensitivity (90.%) and upper bound of specificity (100%)(**?**) (Figure 4).

Denote the estimated number of infections in age bin a, sex g in sampled ward s as:

$$\widehat{inf}_{ags} = SP_{ags} \times pop_{ags}$$

where $SP_{ag,s}$ is the estimated seroprevalence rate, and $pop_{ag,s}$ is population.

Estimating the number of infections in non-sampled wards. BMC death data reported the ward of death, but not the ward of residence. Discussion with government officials and review of the data indicated that the ward of death was not a reliable indicator of ward of residence. This implied that calculating IFR by dividing the number of ward-level deaths by the number of ward-level infections would overestimate deaths in wards with large hospitals and underestimate them elsewhere. Instead, we used the seroprevalence surveys to generate estimates of city-wide infection counts.

To estimate true number of infections in non-sampled wards, we drew on administrative ward-level infection counts (which were universally available from city reports), and assumed that they were proportional to actual infections at similar rates in different wards of the city. Effectively, this amounts to assuming that the BMC underestimated the true population infection count at the same rate in sampled and non-sampled wards within the same zone. This assumption is supported by Table 1, which shows that in the three wards where we obtained seroprevalence data, case multipliers were very similar.

Thus, in each zone z, we calculated a case multiplier based on sampled ward s:

$$\gamma_z = \frac{\sum_a \sum_g \widehat{inf}_{ag,s}}{\text{BMC-reported cases}_s}$$

The multiplier indicates the under-reporting rate in each zone z. The numerator of the expression is calculated from the seroprevalence surveys as above, and the denominator is taken from the BMC reports. BMC-reported cases were measured as of July 19, the last day of seroprevalence sample collection. We then multiplied the BMC's reported number of positive cases in non-sampled ward n in zone z by γ_z. That is,

$$\widehat{inf}_{n,z} = \gamma_z \times \text{BMC-reported cases}_n$$

The benefit of this approach is that it allows pandemic intensity to vary across wards, a realistic assumption given significant ward-level variation in reported cases per capita and in the number of containment zones.

This approach also implicitly assumes that the BMC under-reports cases in slums and non-slums at the same rate, *i.e.* a ward's case multiplier does not depend on share of population living in slums. This assumption is also supported by the consistent multipliers reported in Supplement Table 1, across three wards with different slum shares.

Estimating the number of infections in each age-sex group in non-sampled wards. We did not observe the age and sex distribution of infections outside of the sampled wards, so it was necessary to assume that non-sampled wards had the same age and sex distribution of infections of sample wards. This was supported by similar age and sex distributions of infections in the three wards with seroprevalence surveys. Figure 6 shows the calculated age and sex distribution of infections; note that the distribution of infections measured with seroprevalence skews younger than the number of reported positive cases, which we presume omits many infected but asymptomatic young people. This approach would cause error if the age distribution varied substantially across wards, but it is overall quite similar; even the median age gap between slums and non-slums was less than one year.

The number of infections in non-sampled ward s for sex g in age a was thus calculated as:

$$\widehat{inf}_{ag,n} = \frac{\sum_s \alpha_{ag,s}}{\sum_s \sum_a \sum_g \alpha_{ag,s}} \times \widehat{inf}_n,$$

where $\alpha_{ag,s}$ is the age-sex group's share of total cases in sampled ward s.

Estimating the number of deaths. To map seroprevalence numbers to death numbers, the time between infection and death and the time between infection and seroprevalence are needed. The literature suggests a distribution of delay between symptom onset and death(**?**) that is wider than that between onset and seroconversion(**?**). Linton et al. estimated a median time delay of 13 days (17 days with right truncation) between illness onset to death. Stringhini et al. estimated a mean delay of 11. days between symptom onset and seroconversion. Based on these estimates, we assumed that the delay between infection and death is on average two days longer than the delay between infection and seroconversion. In the main results, the number of deaths was therefore measured as the cumulative deaths reported in each Mumbai ward as of July 21. This is likely to slightly overstate the IFR, since some deaths may have been associated with individuals who contracted the virus after testing negative in the seroprevalence surveys. However, this upward bias is partially balanced out by the fact that the time between seroconversion and death is not uniform and is likely to be longer than 2 days for a non-trivial share of cases.

Rather than model non-uniform delays between infection and death, we bounded our IFR estimates from above by choosing more conservative death dates. In sensitivity analyses reported below (Figure 1), we replicated IFR estimates using deaths from one week (July 28) and two weeks (August 4) after the end of seroprevalence surveying, both of which plausibly overestimated the number of deaths related to the seroprevalence surveys, given the context of steadily increasing case counts in Mumbai from June to August.

The assumption that deaths measured 1 and 2 weeks later will lead to upward biased IFRs is further strengthened by recent evidence from roughly 125,000 cases in two other Indian states(**?**), which found that delays between case report and death were significantly shorter than delays found in China and the United States(**?**).

We used the age distribution of deaths as reported by the BMC up to the date used for measuring deaths, and the sex distribution (65% male, 35% female) up to August 3(**?**) (the sex distribution of deaths was not included in earlier reports). This yields the estimated number of city-wide deaths by age-sex group, d_{ag}.

Estimating city-wide IFR by age in Mumbai. Denote the final city-wide IFR in Mumbai, in age bin a for sex g, as IFR_{ag}:

$$IFR_{ag} = \frac{d_{ag}}{\sum_{ns} \widehat{inf}_{ag,ns} + \sum_s \widehat{inf}_{ag,s}}$$

Standard errors of IFRs were calculated reflecting propagation of the design-based standard errors of the age- and sex-specific seroprevalence estimates with a normal approximation.

Karnataka

Data

Data on seroprevalence were obtained from the Karnataka Seroprevalence Survey (hereafter KSS) a state-wide representative sample of urban and rural areas in 20 out of 30 districts in Karnataka, representing 5 broader regions (see Mohanan et al. *(38)* for a detailed survey description). The sample was collected from June 15 to August 29, 2020. Collection times within individual regions were significantly shorter. The study sample was drawn from an existing representative sample of a panel survey—the Consumer Pyramids Household Survey (CPHS)—collected by the Center for Monitoring Indian Economy (CMIE). Our analytical subsample consists of 1,196 tests for IgG antibodies to the receptor binding domain (RBD) of the SARS-CoV-2 virus using an ELISA test developed by Translational Health Science and Technology Institute, India. The sample was not stratified by age and sex, an issue addressed below.

Data on confirmed COVID-19 deaths by district were drawn from Government of Karnataka Department of Health and Family Welfare bulletins, which are released several times per week. Data on the age distribution of total COVID-19 deaths were given by public reports from the state COVID-19 task force. Data on the sex distribution of deaths by age group were obtained from an individual-level dataset of confirmed COVID-19 deaths which was updated through July. The case-level death data were parsed from covid19india.org. Age- and sex-disaggregated population for districts and regions was drawn from the 2012 Socio-Economic and Caste Census (SECC).

Estimating IFR

Estimating the number of infections. The KSS dataset was designed to be representative of 5 broader regions in Karnataka. We therefore can take the ELISA positive test rate as an unbiased measure of the region-level positivity rate. We pooled the data across regions to obtain a statewide test positivity rate in each age and sex group, weighting by region population in each age-sex group.

We then corrected for the sensitivity (84.%) and specificity (100%) of the ELISA imunoassay(**?**), using the Rogan-Gladen correction(**?**). This yielded the estimated seroprevalence by age-sex group SP_{ag}, which is multiplied by population pop_{ag} in each age-sex bin to generate an estimated number of infections $\widehat{inf}_{ag}$, as was done in Mumbai.

Estimating the number of deaths. The seroprevalence samples were collected at different times in different regions, with the survey period spanning roughly two months (Table 2). To estimate an IFR, we need to match the timing of deaths to the timing of seroprevalence surveying in each region.

Choice of dates for measuring deaths. As in Mumbai, we worked from an assumption that the average time difference between seroconversion and death was two days, while testing sensitivity to alternate assumptions (Figure 2). We therefore matched the estimated number of infections calculated in each region to the number of deaths recorded in administrative data two days after the last date of seroprevalence surveying. As in Mumbai, if the two-day delay between seroconversion and death was uniform, this approach would overestimate the IFR, because it counts the deaths of some people who may have been infected *after* recording negative seroprevalence tests.

In all regions except Bangalore, seroprevalence surveying was conducted over a three week period or less, making it straightforward to match test data to death data. In Bangalore, surveying was begun in mid-June but was interrupted by a lockdown. Survey teams returned to finish sampling in the last week of August. Matching Bangalore deaths to the last date of seroprevalence surveying is therefore likely to overestimate the IFR, because a number of those deaths may have been associated with individuals contracting SARS-CoV-2 after testing negative. It was not possible to disaggregate the early and late surveys because death reporting was at the district level, and the early and late survey groups were not representative in and of themselves. To adjust for increased uncertainty regarding the number of infections in Bangalore, we therefore report a sensitivity analysis for all of Karnataka excluding Bangalore (Figure 8).

On some days, official deaths were not reported; in those cases, we used deaths from the following day.[1]

Estimating the number of deaths in each demographic group: The Karnataka state government released total death counts on a daily basis, but only intermittently published the age distribution of state-wide deaths. To attribute daily deaths to age and sex groups, we used the age distribution of deaths from the nearest date that was available. The largest period between the date used for deaths and the date used for age-shares was 13 days.

Government reports provided age shares of deaths in 10-year bins in the form (e.g.) 51-60, while the seroprevalence surveys provided age bins in the form (e.g.) 50-59. To harmonize the age groups, we use the medians of the provided bins (e.g. median of 51-60 is 55.) to interpolate death data to match the age bins in the seroprevalence data, using an inverse distance weighted average method via the mipolate Stata package. Because the target age bins were very close to the available age bins, the risk of error here is small. As a sensitivity test, we replicated IFRs using piecewise cubic Hermite interpolation For more details, see the discussion on interpolation in Mumbai.

[1] In Belgaum, the target date was July 27th; we used July 28. In the sensitivity test, we used August 11 instead of August 10, which was unavailable.

In the absence of death data disaggregated by age and sex on most dates, we assumed that, within age group, the sex distribution of deaths was uniform across regions and equal to the state-wide sex distribution of deaths reported between April and July. This assumption is supported by the finding that IFRs among males were approximately double those among females, consistent with reports from other countries.

Standard errors of IFRs reflect propagation of design-based standard errors of the age- and sex-specific seroprevalence estimates with a normal approximation.

Tamil Nadu

Data

Data on seroprevalence in Tamil Nadu comes from a state-conducted population-level seroprevalence survey of 26,640 adults aged 18 and older, covering the 37 districts of the state. The sample was collected between October 19 and November 30, 2020. Collection times within districts were often significantly shorter. The sampling frame divided Tamil Nadu's 37 administrative districts (as of February 2020) into health unit districts (HUDs), then formed and randomly sampled urban and rural clusters. Within clusters, enumerators started at a randomly selected GPS starting point, sampling one person from households adjacent to the starting point (using the Kish method) to provide a biosample, until 30 persons were sampled per cluster. Serum was analyzed for IgG antibodies to the SARS-CoV-2 spike protein using either the iFlash-SARS-CoV-2 IgG (Shenzhen YHLO Biotech; sensitivity of 95.% and specificity of 95.% per manufacturer(**?**)) or the Vitros anti-SARS-CoV-2 IgG CLIA kit (Ortho- Clinical Diagnostics; sensitivity of 90% and specificity of 100% per manufacturer). For uniformity, in each district, one type of kit was used; in one district (Chennai) both kits were used. Our analytical subsample consists of 26,107 CLIA antibody tests that could be conclusively determined as positive or negative.

Case-level data on state-wide COVID-19 deaths was collected from daily government reports released on https://stopcorona.tn.gov.in/daily-bulletin/. The data cover all recorded deaths, beginning on March 25 and updated until December 24, 2020. The data was collected and shared by the faculty and staff of the Urban Expansion Observatory at Pillai College, New Panvel, Maharashtra. The dataset contains 12,019 observations, each with information about age, sex, date of reported positive test and death, and district. Age- and sex-disaggregated population data were from the 2012 Socio-Economic and Caste Census.

Estimating IFR

Estimating the number of infections. We estimated the number of state-wide infections associated with measured seroprevalence in three steps. First, we calculated positive test rate

by district-age-sex group, separately for each kit. Positive test rate was estimated by regressing an indicator for positive result on district-age-sex group indicators, clustering standard errors within the randomly sampled clusters. Seroprevalence sample collection was stratified by district, health unit district (HUD), then cluster; within clusters, age and sex of test participants was random. Thus we take the positive test rate for each district-age-sex group as representative.

Second, we adjusted for test inaccuracies for each kit, using the Rogan-Gladen correction (**?**) and the manufacturer-provided sensitivity and specificity. In a sensitivity check, we utilized the lowest estimated sensitivity and corresponding specificity, from any manufacturer-conducted or independent analyses of each kit (Figure 5). Independent analysis of the iFlash kit from Shenzhen YHLO Biotech estimated sensitivity of 93% (95% CI: 84. to 97.) and specificity of 92.% (85. to 97.)(**?**). FDA evaluation of the Vitros kit from Ortho-Clinical Diagnostics suggests 100% sensitivity (95% CI: 88. to 100%) and 100% specificity (95. to 100%)(**?**), while other analysis estimated a sensitivity of 98.% (92. to 100) and specificity of 97.% (85 to 100)(**?**). Note that, unlike in Mumbai and

Karnataka, the minimum specificity of the kits in Tamil Nadu had lower corresponding specificity, leading to lower overall seroprevalence estimates. In the district in which both kits were used, kit-specific seroprevalence estimates were averaged, using proportion of sample size (by age-sex group) as the weight.

Third, we estimated number of infections in each district-age-sex group by multiplying seroprevalence rate by population. Age- and sex-disaggregated population data was available for census districts. Finally, estimated state-wide infections by age-sex group were calculated by simply summing over all districts.

Estimating number of deaths in each demographic group. As in Mumbai and Karnataka, we matched the estimated number of infections calculated in each district to the number of deaths recorded in administrative data two days after the last date of seroprevalence surveying. We test sensitivity to alternative assumptions by measuring cumulative deaths 1 week and 2 weeks after the main date (Figure 3). As explained in the supplement sections on Mumbai and Karnataka, these are all plausible over-estimates of deaths associated with the measured seroprevalence level. Cumulative deaths in each demographic group were measured up to the specified date. Cumulative deaths were

measured from March through December, a longer span than in other locations. This may over-estimate deaths, and therefore over-estimate IFR, if infected individuals gradually become seronegative after recovery. Available evidence suggests that antibody loss varies significantly with symptom severity ((**?**; **?**; **?**)). Because we cannot precisely estimate antibody loss rates across the population, and because IFR estimates in Tamil Nadu are the lowest across the four locations, we simply note that, given available data, our IFR estimates are conservatively high.

Seroprevalence surveying lasted longer than three weeks in 6 out of 37 districts. In these districts, there is a risk that seroprevalence in the population changed during sample collection. During a period of increasing pandemic intensity, this may under-estimate seroprevalence, over-estimating IFR. As a sensitivity check, we limit analysis of both seroprevalence and deaths to the 31 districts in which seroprevalence sample collection was less than three weeks (Figure 9).

Age- and sex-specific IFRs were estimated as the proportion of state-wide deaths divided by estimated infections. Standard errors reflect propagation of error from the HUD-age-sex estimates of positive test rates.

1 Supplementary Tables

eTable 1
Zone-wise
case multipliers for main and higher seroprevalence estimates based on sampled wards

		No. infections			γ_z	
Ward	*Zone*	*BMC report*	*main SP*	*high SP*	*main SP*	*high SP*
(1)	(2)	(3)	(4)	(5)	(6)	(7)
F North	City	4,017	190,652	211,835	47.6	52.3
M West	Eastern	2,965	139,791	155,322	47.5	52.9
R North	Western	2,421	145,413	161,569	60.6	66.4

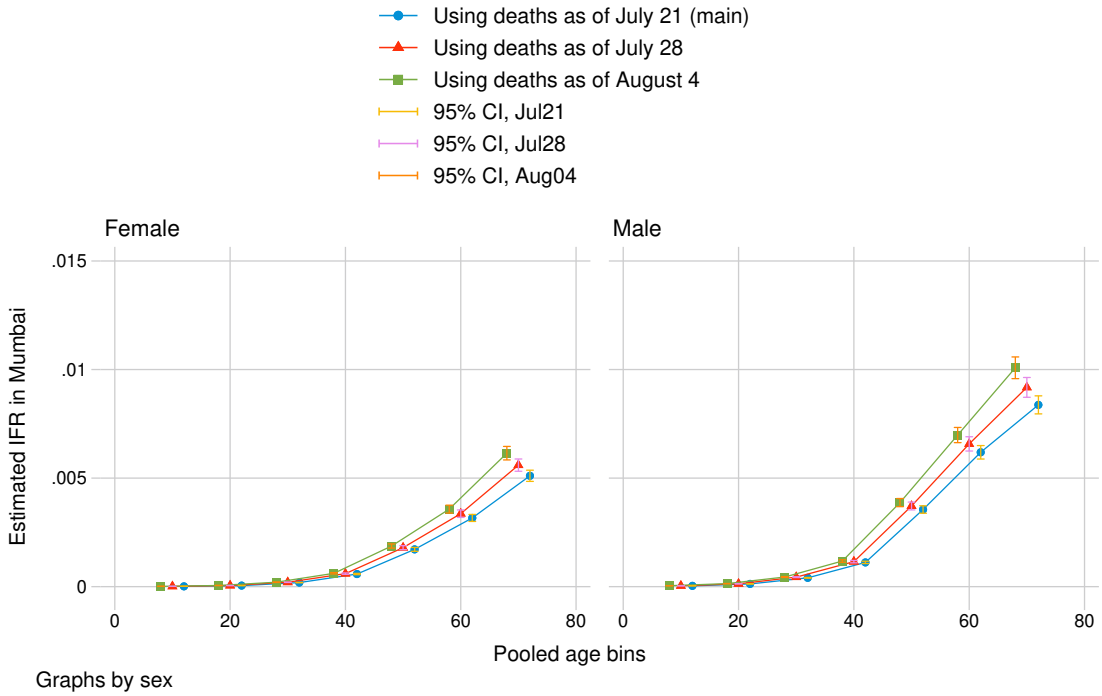
"Number of infections, main SP" refers to the estimated seroprevalence (using the midpoint estimated sensitivity of the antibody test) multiplied by population in each sampled ward. "Number of infections, high SP" uses lowest bound sensitivity of the antibody test. Case multiplier "γ_z, main SP" (Column 6) was calculated by dividing Column 4 by Column 3. "γ_z, high SP" (Column 7) was calculated by dividing Column 5 by Column 3. Main SP indicates seroprevalence estimated from midpoint of two published estimates of sensitivity of the antibody test. High SP indicates seroprevalence was estimated using the minimum sensitivity and maximum specificity of the antibody test, generating a upper-bound estimate.

eTable 2
Karnataka: duration of sample collection by region

Region	Duration of sample collection (days)	Dates of sample collection
Bangalore	73	June 17 – August 29
Mysore	18	August 3 – August 21
Kannada	16	August 6 – August 21
Belgaum	17	July 8 – July 25
Gulbarga	10	July 21 – July 31

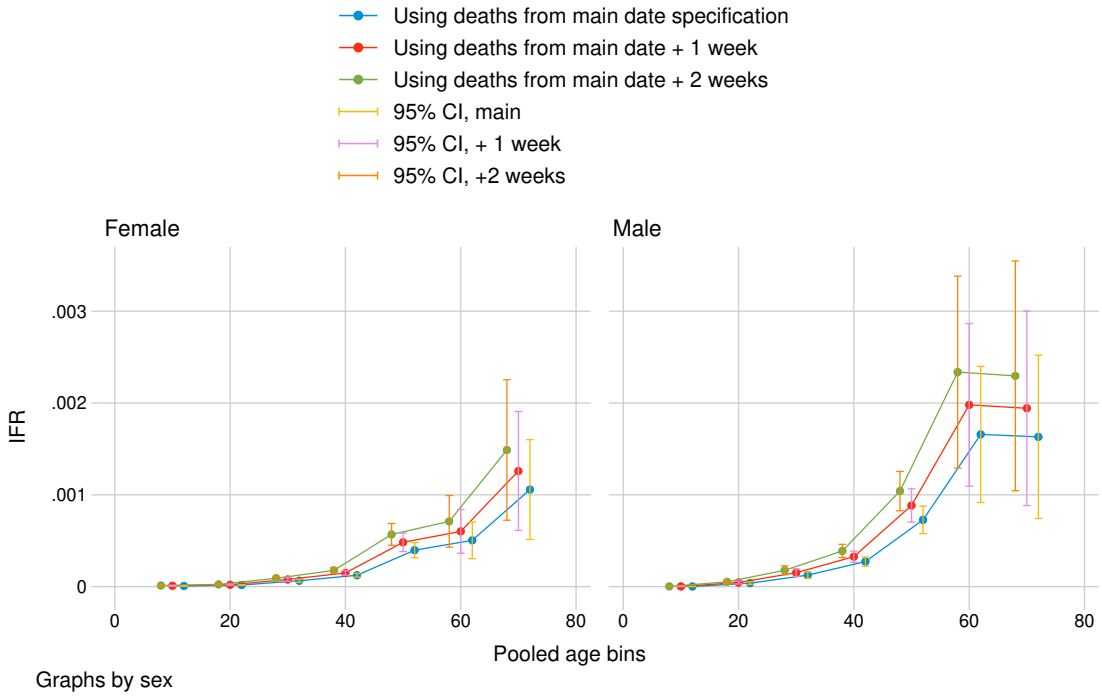
eFigure 1

Mumbai: sensitivity analysis
using number of COVID deaths from 1 and 2 weeks after date of deaths in main estimation



"Date of death" refers to the day on which we measured cumulative deaths as reported by the city government (BMC). The main date specification measured deaths two days after the end of seroprevalence sample collection. Graphs by sex with 95% confidence intervals. Standard errors reflect propagation of error from design-based uncertainty of seroprevalence estimates. IFRs are calculated in age bins 0-19, 20-29, ... 60-69, and 70+.

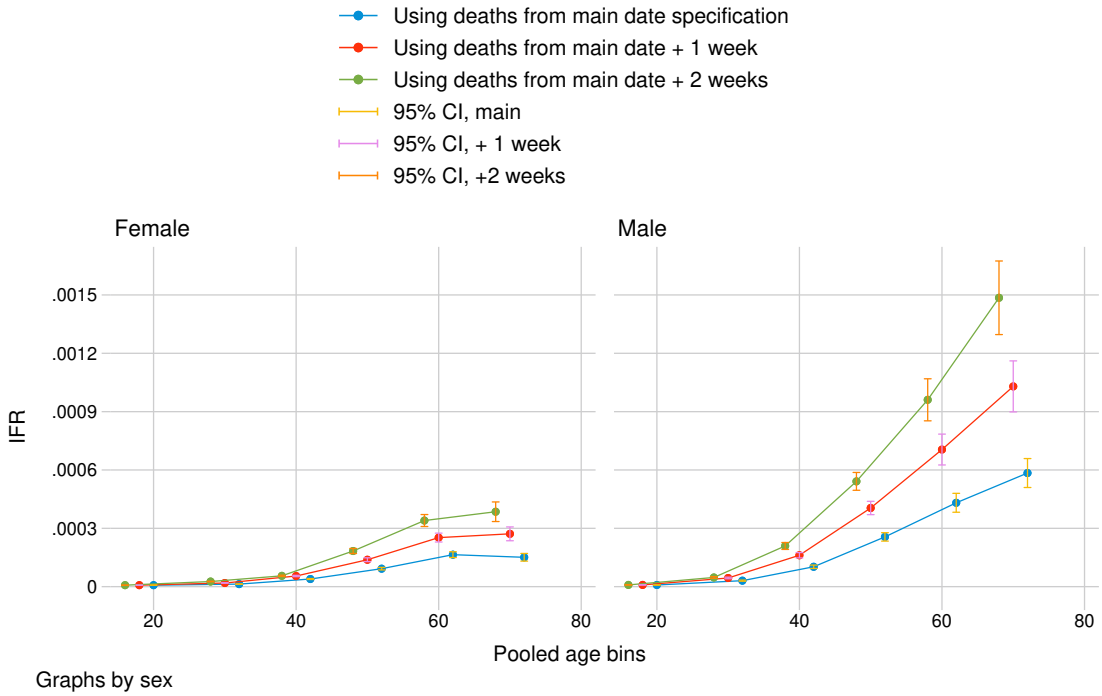
eFigure 2
Karnataka: sensitivity analysis
using number of COVID deaths from 1 and 2 weeks after date of deaths in main estimation



"Date of death" refers to the date on which we measured cumulative COVID-19 deaths. Main date of specification was determined separately for each sampled region as two days after the median date of sample collection. Graphs by sex with 95% confidence intervals. Standard errors reflect propagation of error from design-based uncertainty of seroprevalence estimates. IFRs are calculated in age bins 0-9, ... 60-69, and 70+.

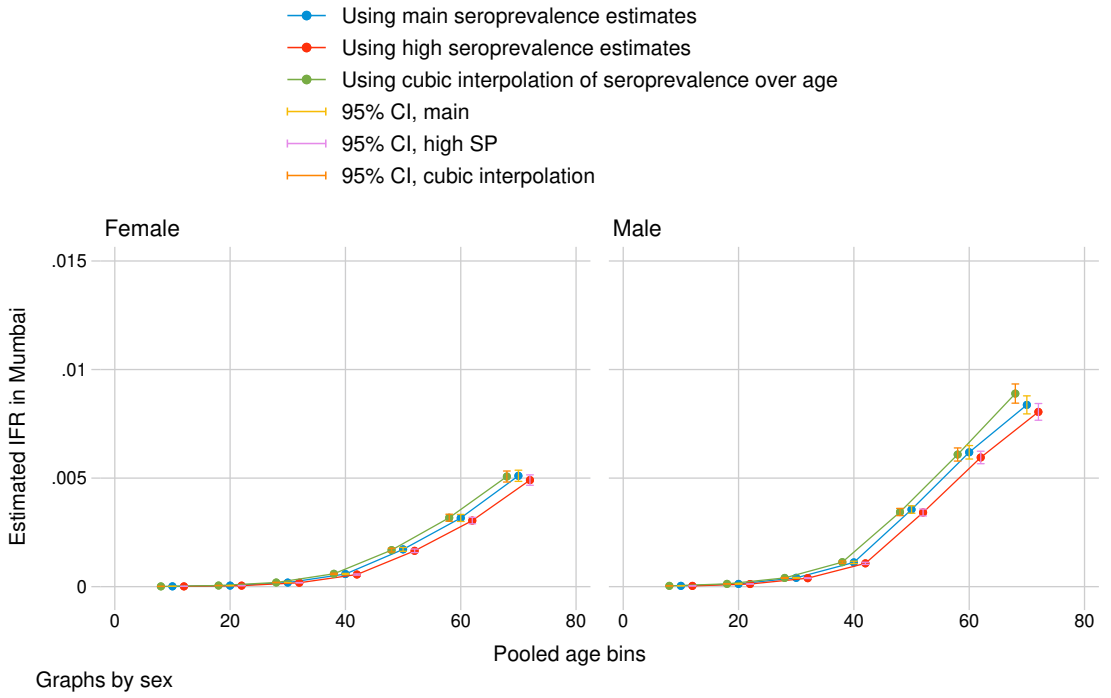
eFigure 3

Tamil Nadu: sensitivity analysis
using number of COVID deaths from 1 and 2 weeks after date of deaths in main estimation



"Date of death" refers to the date on which we measured cumulative COVID-19 deaths. Main date of specification was determined separately for each sampled district (N = 37) as two days after the last date of sample collection. Graphs by sex with 95% confidence intervals. Standard errors reflect propagation of error from uncertainty in estimating positive test rate by HUD-age-sex group. IFRs are calculated in age bins 18-29, 30-39 ... 60-69, and 70+.

eFigure 4
Mumbai: sensitivity
analysis, using alternative estimate of seroprevalence and different interpolation method

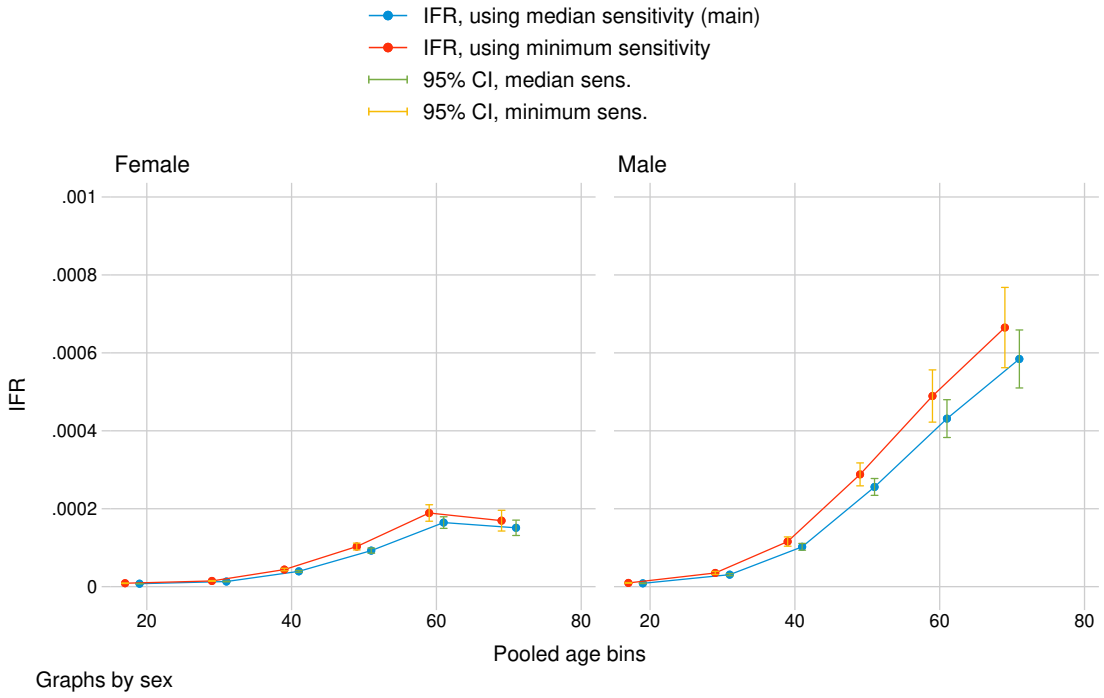


"Main seroprevalence estimates" use midpoint sensitivity estimate of the Abbott antibody test to calculate seroprevalence from seropositivity in sampled wards, then interpolates seroprevalence to finer age bins with inverse distance weighting (IDW). "High seroprevalence estimates" use minimum sensitivity of the Abbott test to calculate seroprevalence from seropositivity and IDW interpolation. The final sensitivity analysis uses midpoint sensitivity, but piecewise cubic Hermite interpolation to estimate seroprevalence in finer bins. Graphs by sex with 95% confidence intervals. Standard errors reflect propagation of error from design-based uncertainty of seroprevalence estimates. IFRs are calculated in age bins 0-19, 20-29, ... 60-69, and 70+.

eFigure 5

Tamil Nadu:

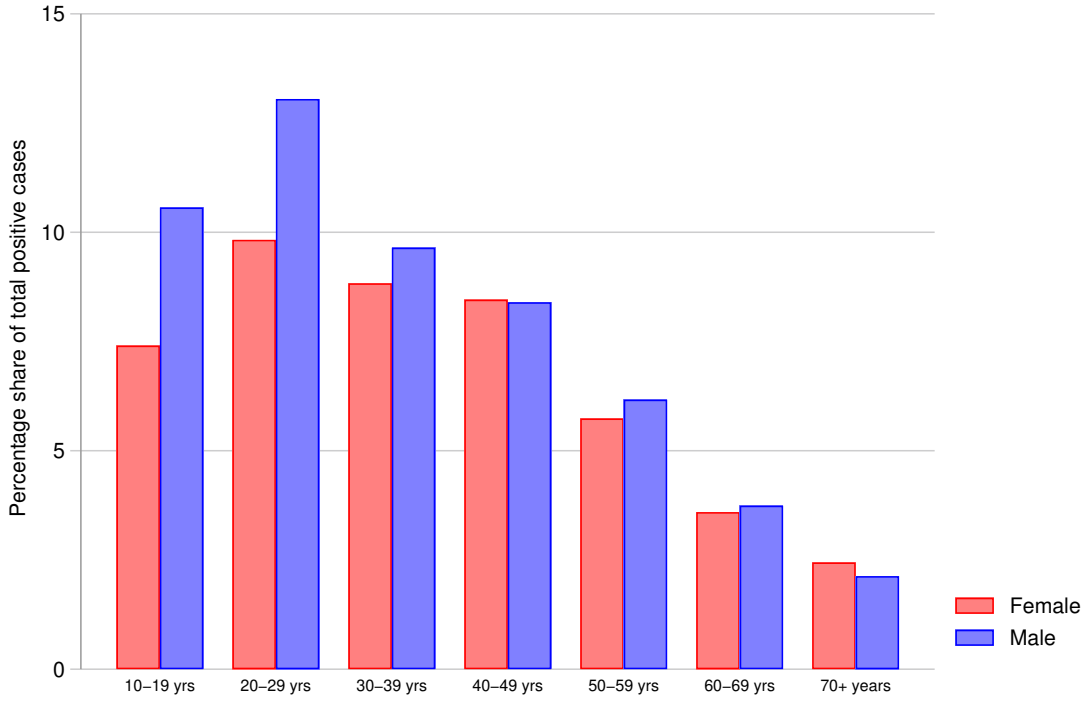
sensitivity analysis using minimum sensitivity and corresponding specificity of immunoassays.



Two kits were used to evaluate seropositivity. Seroprevalence rate was calculated from the seropositivity rate using the Rogan-Gladen correction for imperfect test sensitivity and specificity. Main estimation used the manufacturer-provided sensitivity and corresponding specificity of the kits. The robustness check uses the lowest estimated sensitivity of both kits, which was the manufacturer-provided estimate for the Ortho-Clinical kit. Graphs by sex with 95% confidence intervals. Standard errors reflect propagation of error from uncertainty in estimating positive test rate by HUD-age-sex group. IFRs are calculated in age bins 18-29, 30-39 ... 60-69, and 70+.

eFigure 6

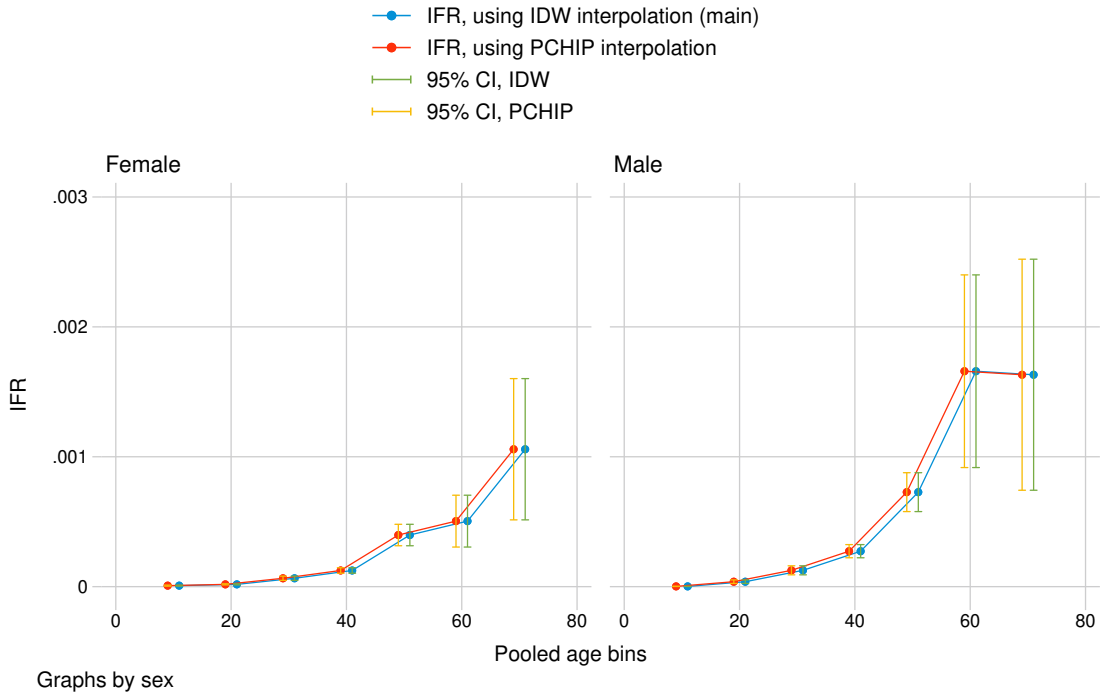
Age-sex cohorts' share of positive cases from sampled wards in Mumbai seroprevalence survey



"Total positive cases" refers to the estimated number of total infections in Mumbai, multiplying age- and sex-specific seroprevalence rate by group population, summed across age-sex groups and wards. The age- and sex-share of total cases refers to estimated number of infections in age-sex group ag, divided by estimated total infections. Age bins are 0-19, 20-29, ...60-69, and 70+.

eFigure 7

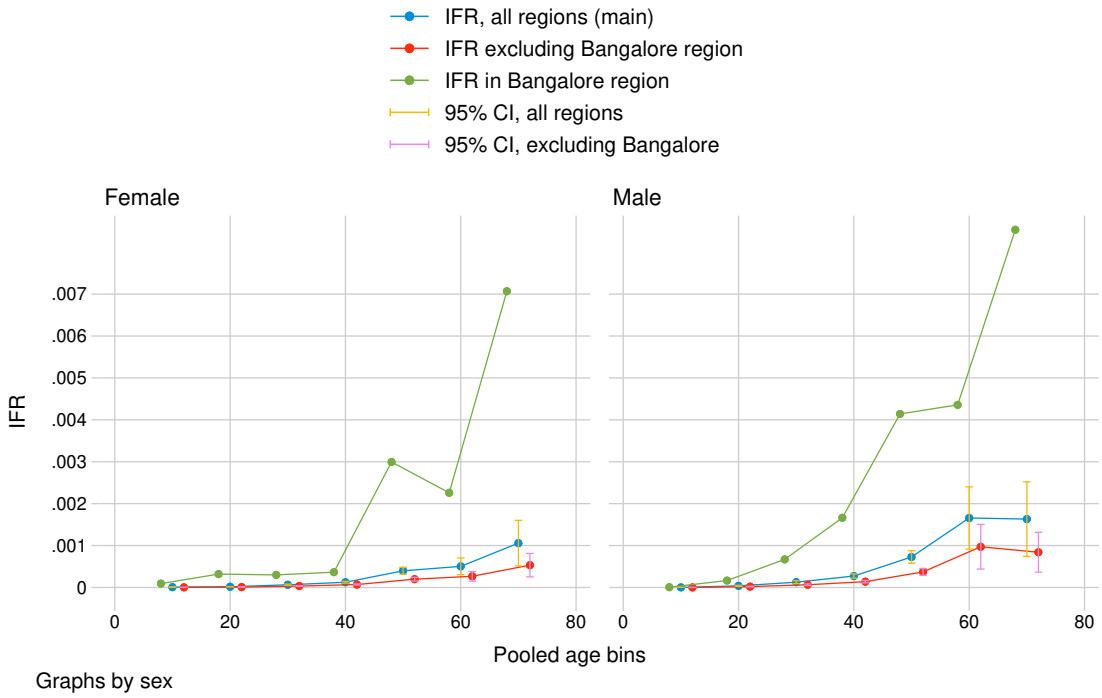
Karnataka: sensitivity analysis using piecewise cubic Hermite interpolation to estimate age bin share of deaths



Government reports provide age-shares of deaths in age bins of the form 11-20, 21-30, etc. To match seroprevalence estimates, we interpolate age-shares of deaths in the form 10-19, 20-29, etc. Main specification uses the inverse distance weighted average (IDW) to interpolate age shares. sensitivity analysis uses piecewise cubic Hermite interpolation. Interpolation was done with Stata package mipolate. Graphs by sex with 95% confidence intervals. Standard errors reflect propagation of error from design-based uncertainty of seroprevalence estimates. IFRs are calculated in age bins 0-9, ... 60-69, and 70+.

eFigure 8

Karnataka: sensitivity analysis isolating Bangalore from other sampled regions

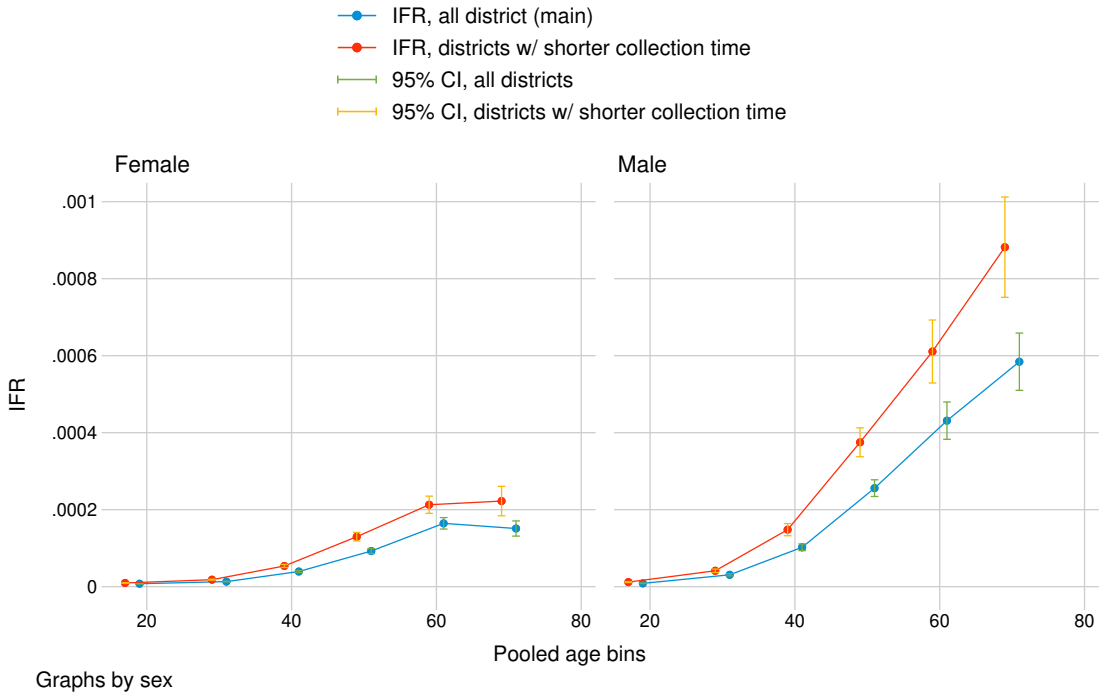


IFRs in main specification are calculated by pooling seroprevalence and death estimates from all five sampled regions of Karnataka. IFRs excluding Bangalore pool from the four remaining regions. Graphs by sex with 95% confidence intervals. Standard errors reflect propagation of error from design-based uncertainty of seroprevalence estimates. Confidence intervals are not reported for Bangalore due to small sample size, and age-specific estimated IFRs in Bangalore should not be interpreted as conclusive. IFRs are calculated in age bins 0-9, ... 60-69, and 70+.

eFigure 9

Tamil Nadu:

sensitivity analysis excluding districts where sample collection duration exceeded 3 weeks



IFRs in main specification are calculated by pooling seroprevalence and death estimates from all 37 districts of Tamil Nadu. IFRs estimated from districts with shorter collection time exclude 6 districts where seroprevalence surveying lasted longer than three weeks. Graphs by sex with 95% confidence intervals. Standard errors reflect propagation of error from uncertainty in estimating positive test rate by HUD-age-sex group. IFRs are calculated in age bins 18-29, 30-39 ... 60-69, and 70+.

2 References

References

[1] Deshingkar Priya, Khandelwal Rajiv, Farrington John. Support for migrant workers: The missing link in India's development tech. rep.Overseas Development Institute 2008.

[2] Malani Anup, Shah Daksha, Kang Gagandeep, et al. Seroprevalence of SARS-CoV-2 in slums versus non-slums in Mumbai, India *The Lancet Global Health.* 2020;0.

[3] USFDA U.S. Food and Drug Administration. Serology Test Evaluation Report for "Architect i1000 SARS-CoV-2 IgG" from Abbott tech. rep. 2020.

[4] Bryan Andrew, Pepper Gregory, Wener Mark H., et al. Performance Characteristics of the Abbott Architect SARS-CoV-2 IgG Assay and Seroprevalence in Boise, Idaho *Journal of Clinical Microbiology.* 2020;58.

[5] Rogan W. J., Gladen B.. Estimating prevalence from the results of a screening test *American Journal of Epidemiology.* 1978;107:71–76.

[6] Linton Natalie M., Kobayashi Tetsuro, Yang Yichi, et al. Incubation Period and Other Epidemiological Characteristics of 2019 Novel Coronavirus Infections with Right Truncation: A Statistical Analysis of Publicly Available Case Data *Journal of Clinical Medicine.* 2020;9.

[7] Stringhini Silvia, Wisniak Ania, Piumatti Giovanni, et al. Seroprevalence of anti-SARS-CoV-2 IgG antibodies in Geneva, Switzerland (SEROCoV-POP): a population-based study *The Lancet.* 2020;396:313–319.

[8] Laxminarayan Ramanan, Wahl Brian, Dudala Shankar Reddy, et al. Epidemiology and transmission dynamics of COVID-19 in two Indian states *Science.* 2020;370:691–697.

[9] Lewnard Joseph A., Liu Vincent X., Jackson Michael L., et al. Incidence, clinical outcomes, and transmission dynamics of severe coronavirus disease 2019 in California and Washington: prospective cohort study *BMJ.* 2020;369.

[10] Debroy Sumitra. Men account for 65% of Mumbai's 6,000+ Covid deaths, CFR high at 6.1%: BMC *The Times of India.* 2020.

[11] Chaudhuri Susmita, Thiruvengadam Ramachandran, Chattopadhyay Souvick, et al. Comparative evaluation of SARS-CoV-2 IgG assays in India *Journal of Clinical Virology.* 2020;131.

[12] Ltd. Shenzhen YHLO Biotech No.. Customer Notification: Sensitivity and Specificity of iFlash-SARS-CoV-2 and IgM kits from Clinical Trials. tech. rep. 2020.

[13] Plebani Mario, Padoan Andrea, Negrini Davide, Carpinteri Benedetta, Sciacovelli Laura. Diagnostic performances and thresholds: The key to harmonization in serological SARS-CoV-2 assays? *Clinica Chimica Acta; International Journal of Clinical Chemistry.* 2020;509:1–7.

[14] Theel Elitza S., Harring Julie, Hilgart Heather, Granger Dane. Performance Characteristics of Four High-Throughput Immunoassays for Detection of IgG Antibodies against SARS-CoV-2 *Journal of Clinical Microbiology.* 2020;58.

[15] Ripperger Tyler J., Uhrlaub Jennifer L., Watanabe Makiko, et al. Orthogonal SARS-CoV-2 Serological Assays Enable Surveillance of Low-Prevalence Communities and Reveal Durable Humoral Immunity *Immunity.* 2020;53:925–933.e4.

[16] Ibarrondo F. Javier, Fulcher Jennifer A., Goodman-Meza David, et al. Rapid Decay of Anti–SARS-CoV-2 Antibodies in Persons with Mild Covid-19 *New England Journal of Medicine.* 2020;383. Publisher: Massachusetts Medical Society pages = 1085–1087.

[17] Long Quan-Xin, Tang Xiao-Jun, Shi Qiu-Lin, et al. Clinical and immunological assessment of asymptomatic SARS-CoV-2 infections *Nature Medicine.* 2020;26:1200–1204.

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

	Item No	Recommendation
Title and abstract	1	(*a*) Indicate the study's design with a commonly used term in the title or the abstract **pg. 2**
		(*b*) Provide in the abstract an informative and balanced summary of what was done and what was found **pg. 2, 3**
Introduction		
Background/rationale	2	Explain the scientific background and rationale for the investigation being reported **pg. 5**
Objectives	3	State specific objectives, including any prespecified hypotheses **pg. 6**
Methods		
Study design	4	Present key elements of study design early in the paper **pg. 6**
Setting	5	Describe the setting, locations, and relevant dates, including periods of recruitment, exposure, follow-up, and data collection **pg. 6-8**
Participants	6	(*a*) Give the eligibility criteria, and the sources and methods of selection of participants **pg. 7-8, supplement**
Variables	7	Clearly define all outcomes, exposures, predictors, potential confounders, and effect modifiers. Give diagnostic criteria, if applicable **pg. 2, 6-9**
Data sources/ measurement	8*	For each variable of interest, give sources of data and details of methods of assessment (measurement). Describe comparability of assessment methods if there is more than one group **pg. 7-8, supplement**
Bias	9	Describe any efforts to address potential sources of bias **pg. 8-11, supplement**
Study size	10	Explain how the study size was arrived at **pg. 7-8**
Quantitative variables	11	Explain how quantitative variables were handled in the analyses. If applicable, describe which groupings were chosen and why **pg. 8-11 supplement**
Statistical methods	12	(*a*) Describe all statistical methods, including those used to control for confounding **pg. 8-11, supplement**
		(*b*) Describe any methods used to examine subgroups and interactions **pg 8-11**
		(*c*) Explain how missing data were addressed **pg. 9, 10, supplement**
		(*d*) If applicable, describe analytical methods taking account of sampling strategy **pg. 9, 10 supplement**
		(*e*) Describe any sensitivity analyses **pg. 9, 10, supplement**
Results		
Participants	13*	(a) Report numbers of individuals at each stage of study—eg numbers potentially eligible, examined for eligibility, confirmed eligible, included in the study, completing follow-up, and analysed **pg. 7, 8**
		(b) Give reasons for non-participation at each stage **pg. 7, 8, 10**
		(c) Consider use of a flow diagram **NA**
Descriptive data	14*	(a) Give characteristics of study participants (eg demographic, clinical, social) and information on exposures and potential confounders **pg. 6-7, 15**
		(b) Indicate number of participants with missing data for each variable of interest **pg. 7, 8, 10**
Outcome data	15*	Report numbers of outcome events or summary measures **NA**
Main results	16	(*a*) Give unadjusted estimates and, if applicable, confounder-adjusted estimates and their precision (eg, 95% confidence interval). Make clear which confounders were adjusted for and why they were included **pg. 11, 12**

		(*b*) Report category boundaries when continuous variables were categorized **NA**
		(*c*) If relevant, consider translating estimates of relative risk into absolute risk for a meaningful time period **NA**
Other analyses	17	Report other analyses done—eg analyses of subgroups and interactions, and sensitivity analyses **pg. 9, 10, 12, figure 2, supplement**
Discussion		
Key results	18	Summarise key results with reference to study objectives **pg. 13, 15**
Limitations	19	Discuss limitations of the study, taking into account sources of potential bias or imprecision. Discuss both direction and magnitude of any potential bias **pg. 13, 14**
Interpretation	20	Give a cautious overall interpretation of results considering objectives, limitations, multiplicity of analyses, results from similar studies, and other relevant evidence **pg. 14, 15**
Generalisability	21	Discuss the generalisability (external validity) of the study results **pg. 10, 15**
Other information		
Funding	22	Give the source of funding and the role of the funders for the present study and, if applicable, for the original study on which the present article is based **pg. 19**

*Give information separately for exposed and unexposed groups.

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

BMJ Open

Representative estimates of covid-19 infection fatality rates from four locations in India: cross-sectional study

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2021-050920.R1
Article Type:	Original research
Date Submitted by the Author:	26-Jul-2021
Complete List of Authors:	Cai, Rebecca; Development Data Lab Novosad, Paul; Dartmouth College, Economics Tandel, Vaidehi; Mercatus Center at George Mason University, Affiliate Scholar Asher, Sam; Johns Hopkins University School of Advanced International Studies, Economics Malani, Anup; University of Chicago Law School
Primary Subject Heading:	Epidemiology
Secondary Subject Heading:	Global health
Keywords:	COVID-19, Epidemiology < INFECTIOUS DISEASES, Public health < INFECTIOUS DISEASES

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

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

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

Title: Representative estimates of covid-19 infection fatality rates from four locations in India: cross-sectional study

Authors: R. Cai[1], P. Novosad[2], V. Tandel[3], S. Asher[4], A. Malani[5,*]

Affiliations:
[1] B.A. (0000-0003-2295-0667)
Research Associate
Development Data Lab
1807 California St. NW, #106
Washington D.C., USA.
cai@devdatalab.org

[2] Ph.D. (0000-0001-8331-6630)
Associate Professor of Economics
Department of Economics
Dartmouth College 311Rockefeller Hall
Hanover, NH 03755
paul.novosad@dartmouth.edu

[3] Ph.D (0000-0001-6322-0068)
Affiliate Scholar
Mercatus Center at George Mason University
3434 Washington Blvd. 4th Floor
Arlington, VA 22201
vaidehi.tandel@gmail.com

[4] Ph.D. (0000-0002-2584-8011)
Assistant Professor of Economics
Department of Economics
Johns Hopkins University School of Advanced International Studies
Wyman Park Building 5th Floor, 3100 Wyman Park Drive
Baltimore, MD 21211
sasher2@jhu.edu

Corresponding author:
[5] Ph.D, J.D. (0000-0002-2594-5778)
Lee and Brena Freeman Professor of Law
University of Chicago Law School
University of Chicago Pritzker School of Medicine
1111 E. 60th Street
Chicago, IL 60637
*Correspondence to: amalani@uchicago.edu

Manuscript Word Count: 3,673

Structured Abstract

Objectives: To estimate age- and sex-specific mortality risk among all SARS-CoV-2 infections in four settings in India, a major lower-middle-income country, and to compare age trends in mortality with similar estimates in high-income countries.

Design: Cross-sectional study.

Setting: India, multiple regions representing combined population >150 million.

Participants: Aggregate infection counts were drawn from four large population-representative prevalence/seroprevalence surveys. Data on corresponding number of deaths were drawn from official government reports of confirmed SARS-CoV-2 deaths.

Primary and secondary outcome measures: The primary outcome was age- and sex-specific infection fatality rate (IFR), estimated as the number of confirmed deaths per infection. The secondary outcome was the slope of the IFR-by-age function, representing increased risk associated with age.

Results: Among males aged 50–89, measured IFR was 0.12% in Karnataka (95 % CI: 0.09%, 0.15%), 0.42% in Tamil Nadu (0.39%, 0.45%), 0.53% in Mumbai (0.52%, 0.54%), and an imprecise 5.64% (0, 11.16%) among migrants returning to Bihar. Estimated IFR was approximately twice as high for males as for females, heterogeneous across contexts, and rose less dramatically at older ages compared to similar studies in high-income countries.

Conclusions: Estimated age-specific IFRs during the first wave varied substantially across India. While estimated IFRs in Mumbai, Karnataka, and Tamil Nadu were considerably lower than comparable estimates from high-income countries, adjustment for under-reporting based on

crude estimates of excess mortality puts them almost exactly equal with higher-income country benchmarks. In a marginalized migrant population, estimated IFRs were much higher than in other contexts around the world. Estimated IFRs suggest that the elderly in India are at an advantage relative to peers in high-income countries. Our findings suggest that the standard estimation approach may substantially underestimate IFR in low-income settings due to under-reporting of COVID deaths, and that COVID-19 IFRs may be similar in low- and high-income settings.

Article Summary

Strengths and limitations of this study

- This study provides representative estimates of the age-specific COVID-19 infection fatality rate (IFR) in four socio-economically diverse regions of India, a major lower-middle-income country, using the standard method for estimating IFR.
- Due to high measurement cost, there are very few age-specific IFR estimates in low- and middle-income countries (LMIC), despite concerns that LMIC are more vulnerable and plausibly have different mortality patterns.
- This study utilizes the primary method of estimating IFR in settings around the world, combining population-representative prevalence/seroprevalence surveys with official death reports, allowing direct methodological comparison with dozens of similar estimates from high-income countries.
- We provide population-representative estimates for over 150 million people using the largest sample to date in a low- or middle-income country, and the first documentation of IFR among the large, highly vulnerable population of migrant workers.
- The main limitation is our reliance on official reports of confirmed COVID-19 deaths, which, due to under-reporting and under-testing, likely underestimate the true number of deaths.

Introduction

Measuring the infection fatality rate (IFR) for severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) has been a major research objective since the beginning of the global pandemic. Reliable IFR estimates are essential for policy decisions on non-pharmaceutical interventions and vaccine allocation,[1–3], and comparison of waves and variants. IFR estimates almost universally rely upon large-scale seroprevalence samples drawn from the general population, matched to official death data. Because of these data requirements, the vast majority of age-specific IFR estimates are based on data from high-income countries (HICs);[2–6] meta-analyses estimating age-specific IFR in low- and middle-income countries (LMICs)[7,8] rely on untested assumptions that key epidemiological characteristics (*e.g.* transmission dynamics, age-specific death rate) in HICs are generalizable to low-income settings. Studies measuring IFR in LMICs mostly report age-aggregated IFR,[9–13] which are difficult to compare across contexts; the age pattern of infection may vary and aggregate IFRs skew higher where older people contract a larger share of infections. Estimates of age-specific IFR in LMICs have only been made from small or non-representative samples.[14,15]

Early modelers of lower-income settings warned that IFRs could be higher, due to worse baseline population health and under-resourced healthcare systems.[8,15,16] Other researchers observed low case fatality rates (CFRs) in Sub-Saharan Africa and proposed that vaccination, past infection history, and effective mitigation strategies might have reduced mortality.[17,18] The age pattern of deaths in lower-income countries has skewed younger than in high-income countries, more so than can be explained by age distribution alone.[19–21]

We calculated age-specific IFRs from four samples in India representing a combined population exceeding 150 million. We used population-representative seroprevalence surveys in

the city of Mumbai (N~=7000, population 12.5 million) and in the states of Karnataka (N~=1200, population 61 million) and Tamil Nadu (N~=26000, population 71 million). By matching these surveys to age-specific administrative death data, we calculated IFR without relying on non-representative testing data. Additionally we drew on a survey of COVID-19 prevalence among randomly sampled short-term outmigrants (N ~=4000 infections, population minimum 10 million), mostly working-age males, returning home to the state of Bihar with mortality followup. Because these migrants were randomly sampled and tracked until recovery or death, the death rate among those who tested positive is interpretable as an IFR.

Our objective was to calculate age-specific IFRs in four locations and compare them to international estimates, which are based mostly on high-income countries. We further examined heterogeneity of IFR within India and by age and sex.

Importantly, data collection took place during India's first wave of COVID-19 between March-December, 2020. India has since undergone a second, more severe wave between March-June, 2021, characterized by much higher case counts, new and potentially more transmissible variants, and a health system crisis.[22] Excess mortality and reports suggest more severe infections and higher mortality in the second wave.[22] Our IFR estimates apply to the first wave, and should not be interpreted as representative for the second.

Methods

We studied three states and one mega-city with disparate demographic and health characteristics (Table 1). Qualitatively, Tamil Nadu and Karnataka are large, relatively wealthy, southern Indian states. Mumbai is India's most populous city, and the capital of the western state Maharashtra. Tamil Nadu, Karnataka, and Maharashtra have relatively robust health care

infrastructure and vital registration.[23] In contrast, the northern state Bihar is one of the poorest in India, with the lowest stock of hospital beds per capita.[24]

The Bihar sample is limited to a sub-population of returning migrants, primarily young male laborers who lost work opportunities during lockdown. The returning migrants to Bihar are part of a large population of internal labor migrants in India; a conservative estimate from the 2001 Census found that nearly 30 million workers migrated within India for employment.[25] Tens of millions of migrants exited cities immediately after lockdown, including 6.3 million travelling on specially designated trains ("Shramik Specials") between May-August, 2020.[26,27] Short-term migrants were on average very poor even before the pandemic.[28] India's sudden lockdown left them unemployed, and many experienced extreme physical and economic duress on the long journey home.[29,30]

India began its first nation-wide lockdown on March 24, 2020, and by July 2021 had the second-highest number of country-wide confirmed COVID-19 cases in the world. The Indian government spends roughly 1.5% of GDP on healthcare, one of the world's lowest rates.[31] Discussion of India's COVID-19 preparedness has focused on under-resourced public hospitals, a largely unregulated private healthcare sector, and fear and stigma among the public surrounding infection.[31]

Data sources and study design

In Mumbai, Karnataka, and Tamil Nadu, we matched representative seroprevalence surveys to administrative reports of confirmed COVID-19 deaths.

In Mumbai, seroprevalence surveys were conducted for two weeks in July 2020 with representative sampling of three wards, one from each of the city's three zones, stratified by age,

sex, and slum/non-slum dwellers.[10] Enumerators sought voluntary consent to sample one member per household, rotating through age-gender groups. Thus, the sample composition is representative for city-wide age and sex, subject to consent rates. The sample consisted of 6,904 participants (4,202 from slums and 2,702 from non-slums), tested for IgG antibodies to the SARS-CoV-2 N-protein using the Abbott Diagnostics Architect™ test. Data on cumulative deaths were collected from daily reports from the municipal governing body.

In Karnataka, seroprevalence surveys were conducted from June 15 to August 29, 2020, in representative samples of urban and rural areas in 20 out of 30 districts, stratified to generalize to five regions spanning all districts.[32] We can therefore take the ELISA positive test rate as an unbiased measure of region-level positivity rate. The sampling frame was not age- or sex-stratified, and older individuals were over-sampled relative to population age composition. We assume that ELISA positive test rate is representative by age-sex-region group, because there was no evidence that the age of the consenting member of each household was associated with seropositivity in the home. 1,196 participants were tested with an ELISA for antibodies to the receptor binding domain of the SARS-CoV-2 virus, developed by Translational Health Science and Technology Institute in India. We collected district-level death data from the Government of Karnataka Department of Health and Family Welfare bulletins.

In Tamil Nadu, a representative seroprevalence survey was conducted between October 19 to November 30, 2020, of adults aged 18 and older, covering the state's 37 districts.[33] Collection times within districts were often significantly shorter. Enumerators divided districts into health unit districts, then randomly sampled urban and rural clusters. Within clusters, enumerators started at a randomly selected GPS starting point, sampling one person from households adjacent to the starting point (using the Kish method) to provide a biosample.

Because household members were selected randomly, we similarly assume seropositivity is representative at the age-sex-district level. Seropositivity was tested using either the iFlash-SARS-CoV-2 IgG or the Vitros anti-SARS-CoV-2 IgG CLIA kit. The analytical subsample was 26,107 antibody tests that could be conclusively determined as positive or negative. Case-level data on 12,019 recorded state-wide COVID-19 deaths, from March to December, 2020, was collected from daily government reports.

In Bihar, the state government began COVID-19 testing among returning out-of-state migrants soon after the first positive case was identified in a migrant on March 22, 2020. On May 4, Bihar began to randomly select migrants for testing. Random testing continued until July 21, though for a brief window (May 22–31) only migrants returning from seven major cities were sampled. We isolated the subsample of randomly selected migrants, yielding 4,362 individuals with positive tests.[29] Tests were conducted with TrueNat machines manufactured by MolBio Diagnostics in Goa, with positive tests confirmed by real-time PCR kits.[34] Bihar attempted to track all migrants who tested positive until they eventually recovered or died.

In all locations, population data came from the 2012 Socio-Economic and Caste Census.

Statistical analysis

In Mumbai, Karnataka, and Tamil Nadu, we estimated infection counts from representative seroprevalence surveys. Methods for estimating infection counts are described in detail below. We matched infection counts to deaths assuming that the infection-seroconversion delay is on average two days shorter than the infection-death delay.[35,36] To implement this, we calculated IFR as the cumulative number of deaths reported as of two days after the end of seroprevalence testing, divided by the number of infections. Testing sensitivity to this

assumption, we replicate results using deaths from one and two weeks after last day of seroprevalence testing, effectively generating upper bounds for the number of deaths (eFigures 1-3 in the Supplement). Where multiple evaluations of the antibody tests' sensitivity/specificity existed, we tested robustness to assuming minimum sensitivity (eFigures 4 and 5 in the Supplement).

In Mumbai, we first adjusted for test sensitivity and specificity using the Rogan-Gladen correction,[37] then calculated aggregate seroprevalence for each sampled ward and multiplied by ward population to estimate infection count. We estimated infection counts in non-sampled wards by assuming a constant rate of government under-reporting in wards in the same zone. This approach was supported by very similar case-to-seroprevalence ratios in the three wards with seroprevalence data (eTable 1). Age- and sex-specific infection shares were based on the seroprevalence survey (eFigure 6).

In Karnataka we adjusted for test inaccuracies,[37] then used census population counts to aggregate from regional to state-level infection counts, reweighting to match regional age-sex distributions. Methods for matching dates and deaths to infections is described in detail in the Supplement (eFigure 7). Because the seroprevalence survey period in Bangalore spanned two months (compared with less than three weeks in the other regions), we show results excluding Bangalore, where deaths may have been overestimated due to the longer survey period (eFigure 8).

In Tamil Nadu, we first calculated the population-representative seropositivity rate by district-age-sex group and type of test kit, then adjusted for test inaccuracies. We estimated the number of state-wide infections per district-age-sex group by combining kit-specific

seroprevalence estimates and multiplying by population, then summing across districts. In sensitivity checks, we re-estimated IFR limiting samples to districts where seroprevalence surveillance lasted less than three weeks (eTable 2, eFigure 9).

In Bihar, although enumerators attempted to track outcomes for all migrants, 1,530 (35%) infected individuals could not be tracked. In main estimates, we assumed that their fatality rates were the same as successfully tracked individuals; in sensitivity checks, we considered the possibility that all survived. High attrition is common in studies of migrant workers,[29] with followup in this case complicated by the ongoing crisis. We limited our analytic sample to 3,921 randomly-sampled male migrants, for whom 2,536 outcomes are known.

Information on underlying sample size, seroprevalence rate, and number of deaths used to calculate IFRs in each location are in eTables 3–6 in the Supplement.

Matching representative seroprevalence surveys to administrative death data is the primary method of IFR measurement everywhere in the world.[2,4,5] In Bihar, because migrants were randomly sampled, there was no selection on symptomatic or severe cases, and mortality rates among positive cases can be interpreted as IFRs. As noted above, short-term migrants from Bihar are economically marginalized; their IFRs can be understood as representative for migrants, but not necessarily the general population.

We calculated IFRs in 10-year age bins, plus bins 10–49 and 50–89, in all locations. We used two large-scale meta-analyses[1,7] of age-specific SARS-CoV-2 IFRs as reference groups. Both Levin et al.[1] and O'Driscoll et al.[7] draw almost exclusively from seroprevalence samples from Europe and the United States. The application of these samples to mortality in LMICs (as in O'Driscoll et al.[7]) requires the as-yet untested assumption that multiple

epidemiological factors (*e.g.*, transmission dynamics) are uniform between HIC and LMIC. Levin et al.[1] do not report IFR by sex; we estimated sex-specific IFRs in Levin et al. by assuming the same sex ratio in IFR as reported in O'Driscoll et al. For the larger age bins, we weighted age-specific IFR estimates from sample populations and meta-analyses by the Indian national population distribution, to ensure differences across contexts were driven by differences in age-specific IFRs, rather than population age distribution.

We calculated the slope of the natural log of IFR as a function of age by fitting a linear function to the most granular age-specific IFR data that could be obtained in each location. Additional details on the underlying samples and the methodology are in the Supplementary Materials. All analyses were conducted in Stata 16.0.

Patient and public involvement

No patients were directly involved in this study. Patients would not be able to identify themselves in the data.

Ethics statement

Because there were no patients or human subjects, the study was exempt from ethics committee approval. There was no direct data collection for this study; all data were gathered secondhand from public or published sources. The data used for measuring seroprevalence, COVID-19 deaths, and population were all anonymized and aggregated before we accessed it. We retrieved seroprevalence rate data in all locations from public sources, aggregated by age and sex[1].[10,29,33,38] Seroprevalence studies were designed and implemented in partnership with

[1] Details on public sources for seroprevalence data. Bihar migrant data may be requested from the Government of Bihar. Positive test rates by age, gender, ward, and slum in Mumbai can be found in the online supplement of [10]. The same rates by district in Tamil Nadu can be found in the online supplement of [33]. The same rates by region in Karnataka can be found in the supplement of [38].

local city and state governments. Details of patient involvement, protocols, and institutional ethics approval for each seroprevalence study have been published in separate papers, and in reports from the respective governments.[10,29,32,33]

Results

We plotted age-specific IFR for each location on a log scale, to enable comparison at all ages despite exponential increases at higher ages found in all countries (Figures 1a and 1b). For both males and females, there is substantial variation in IFR across the four locations in India. In Karnataka, age-specific IFRs are 10 times lower than those reported in the meta-analyses, and 25 times lower over age 70. In Tamil Nadu, estimates were 2–4 times lower than those in the meta-analyses. In Mumbai, estimates were close to the lower of the two meta-analyses at younger ages,[7] but were considerably lower than meta-analyses after age 60. For 60–69-year-old men, for example, we measured an IFR of 0.17% (95% CI: 0.092 to 0.240) in Karnataka, 0.45% (0.397 to 0.0.497) in Tamil Nadu, and 0.62% (0.591 to 0.647) in Mumbai (Table 2); the two meta-analyses reported male IFR of 1.02%[7] and 1.86%[1] in this age group.

In contrast, mortality among male migrants returning to Bihar was an order of magnitude higher. Mortality among males aged 60–69 was extremely high but measured imprecisely due to the small sample of older males (4.26% [95% CI: 0.0 to 10.0%]). The larger age bins allowed a more precise measure of IFR in Bihar (Table 3). In both the 10–49 and 50–89 age bins, mortality in Bihar was an order of magnitude higher than in the other Indian locations and at least twice as high as rates in meta-analyses, after weighting to the Indian age distribution to ensure cross-context comparability. For the 50–89 age group, estimates were not precise enough to rule out

equality between Bihar and the other locations. For the 10–49 age group, we can rule out equality ($p < 0.01$).

To the extent that an IFR advantage exists in India, it appears more strongly among the elderly. In most cases, the overall increase in IFR with age was considerably less steep than in the reference meta-analyses (Figure 1), particularly at older ages. The meta-analyses suggest that an 80-year-old has about 100x the IFR of a 40-year-old; in Mumbai, the increase in risk factor is 40x and in Bihar it is only 10x. Specifically, male IFR increased on average by 4.7%, 9.6%, 10.3%, and 11.6% with each year of age in Bihar, Mumbai, Karnataka, and Tamil Nadu respectively. We calculated comparable figures in the meta-analyses as 11.4%[7] and 12.3%.[1] Slopes for Indian females were uniformly flatter than those for the reference groups (Figure 1b).

The main estimates are replicated in the Supplementary Materials under a range of different scenarios and assumptions; the ordering of IFRs across regions and with respect to the reference groups is highly robust (Figures 2a–d).

Discussion

Principal findings

Using best-practice methods applied in many high-income countries, we found substantial heterogeneity in age-specific COVID-19 infection fatality rate in India. In all four locations, we found a weaker increase in IFR over age than seen in other countries.

In Mumbai, Karnataka, and Tamil Nadu, estimated IFRs were considerably lower than those measured in richer countries. These results are qualified by the fact that COVID deaths are known to be under-reported in these locations, as we discuss below. In a tracked sample of male

migrants returning to Bihar, IFR estimates were an order of magnitude higher than the other two locations and twice as high as the international reference groups.

Our Mumbai IFR estimates are representative for the city while Tamil Nadu and Karnataka estimates are representative for the state. IFR estimates for migrants returning to Bihar are plausibly generalizable to the tens of millions of migrant workers who exited cities, returning primarily to poorer rural areas, in the first months of the pandemic. Migrant workers differ from the general population, typically living in dense quarters that increase disease transmission,[25] with higher poverty rates,[28] lower baseline health, and higher prevalence of malaria, respiratory infections, and acute febrile illness.[25] In these aspects, our findings on migrants have some generalizability to other extremely disadvantaged populations. However, the actual journey migrants undertook is a unique risk factor. Over-packed trains likely heightened transmission and long travel distances, often on foot, increased physical vulnerability.[27]

Strengths and weaknesses of the study

The strength of this study was the use of seroprevalence data representing over 150 million people, with a sufficiently large sample to calculate age-disaggregated IFR in a lower middle-income country. The main weakness of the study is that, like all COVID-19 population estimates, our results depend on the quality of underlying mortality data. The largest potential source of bias was our use of official reports of COVID-19 deaths, which undercount the true number of deaths in all contexts.[23,39]

Though estimates of under-reporting are highly uncertain, accounting for misreporting of deaths brings IFRs in three of the study locations close to estimates from high-income countries. Focusing on the 50–89 age group, in Mumbai, a doubling of COVID-19 deaths is required to put

estimated IFR in the range of the meta-analyses. It is plausible that deaths in Mumbai were undercounted by a factor of 2; between March and July, Mumbai recorded 6,600 excess deaths in addition to the 6,400 COVID-19 deaths used in this study.[39]

In Karnataka and Tamil Nadu, COVID-19 deaths would have to be under-reported by factors of 10 and 3 respectively to bring IFR in line with international estimates. Crude estimates from recently published data from India's Civil Registration System suggest excess mortality rates during the first COVID-19 wave were approximately six times higher than official COVID deaths in both Karnataka and Tamil Nadu.[40] If this ratio between excess mortality and reported COVID-19 deaths is an accurate measure of the death under-reporting rate, then this puts IFRs in Mumbai and Tamil Nadu close to the range of the high-income country results, and Karnataka only slightly lower.

While these IFR estimates remain subject to bias, we note that we calculated IFR with the standard methodology used in many cross-national settings, many of which are also characterized by under-reporting of COVID-19 deaths. As described in the Supplement, wherever possible we made conservative choices that would bias our IFR estimates upward rather than downward. In particular, antibodies may fade over time, so seroprevalence tests provide a lower bound on the cumulative infection rate.[41]

Official misreporting of COVID-19 deaths would not bias our IFR estimates in Bihar, due to the mortality followup methodology underlying these estimates. For our Bihar estimates to match the range of meta-analyses, deaths would need to have been *overcounted* by a factor of 2 for ages 50–89, and by 10 for ages 10–49. However, we do not know the base rate of migrant death. If migrant deaths would be high in absence of COVID-19, due to migrants' arduous return journeys, we may overstate the mortality attributable to COVID-19 in this group.

Comparison with other studies

Few other studies have utilized sufficiently large seroprevalence samples to estimate age-specific IFR for a large lower-income population. Seroprevalence-based IFR estimates for older individuals in a Brazilian city[14] were slightly lower than our estimate for Bihari migrants, and much higher than our seroprevalence-based estimates. However, seroprevalence samples of non-representative groups in Sub-Saharan Africa implied high infection rates, suggesting either low overall mortality or substantial under-reporting of deaths, consistent with our findings in India.[11,17,42]

Studies have noted that the pattern of mortality in low- and middle-income countries skews younger than would be predicted from the age distributions of death in high-income countries.[19,21] Our study suggests that a flatter age profile in infection fatality rates in lower-income settings could be a major factor driving this difference.

Conclusion and further research

In large samples representing India's higher-income South, we found infection fatality rates that broadly corresponded to those reported in richer countries, after adjusting for undercounting. Among a sample of economically distressed migrants, we found IFRs that were twice as high, plausibly due to severe economic and physical distress. Migrant workers have worse health than the general population at baseline;[25,43] the circumstances at the beginning of the pandemic may have made this group exceptionally vulnerable to adverse health events following viral infection.

At the time of writing, these estimates are among the best available in a lower-income setting. Improved surveillance and accounting of SARS-CoV-2 are critical investments that

would improve our understanding of the fatality risk of the virus in lower-income settings. Further research is necessary to determine if infection fatality rates are similar in high- and low-income settings.

Figure legends:

Figure 1: Age-specific infection fatality rate, comparing four locations in India with international estimates. Point estimates of age-specific infection fatality rate in (a) males and (b) females combining representative prevalence/seroprevalence studies and government-reported COVID-19 deaths. IFRs were estimated for age bins 10–19 (Mumbai and Karnataka only), 20–29,…,60–69, and 70+ in India. Slope of IFR age trends from the meta-analyses calculated by fitting a linear regression between age and natural log of IFR.

Figure 2: Age-specific infection fatality rates in India: sensitivity checks. Main estimates and sensitivity checks of infection fatality rate of (a) males aged 10-49 years, (b) males aged 50-89, (c) females aged 10-49, and (d) females aged 50-89. 95% confidence intervals shown in grey. In all locations, including meta-analyses, age-specific IFRs in smaller age bins have been weighted to India's national age distribution, controlling for cross-location differences in population age. See Supplement for details of sensitivity checks.

Table 1: Health and demographic context of sample locations

	Median age	**GDP/capita**	**Cumulative infections on July 31**	**Cumulative COVID-19 deaths on July 31**	**Hospital beds per 100,000 population.**
	Population Census 2011	NSDP nominal (2018-19 INT$)	JHU CSSE Covid-19 data[44]	JHU CSSE Covid-19 data[44]	Kapoor et al.[24]
Bihar	19.9	640	51,233	296	25.55
Maharashtra*	28.2	2,802	411,798	14,994	172.94
Karnataka	27.4	3,082	124,115	2,314	391.62
Tamil Nadu	29.9	2,831	245,859	3,935	174.83
India	24.0	1,964	1,695,988	36,511	137.62

Row 2 indicates the data source. *Mumbai is the capital city of Maharashtra.

Table 2: Age-specific infection fatality rates (%) from four locations in India

	Mumbai	Mumbai	Karnataka	Karnataka	Tamil Nadu	Tamil Nadu	Bihar migrants
Age	Male	Female	Male	Female	Male	Female	Male
10-19	0.004 (0.004,0.004)	0.001 (0.001,0.002)	0.000 (0.000,0.000)	0.001 (0.000,0.001)	NA	NA	0.000 (0.000,0.000)
20-29	0.013 (0.012,0.013)	0.005 (0.005,0.005)	0.004 (0.002,0.005)	0.002 (0.001,0.002)	0.003 (0.003,0.004)	0.003 (0.003,0.003)	0.649 (0.131,1.166)
30-39	0.041 (0.039,0.043)	0.019 (0.018,0.020)	0.013 (0.009,0.016)	0.006 (0.005,0.008)	0.013 (0.012,0.014)	0.005 (0.005,0.006)	1.810 (0.795,2.825)
40-49	0.112 (0.106,0.117)	0.058 (0.055,0.061)	0.027 (0.022,0.032)	0.012 (0.010,0.015)	0.050 (0.046,0.054)	0.020 (0.018,0.021)	1.529 (0.199,2.859)
50-59	0.355 (0.339,0.372)	0.172 (0.163,0.180)	0.073 (0.058,0.088)	0.040 (0.032,0.048)	0.177 (0.162,0.192)	0.066 (0.061,0.071)	2.381 (0.000,5.043)
60-69	0.619 (0.589,0.649)	0.317 (0.301,0.332)	0.166 (0.092,0.240)	0.050 (0.030,0.070)	0.447 (0.397,0.497)	0.172 (0.157,0.188)	4.255 (0.000,10.026)
70-89	0.837 (0.797,0.878)	0.511 (0.486,0.536)	0.163 (0.074,0.252)	0.106 (0.051,0.160)	1.024 (0.894,1.155)	0.267 (0.233,0.301)	12.500 (0.000,35.418)

Infection fatality rates as percentages. 95% confidence intervals in parentheses.

Table 3: Age-specific infection fatality rates in India ages 10—49 and 50—89

	Mumbai	Mumbai	Karnataka	Karnataka	Tamil Nadu	Tamil Nadu	Bihar migrants
Age	Male	Female	Male	Female	Male	Female	Male
10-49	0.033 (0.032,0.034)	0.016 (0.016,0.017)	0.009 0.007,0.010)	0.004 (0.004,0.005)	0.013* (0.013,0.014)	0.006* (0.006,0.007)	0.851 (0.467,1.235)
50-89	0.530 (0.517,0.543)	0.285 (0.278,0.292)	0.120 (0.090,0.150)	0.056 (0.043,0.069)	0.420 (0.390,0.459)	0.140 (0.132,0.149)	5.393 (0.000,11.156)

Infection fatality rates as percentages. 95% confidence intervals in parentheses. *In Tamil Nadu, seroprevalence collection and deaths were restricted to adults aged 18+. The 10-49 group assumes IFR in the 10—17 and 18—29 groups were equal, for weighting purposes.

Funding: This paper was partially supported by Emergent Ventures grant #466, awarded to Malani, Asher, and Novosad. The funder of the study had no role in the following: study design; collection, analysis, management, or interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit for publication. The corresponding author had full access to all of the data, and takes responsibility for the integrity of the data and accuracy of data analysis.

Author contributions: All authors (Cai, Novosad, Asher, Tandel, and Malani) participated in idea generation and development, empirical strategy design, and manuscript development. Malani and Tandel provided data on seroprevalence and mortality, and contextual knowledge regarding government sampling schemes and mortality registration. Cai and Novosad conducted the data analysis. All authors saw and approved the final version of the manuscript. The corresponding author attests that all listed authors meet authorship criteria and that no others meeting the criteria have been omitted.

Competing interests: Authors declare no competing interests.

Data and materials sharing: Replication code, data dictionary, and data will be posted in a public repository on Github. The repository will include all data on demographics and COVID-19 deaths by location, seroprevalence aggregates for Mumbai, Karnataka, and Tamil Nadu, and mortality rates by age and gender for migrants from Bihar. We do not have permission to share seroprevalence microdata. Replication code will be provided to reconstruct all results in the paper from these data.

References:

[1] Levin AT, Hanage WP, Owusu-Boaitey N, Cochran KB, Walsh SP, Meyerowitz-Katz G. Assessing the age specificity of infection fatality rates for COVID-19: systematic review, meta-analysis, and public policy implications. Eur J Epidemiol 2020. https://doi.org/10.1007/s10654-020-00698-1.

[2] Pastor-Barriuso R, Pérez-Gómez B, Hernán MA, Pérez-Olmeda M, Yotti R, Oteo-Iglesias J, et al. Infection fatality risk for SARS-CoV-2 in community dwelling population of Spain: nationwide seroepidemiological study. BMJ 2020;371:m4509. https://doi.org/10.1136/bmj.m4509.

[3] Yang W, Kandula S, Huynh M, Greene SK, Wye GV, Li W, et al. Estimating the infection-fatality risk of SARS-CoV-2 in New York City during the spring 2020 pandemic wave: a model-based analysis. Lancet Infect Dis 2020;0. https://doi.org/10.1016/S1473-3099(20)30769-6.

[4] Perez-Saez J, Lauer SA, Kaiser L, Regard S, Delaporte E, Guessous I, et al. Serology-informed estimates of SARS-CoV-2 infection fatality risk in Geneva, Switzerland. Lancet Infect Dis 2020. https://doi.org/10.1016/S1473-3099(20)30584-3.

[5] Salje H, Kiem CT, Lefrancq N, Courtejoie N, Bosetti P, Paireau J, et al. Estimating the burden of SARS-CoV-2 in France. Science 2020;369:208–11. https://doi.org/10.1126/science.abc3517.

[6] Ward H, Atchison CJ, Whitaker M, Ainslie KEC, Elliott J, Okell LC, et al. Antibody prevalence for SARS-CoV-2 in England following first peak of the pandemic: REACT2 study in 100,000 adults. MedRxiv 2020. https://doi.org/10.1101/2020.08.12.20173690.

[7] O'Driscoll M, Ribeiro Dos Santos G, Wang L, Cummings DAT, Azman AS, Paireau J, et al. Age-specific mortality and immunity patterns of SARS-CoV-2. Nature 2020:1–6. https://doi.org/10.1038/s41586-020-2918-0.

[8] Ghisolfi S, Almås I, Sandefur JC, Carnap T von, Heitner J, Bold T. Predicted COVID-19 fatality rates based on age, sex, comorbidities and health system capacity. BMJ Glob Health 2020;5:e003094. https://doi.org/10.1136/bmjgh-2020-003094.

[9] Hallal PC, Hartwig FP, Horta BL, Silveira MF, Struchiner CJ, Vidaletti LP, et al. SARS-CoV-2 antibody prevalence in Brazil: results from two successive nationwide serological household surveys. Lancet Glob Health 2020;8:e1390–8. https://doi.org/10.1016/S2214-109X(20)30387-9.

[10] Malani A, Shah D, Kang G, Lobo GN, Shastri J, Mohanan M, et al. Seroprevalence of SARS-CoV-2 in slums versus non-slums in Mumbai, India. Lancet Glob Health 2020;0. https://doi.org/10.1016/S2214-109X(20)30467-8.

[11] Uyoga S, Adetifa IMO, Karanja HK, Nyagwange J, Tuju J, Wanjiku P, et al. Seroprevalence of anti–SARS-CoV-2 IgG antibodies in Kenyan blood donors. Science 2020. https://doi.org/10.1126/science.abe1916.

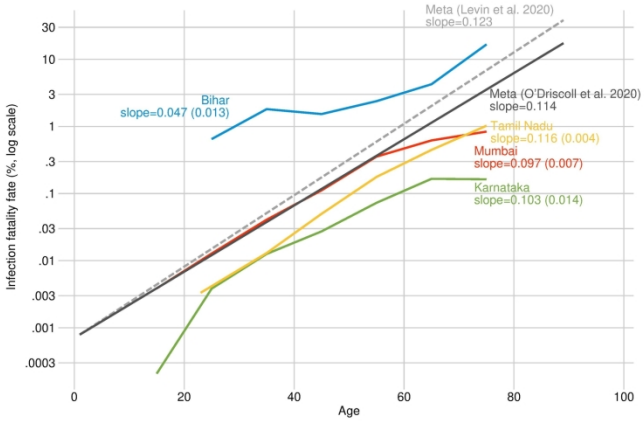
[12] Ioannidis J. The infection fatality rate of COVID-19 inferred from seroprevalence data. Bull World Health Organ 2020;BLT.20.265892.

[13] Xu X, Sun J, Nie S, Li H, Kong Y, Liang M, et al. Seroprevalence of immunoglobulin M and G antibodies against SARS-CoV-2 in China. Nat Med 2020;26:1193–5. https://doi.org/10.1038/s41591-020-0949-6.

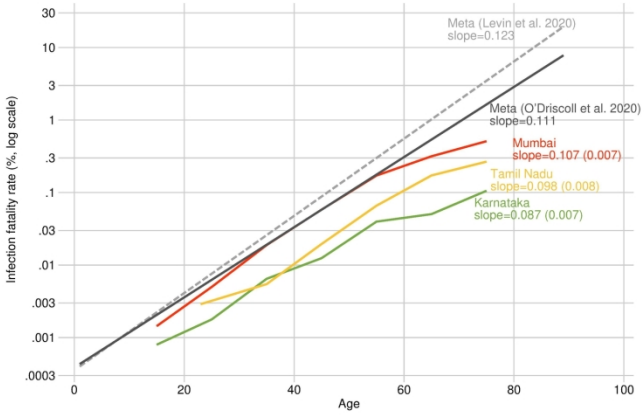
[14] Picon RV, Carreno I, da Silva AA, Mossmann M, Laste G, Domingues G de C, et al. Coronavirus disease 2019 population-based prevalence, risk factors, hospitalization, and

fatality rates in southern Brazil. Int J Infect Dis 2020;100:402–10. https://doi.org/10.1016/j.ijid.2020.09.028.
[15] Verity R, Okell LC, Dorigatti I, Winskill P, Whittaker C, Imai N, et al. Estimates of the severity of coronavirus disease 2019: a model-based analysis. Lancet Infect Dis 2020;20:669–77. https://doi.org/10.1016/S1473-3099(20)30243-7.
[16] Walker PGT, Whittaker C, Watson OJ, Baguelin M, Winskill P, Hamlet A, et al. The impact of COVID-19 and strategies for mitigation and suppression in low- and middle-income countries. Science 2020;369:413–22. https://doi.org/10.1126/science.abc0035.
[17] Nordling L. The pandemic appears to have spared Africa so far. Scientists are struggling to explain why. Sci AAAS 2020.
[18] Njenga MK, Dawa J, Nanyingi M, Gachohi J, Ngere I, Letko M, et al. Why is There Low Morbidity and Mortality of COVID-19 in Africa? Am J Trop Med Hyg 2020;103:564–9. https://doi.org/10.4269/ajtmh.20-0474.
[19] Demombynes G. COVID-19 Age-Mortality Curves Are Flatter in Developing Countries. The World Bank; 2020.
[20] Philip M, Ray D, Subramanian S. Decoding India's Low Covid-19 Case Fatality rate. National Bureau of Economic Research; 2020. https://doi.org/10.3386/w27696.
[21] Novosad P, Jain R, Campion A, Asher S. COVID-19 mortality effects of underlying health conditions in India: a modelling study. BMJ Open 2020;10. https://doi.org/10.1136/bmjopen-2020-043165.
[22] Samarasekera U. India grapples with second wave of COVID-19. Lancet Microbe 2021;2:e238. https://doi.org/10.1016/S2666-5247(21)00123-3.
[23] Chatterjee P. Is India missing COVID-19 deaths? The Lancet 2020;396:657. https://doi.org/10.1016/S0140-6736(20)31857-2.
[24] Kapoor G, Hauck S, Sriram A, Joshi J, Schueller E, Frost I, et al. State-wise estimates of current hospital beds, intensive care unit (ICU) beds and ventilators in India: Are we prepared for a surge in COVID-19 hospitalizations? MedRxiv 2020:2020.06.16.20132787. https://doi.org/10.1101/2020.06.16.20132787.
[25] Adsul BB, Laad PS, Howal PV, Chaturvedi RM. Health problems among migrant construction workers: A unique public–private partnership project. Indian J Occup Environ Med 2011;15:29–32. https://doi.org/10.4103/0019-5278.83001.
[26] Dutta A. Parliament monsoon session: 97 people died on-board Shramik trains, govt tells Rajya Sabha. Hindustan Times 2020.
[27] Gettleman J, Raj S, Yasir S, Singh KD, Loke A. The Virus Trains: How Lockdown Chaos Spread Covid-19 Across India. N Y Times 2020.
[28] Deshingkar P, Khandelwal R, Farrington J. Support for migrant workers: The missing link in India's development. Overseas Development Institute; 2008.
[29] Malani A, Mohanan M, Kumar C, Kramer J, Tandel V. Prevalence of SARS-CoV-2 among workers returning to Bihar gives snapshot of COVID across India. MedRxiv 2020. https://doi.org/10.1101/2020.06.26.20138545.
[30] Abi-Habib M, Yasir S. India's Coronavirus Lockdown Leaves Vast Numbers Stranded and Hungry. N Y Times 2020.
[31] Chetterje P. Gaps in India's preparedness for COVID-19 control. Lancet Infect Dis 2020;20:544. https://doi.org/10.1016/S1473-3099(20)30300-5.
[32] Mohanan M, Malani A, Krishnan K, Acharya A. Prevalence of SARS-CoV-2 in Karnataka, India. JAMA 2021. https://doi.org/10.1001/jama.2021.0332.

[33] Malani A, Tandel V, Parasa R, Ramachandran S, Sudharshini S, Prakash V, et al. Seroprevalence in Tamil Nadu in October-November 2020. MedRxiv n.d.
[34] CMO-PRC. Pre-release cm-243: Directions for random testing 2020.
[35] Linton NM, Kobayashi T, Yang Y, Hayashi K, Akhmetzhanov AR, Jung S-M, et al. Incubation Period and Other Epidemiological Characteristics of 2019 Novel Coronavirus Infections with Right Truncation: A Statistical Analysis of Publicly Available Case Data. J Clin Med 2020;9. https://doi.org/10.3390/jcm9020538.
[36] Stringhini S, Wisniak A, Piumatti G, Azman AS, Lauer SA, Baysson H, et al. Seroprevalence of anti-SARS-CoV-2 IgG antibodies in Geneva, Switzerland (SEROCoV-POP): a population-based study. The Lancet 2020;396:313–9. https://doi.org/10.1016/S0140-6736(20)31304-0.
[37] Rogan WJ, Gladen B. Estimating prevalence from the results of a screening test. Am J Epidemiol 1978;107:71–6. https://doi.org/10.1093/oxfordjournals.aje.a112510.
[38] Mohanan M, Malani A, Krishnan K, Acharya A. Prevalence of COVID-19 In Rural Versus Urban Areas in a Low-Income Country: Findings from a State-Wide Study in Karnataka, India. MedRxiv 2020. https://doi.org/10.1101/2020.11.02.20224782.
[39] Banaji M. Mumbai Had 13k "Excess" Deaths in March-July. How Does Its COVID Story Change? Wire Sci 2020. https://science.thewire.in/the-sciences/covid-19-mumbai-all-cause-mortality-data-ifr-bmc-seroprevalence-survey/ (accessed December 21, 2020).
[40] Asher S, Lunt T, Matsuura P, Novosad P. The Socioeconomic High-resolution Rural-Urban Geographic Dataset on India (SHRUG). World Bank Econ Rev Revise Resubmit 2020.
[41] Ripperger TJ, Uhrlaub JL, Watanabe M, Wong R, Castaneda Y, Pizzato HA, et al. Orthogonal SARS-CoV-2 Serological Assays Enable Surveillance of Low-Prevalence Communities and Reveal Durable Humoral Immunity. Immunity 2020;53:925-933.e4. https://doi.org/10.1016/j.immuni.2020.10.004.
[42] Mukherjee S. Why Does the Pandemic Seem to Be Hitting Some Countries Harder Than Others? New Yorker n.d.
[43] Choudhari R. COVID 19 pandemic: Mental health challenges of internal migrant workers of India. Asian J Psychiatry 2020;54:102254. https://doi.org/10.1016/j.ajp.2020.102254.
[44] Dong E, Du H, Gardner L. An interactive web-based dashboard to track COVID-19 in real time. Lancet Infect Dis 2020;20:533–4. https://doi.org/10.1016/S1473-3099(20)30120-1.



(a) Male



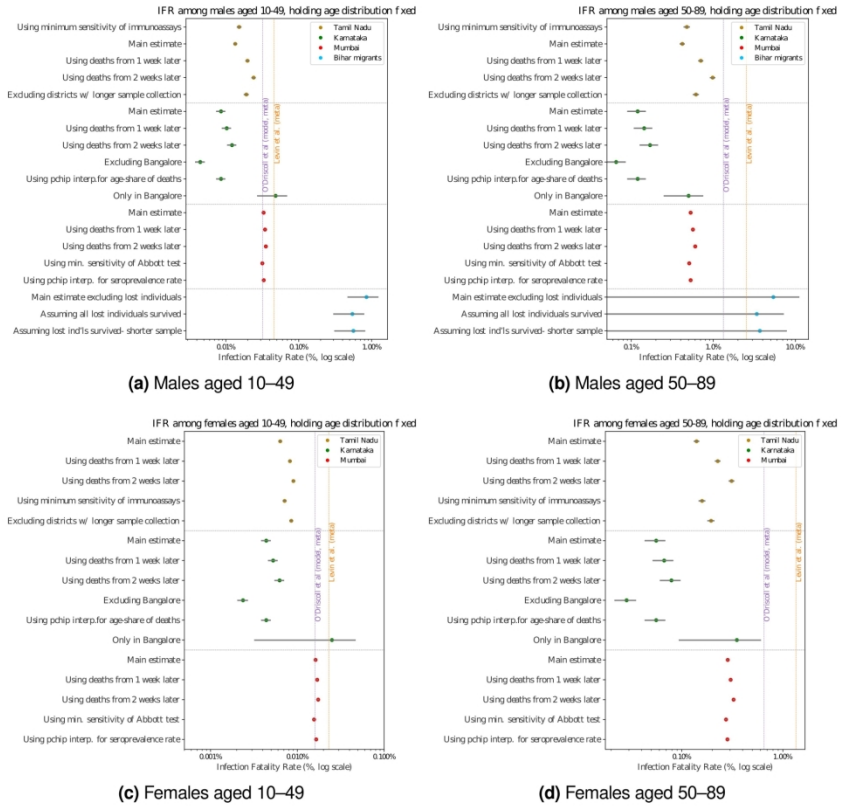
(b) Female

Figure 1
Age-specific infection fatality rate,
comparing four locations in India with international estimates

1

Age-specific infection fatality rate, comparing four locations in India with international estimates

132x239mm (300 x 300 DPI)



(a) Males aged 10–49 **(b)** Males aged 50–89

(c) Females aged 10–49 **(d)** Females aged 50–89

Figure 2
Age-specific infection fatality rates in India:
sensitivity tests

Markers indicate population-weighted pooled IFRs with India as the reference population. 95% confidence intervals are in grey.

Age-specific infection fatality rates in India: sensitivity tests

165x235mm (300 x 300 DPI)

Supplementary materials for:

Representative estimates of covid-19 infection fatality rates from four locations in India: cross-sectional study

Rebecca Cai, Paul Novosad, Vaidehi Tandel, Sam Asher, Anup Malani*
* Correspondence to: amalani@uchicago.edu

This PDF file includes:

Detailed Materials and Methods
eFigures 1 to 9
eTables 1 to 6

Materials and Methods

Bihar

Data

We made use of data on all positive cases in the state of Bihar found during random testing of incoming migrants during an early phase of the pandemic. The data was provided by the Health Department of the Government of Bihar. The data contained a sample of 4,954 active infections and their outcomes, reported between March 22 (the date on which the first positive case in Bihar was detected) and July 21, 2020. The vast majority of the sample (over 99%) consisted of migrants travelling from within India into Bihar, most on designated trains. Migrants were more likely to be sampled if they presented symptoms between March 22 and May 3. State policy beginning May 4 during the sample collection period mandated that travellers from within or outside India (mainly migrant workers returning home due to travel restrictions) be randomly sampled and tested for COVID-19 infection from March 20 to May 22, and after May 31. Between May 22–31, only migrants from seven high-infection cities (National Capital Region, Mumbai, Ahmedabad, Pune, Surat, Kolkata, and Bangalore) in India were randomly sampled. We isolated the subsample of migrants who were randomly selected for testing, yielding 4,362 cases.

During the sample period, migrants were tested with TrueNat machines manufactured by MolBio Diagnostics in Goa (India), and positive tests were confirmed with real-time reverse polymerase chain reaction (RT-PCR) kits (CMO-PRC, 2020). Importantly, all infected migrants were tracked by the monitoring team, to determine whether they eventually recovered or died. Among randomly sampled male migrants, 1,385 infected individuals (35%), whom we call "lost", could not be tracked and thus their final outcome is uncertain. The high level of attrition is common in studies of migrant workers, whose frequent movement complicates administrative registration and tracking, particularly during a crisis (Deshingkar et al., 2008). We considered several approaches to adjusting for attrition, described below. The migrant sample, reflecting typical labor migration patterns in India, was overwhelmingly male (90%). Thus we limited our final analytical sample to 3,921 randomly sampled male migrants, for 2,536 of whom outcomes (recovery or death) are known.

Estimating infection fatality rate

Because everyone in the sample had tested positive for SARS-CoV-2, IFRs were estimated as the share of deaths among non-lost individuals in each age group. To account for potential biases due to attrition and delays between infection and recovery/death/reporting, we estimated IFRs using three separate methods, and report estimates from all three.

In age group a, denote the number of lost cases as $n_{a,lost}$, the number of recovered cases as $n_{a,recovered}$, and the number of cases ending in death as $n_{a,died}$.

Method 1 (main estimation): In our main estimation, we assumed that lost cases had the same IFR as successfully tracked cases, within each age group. This assumption was implemented by excluding lost individuals from the IFR calculation. Method 1 provided a midline IFR estimate:

$$IFR1_a = \frac{n_{a,died}}{n_{a,died} + n_{a,recovered}}$$

Method 2: In this estimation, we assumed that all lost cases eventually recovered. Thus Method 2 provided a lower bound IFR estimate:

$$IFR2_a = \frac{n_{a,died}}{n_{a,died} + n_{a,recovered} + n_{a,lost}}$$

Method 3: The share of cases with successful followup declined in late July as the volume of migrants increased. In the third method, to account for potential right-censoring of reported outcome rate due to delays between report of initial infection and report of recovery/death, we dropped all cases reported within two weeks of the last report date (July 21st):

$$IFR3_a = \left(\frac{(n_{a,died})}{n_{a,died} + n_{a,recovered} + n_{a,lost}} \middle| \text{ infection reported on or before July 7}\right)$$

Standard errors were estimated with the normal approximation for a proportion from multiple draws from a binomial distribution.

Mumbai

Data

Data on seroprevalence were obtained from a representative, stratified, random sample of slum and non-slum populations in three of twenty-four wards of Mumbai (see Malani et al. (2020) for full survey design). Sample collection lasted two weeks and ended on July 14th in slums and July 19th in non-slums. The three wards were selected to represent the city's three broad zones (city, eastern suburbs, western suburbs); choice of sampled ward within each zone was by convenience. The sample consists of 6,904 participants (4,202 from slums and 2,702 from non-slums), who were tested for IgG antibodies to the SARS-CoV-2 N-protein using the Abbott Diagnostics Architect^{TM} N-protein based test. The samples were stratified by four age groups, sex, ward, and slum/non-slum residence.

Data on reported infections and deaths by ward and age distribution of deaths were provided in reports released by the municipal governing body (Brihanmumbai Municipal Corporation, hereafter BMC). Data on ward population in slums and non-slums came from the 2011 Population Census. Data on shares of population by age and sex in each ward-slum came from the 2012 Socio-Economic and Caste Census.

Estimating IFR

Estimating number of infections. The seroprevalence survey reported seropositivity in four age groups (12–24, 25–39, 40–60, 61+), called "coarse bins". To generate infection counts that could be compared with city death statistics (which are reported in 10-year age bins), seropositivity by 10-year age bin was interpolated by fitting a non-linear function over seropositivity in the coarse bins. For the main estimation, we interpolated seropositivity in 10-year bins, using the inverse distance-weighted mean of non-missing values (using the Stata package mipolate), weighting with the squared inverse of distance. In each coarse bin, the median age of residents in Mumbai City was used as the non-missing value for age. As a sensitivity analysis, we report IFR estimates using a piecewise cubic Hermite ("pchip") interpolation for seropositivity. Interpolation predicted seroprevalence for the midpoint of each 10-year age bin, separately by sex, ward, and slum status.

The estimated sensitivity of the chemiluminiscence immunoassay ranges from 90% (95% CI: 74 to 96) (USFDA, 2020) to 96% (89 to 99) (Bryan et al., 2020) while specificity in those studies was 100% (95% CI: 95 to 100) and 99.0%, respectively. We estimated seroprevalence from seropositivity using the Rogan-Gladen correction (Rogan and Gladen, 1978) to account for imperfect accuracy of tests. In the main results, we used the midpoint of mean sensitivity estimates (93.5%) and the midpoint of corresponding specificities (99%). As a sensitivity analysis, we replicated results with an upper bound for seroprevalence based on the Abbott test's lower bound of sensitivity (90%) and upper bound of specificity (100%) (Bryan et al., 2020) (Figure 4).

Denote the estimated number of infections in age bin a, sex g in sampled ward s as:

$$\widehat{inf}_{ags} = SP_{ags} \times pop_{ags}$$

where $SP_{ag,s}$ is the estimated seroprevalence rate, and $pop_{ag,s}$ is population.

Estimating the number of infections in non-sampled wards. BMC death data reported the ward of death, but not the ward of residence. Discussion with government officials and review of the data indicated that the ward of death was not a reliable indicator of ward of residence. This implied that calculating IFR by dividing the number of ward-level deaths by the number of ward-level infections would overestimate deaths in wards with large hospitals and underestimate them elsewhere. Instead, we used the seroprevalence surveys to generate estimates of city-wide infection counts.

To estimate true number of infections in non-sampled wards, we drew on administrative ward-level infection counts (which were universally available from city reports), and assumed that the BMC underestimated the true population infection count at the same rate in sampled and non-sampled wards within the same zone. This assumption is supported by Table 1, which shows that in the three wards where we obtained seroprevalence data, case multipliers were very similar.

Thus, in each zone z, we calculated a case multiplier based on sampled ward s:

$$\gamma_z = \frac{\sum_a \sum_g \widehat{inf}_{ag,s}}{\text{BMC-reported cases}_s}$$

The multiplier indicates the under-reporting rate in each zone z. The numerator of the expression is calculated from the seroprevalence surveys as above, and the denominator is taken from the BMC reports. BMC-reported cases were measured as of July 19, the last day of seroprevalence sample collection. We then multiplied the BMC's reported number of positive cases in non-sampled ward n in zone z by γ_z. That is,

$$\widehat{inf}_{n,z} = \gamma_z \times \text{BMC-reported cases}_n$$

The benefit of this approach is that it allows pandemic intensity to vary across wards, a realistic assumption given significant ward-level variation in reported cases per capita and number of containment zones.

This approach also implicitly assumes that the BMC under-reports cases in slums and non-slums at the same rate, *i.e.* a ward's case multiplier does not depend on share of population living in slums. This assumption is also supported by the consistent multipliers reported in Supplement Table 1, across three wards with different slum shares.

Estimating the number of infections in each age-sex group in non-sampled wards. We did not observe the age and sex distribution of infections outside of the sampled wards, so it was necessary to assume that non-sampled wards had the same age and sex distribution of infections of sample wards. This was supported by similar age and sex distributions of infections in the three wards with seroprevalence surveys. Figure 9 shows the calculated age and sex distribution of infections; note that the distribution of infections measured with seroprevalence skews younger than the number of reported positive cases, which we presume omits many infected but asymptomatic young people. This approach would cause error if the age distribution varied substantially across wards, but it is overall quite similar; even the median age gap between slums and non-slums was less than one year.

The number of infections in non-sampled ward s for sex g in age a was thus calculated as:

$$\widehat{inf}_{ag,n} = \frac{\sum_s \alpha_{ag,s}}{\sum_s \sum_a \sum_g \alpha_{ag,s}} \times \widehat{inf}_n,$$

where $\alpha_{ag,s}$ is the age-sex group's share of total cases in sampled ward s.

Estimating the number of deaths. To map infection counts to death counts, we must make assumptions about the delays between infection and death and between infection and seroprevalence. The literature suggests the distribution of delay between symptom onset and death (Linton et al., 2020) that is wider than that between onset and seroconversion (Stringhini et al., 2020). Linton et al. estimated a median time delay of 13 days (17 days with right truncation) between illness onset to death. Stringhini et al. estimated a mean delay of 11. days between symptom onset and seroconversion. Based on these estimates, we assumed that the delay between infection and death is on average two days longer than the delay between infection and seroconversion. In the main results, the number of deaths was therefore measured as the cumulative deaths reported in each Mumbai ward as of July 21. This is likely to slightly overstate the IFR, since some deaths may have been associated with individuals who contracted the virus after testing negative in the seroprevalence surveys. However, this upward bias is partially balanced out by the fact that the time between seroconversion and death is not uniform and is likely to be longer than 2 days for a non-trivial share of cases.

Rather than model non-uniform delays between infection and death, we bounded our IFR estimates from above by choosing more conservative death dates. In sensitivity analyses reported below (Figure 1), we replicated IFR estimates using deaths from one week (July 28) and two weeks (August 4) after the end of seroprevalence surveying, both of which plausibly overestimated the number of deaths related to the seroprevalence surveys, given the context of steadily increasing case counts in Mumbai from June to August.

The assumption that deaths measured 1 and 2 weeks later will lead to upward biased IFRs is further strengthened by recent evidence from roughly 125,000 cases in two other Indian states (Laxminarayan et al., 2020), which found that delays between case report and death were significantly shorter than delays found in China and the United States (Lewnard et al., 2020).

We used the age distribution of deaths as reported by the BMC up to the date used for measuring deaths, and the sex distribution (65% male, 35% female) up to August 3 (Debroy, 2020) (the sex distribution of deaths was not included in earlier reports). This yields the estimated number of city-wide deaths by age-sex group, d_{ag}.

Estimating city-wide IFR by age in Mumbai. Denote the final city-wide IFR in Mumbai, in age bin a for sex g, as IFR_{ag}:

$$IFR_{ag} = \frac{d_{ag}}{\sum_{ns} \widehat{inf}_{ag,ns} + \sum_s \widehat{inf}_{ag,s}}$$

Standard errors of IFRs were calculated reflecting propagation of the design-based standard errors of the age- and sex-specific seroprevalence estimates with a normal approximation.

Karnataka

Data

Data on seroprevalence were obtained from the Karnataka Seroprevalence Survey (hereafter KSS) a state-wide representative sample of urban and rural areas in 20 out of 30 districts in Karnataka, representing 5 broader regions (see Mohanan et al. (2021) for a detailed survey description). The sample was collected from June 15 to August 29, 2020. Collection times within individual regions were significantly shorter. The study sample was drawn from an existing representative sample of a panel survey—the Consumer Pyramids Household Survey (CPHS)—collected by the Center for Monitoring Indian Economy (CMIE). Our analytical subsample consists of 1,196 tests for IgG antibodies to the receptor binding domain (RBD) of the SARS-CoV-2 virus using an ELISA test developed by Translational Health Science and Technology Institute, India. The sample was not stratified by age and sex, an issue addressed below.

Data on confirmed COVID-19 deaths by district were drawn from Government of Karnataka Department of Health and Family Welfare bulletins, which are released several times per week. Data on the age distribution of total COVID-19 deaths were given by public reports from the state COVID-19 task force. Data on the sex distribution of deaths by age group were obtained from an individual-level dataset of confirmed COVID-19 deaths which was updated through July. The case-level death data were parsed from covid19india.org. Age- and sex-disaggregated population for districts and regions was drawn from the 2012 Socio-Economic and Caste Census (SECC).

Estimating IFR

Estimating the number of infections. The KSS dataset was designed to be representative of 5 broader regions in Karnataka. We therefore can take the ELISA positive test rate as an unbiased measure of the region-level positivity rate. We pooled the data across regions to obtain a statewide test positivity rate in each age and sex group, weighting by region population in each age-sex group.

We then corrected for the sensitivity (84%) and specificity (100%) of the ELISA imunoassay (Chaudhuri et al., 2020), using the Rogan-Gladen correction (Rogan and Gladen, 1978). This yielded the estimated seroprevalence by age-sex group SP_{ag}, which is multiplied by population pop_{ag} in each age-sex bin to generate an estimated number of infections $\widehat{inf}_{ag}$, as was done in Mumbai.

Estimating the number of deaths. The seroprevalence samples were collected at different times in different regions, with the survey period spanning roughly two months (Table 2). To estimate an IFR, we need to match the timing of deaths to the timing of seroprevalence surveying in each region.

Choice of dates for measuring deaths. As in Mumbai, we worked from an assumption that the average time difference between seroconversion and death was two days, while testing sensitivity to alternate assumptions (Figure 2). We therefore matched the estimated number of infections calculated in each region to the number of deaths recorded in administrative data two days after the last date of seroprevalence surveying. As in Mumbai, if the two-day delay between seroconversion and death was uniform, this approach would overestimate the IFR, because it counts the deaths of some people who may have been infected *after* recording negative seroprevalence tests.

In all regions except Bangalore, seroprevalence surveying was conducted over a three week period or less, making it straightforward to match test data to death data. In Bangalore, surveying was begun in mid-June but was interrupted by a lockdown. Survey teams returned to finish sampling in the last week of August. Matching Bangalore deaths to the last date of seroprevalence surveying is therefore likely to overestimate the IFR, because a number of those deaths may have been associated with individuals contracting SARS-CoV-2 after testing negative. It was not possible to disaggregate the early and late surveys because death reporting was at the district level, and the early and late survey groups were not representative in and of themselves. To adjust for increased uncertainty regarding the number of infections in Bangalore, we therefore report a sensitivity analysis for all of Karnataka excluding Bangalore (Figure 5).

On some days, official deaths were not reported; in those cases, we used deaths from the following day.[1]

Estimating the number of deaths in each demographic group: The Karnataka state government released total death counts on a daily basis, but only intermittently published the age distribution of state-wide deaths. To attribute daily deaths to age and sex groups, we used the age distribution of deaths from the nearest available date. The longest period between the date used for deaths and the date used for age-shares was 13 days.

Government reports provided age shares of deaths in 10-year bins in the form (e.g.) 51-60, while the seroprevalence surveys provided age bins in the form (e.g.) 50-59. To harmonize the age groups, we use the medians of the provided bins (e.g. median of 51-60 is 55.) to interpolate death data to match the age bins in the seroprevalence data, using an inverse distance weighted average method via the mipolate Stata package. Because the target age bins were very close to the available age bins, the risk of error here is small. As a sensitivity test, we replicated IFRs using piecewise cubic Hermite interpolation. For more details, see the discussion on interpolation in Mumbai.

[1] In Belgaum, the target date was July 27th; we used July 28. In the sensitivity test, the target date was August 10; we used August 11.

In the absence of death data disaggregated by age and sex on most dates, we assumed that, within age group, the sex distribution of deaths was uniform across regions and equal to the state-wide sex distribution of deaths reported between April and July. This assumption is supported by the finding that IFRs among males were approximately double those among females, consistent with reports from other countries.

Standard errors of IFRs reflect propagation of design-based standard errors of the age- and sex-specific seroprevalence estimates with a normal approximation.

Tamil Nadu

Data

Data on seroprevalence in Tamil Nadu comes from a state-conducted population-level seroprevalence survey of 26,640 adults aged 18 and older, covering the 37 districts of the state. The sample was collected between October 19 and November 30, 2020. Collection times within districts were often significantly shorter. The sampling frame divided Tamil Nadu's 37 administrative districts (as of February 2020) into health unit districts (HUDs), then formed and randomly sampled urban and rural clusters. Within clusters, enumerators started at a randomly selected GPS starting point, sampling one person from households adjacent to the starting point (using the Kish method) to provide a biosample, until 30 persons were sampled per cluster. Serum was analyzed for IgG antibodies to the SARS-CoV-2 spike protein using either the iFlash-SARS-CoV-2 IgG (Shenzhen YHLO Biotech; sensitivity of 95% and specificity of 95% per manufacturer (Shenzhen YHLO Biotech No. Ltd., 2020)) or the Vitros anti-SARS-CoV-2 IgG CLIA kit (Ortho- Clinical Diagnostics; sensitivity of 90% and specificity of 100% per manufacturer). For uniformity, in each district, one type of kit was used; in one district (Chennai) both kits were used. Our analytical subsample consists of 26,107 CLIA antibody tests that could be conclusively determined as positive or negative.

Case-level data on state-wide COVID-19 deaths was collected from daily government reports released on https://stopcorona.tn.gov.in/daily-bulletin/. The data cover all recorded deaths, beginning on March 25 and updated until December 24, 2020. The data was collected and shared by the faculty and staff of the Urban Expansion Observatory at Pillai College, New Panvel, Maharashtra. The dataset contains 12,019 observations, each with information about age, sex, dates of reported positive test and death, and district. Age- and sex-disaggregated population data were from the 2012 Socio-Economic and Caste Census.

Estimating IFR

Estimating the number of infections. We estimated the number of state-wide infections associated with measured seroprevalence in three steps. First, we calculated positive test rate by district-age-sex group, separately for each kit. Positive test rate was estimated by regressing an indicator for positive result on district-age-sex group indicators, clustering standard errors within the randomly sampled clusters. Seroprevalence sample collection was stratified by district, health unit district (HUD), then cluster; within clusters, age and sex of test participants was random. Thus we take the positive test rate for each district-age-sex group as representative.

Second, we adjusted for test inaccuracies for each kit, using the Rogan-Gladen correction (Rogan and Gladen, 1978) and the manufacturer-provided sensitivity and specificity. In a sensitivity check, we utilized the lowest estimated sensitivity and corresponding specificity, from any manufacturer-conducted or independent analyses of each kit (Figure 7). Independent analysis of the iFlash kit from Shenzhen YHLO Biotech estimated sensitivity of 93% (95% CI: 84 to 97) and specificity of 92% (85 to 97) (Plebani et al., 2020). FDA evaluation of the Vitros kit from Ortho-Clinical Diagnostics suggests 100% sensitivity (95% CI: 88 to 100) and 100% specificity (95 to 100) (USFDA, 2020), while other analysis estimated a sensitivity of 98% (92 to 100) and specificity of 97% (85 to 100) (Theel et al., 2020). Note that, unlike in Mumbai and Karnataka, the minimum sensitivity of the kits in Tamil Nadu had lower corresponding specificity, leading to lower overall seroprevalence estimates. In the district in which both kits were used, kit-specific seroprevalence estimates were averaged, using proportion of sample size (by age-sex group) as the weight.

Third, we estimated number of infections in each district-age-sex group by multiplying seroprevalence rate by population. Age- and sex-disaggregated population data was available for census districts. Finally, estimated state-wide infections by age-sex group were calculated by simply summing over all districts.

Estimating number of deaths in each demographic group. As in Mumbai and Karnataka, we matched the estimated number of infections calculated in each district to the number of deaths recorded in administrative data two days after the last date of seroprevalence surveying. We test sensitivity to alternative assumptions by measuring cumulative deaths 1 week and 2 weeks after the main date (Figure 3). As explained in the supplement sections on Mumbai and Karnataka, these are all plausible over-estimates of deaths associated with the measured seroprevalence level. Cumulative

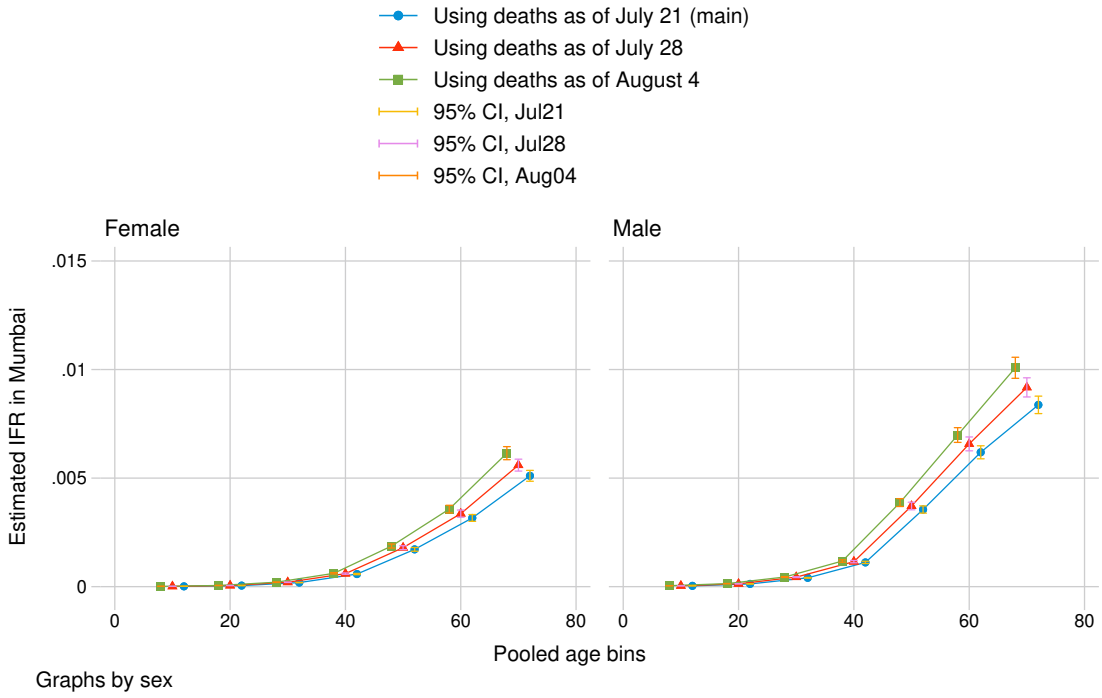
deaths in each demographic group were measured up to the specified date. Cumulative deaths were measured from March through December, a longer span than in other locations. This may over-estimate deaths, and therefore over-estimate IFR, if infected individuals gradually become seronegative after recovery. Available evidence suggests that antibody loss varies significantly with symptom severity ((Ripperger et al., 2020; Ibarrondo et al., 2020; Long et al., 2020)). Because we cannot precisely estimate antibody loss rates across the population, and because IFR estimates in Tamil Nadu are the lowest across the four locations, we simply note that, given available data, our IFR estimates are conservatively high.

Seroprevalence surveying lasted longer than three weeks in 6 out of 37 districts. In these districts, there is a risk that seroprevalence in the population changed during sample collection. During a period of increasing pandemic intensity, this may under-estimate seroprevalence, over-estimating IFR. As a sensitivity check, we limit analysis of both seroprevalence and deaths to the 31 districts in which seroprevalence sample collection was less than three weeks (Figure 8).

Age- and sex-specific IFRs were estimated as the proportion of state-wide deaths divided by estimated infections. Standard errors reflect propagation of error from the HUD-age-sex estimates of positive test rates.

eFigure 1

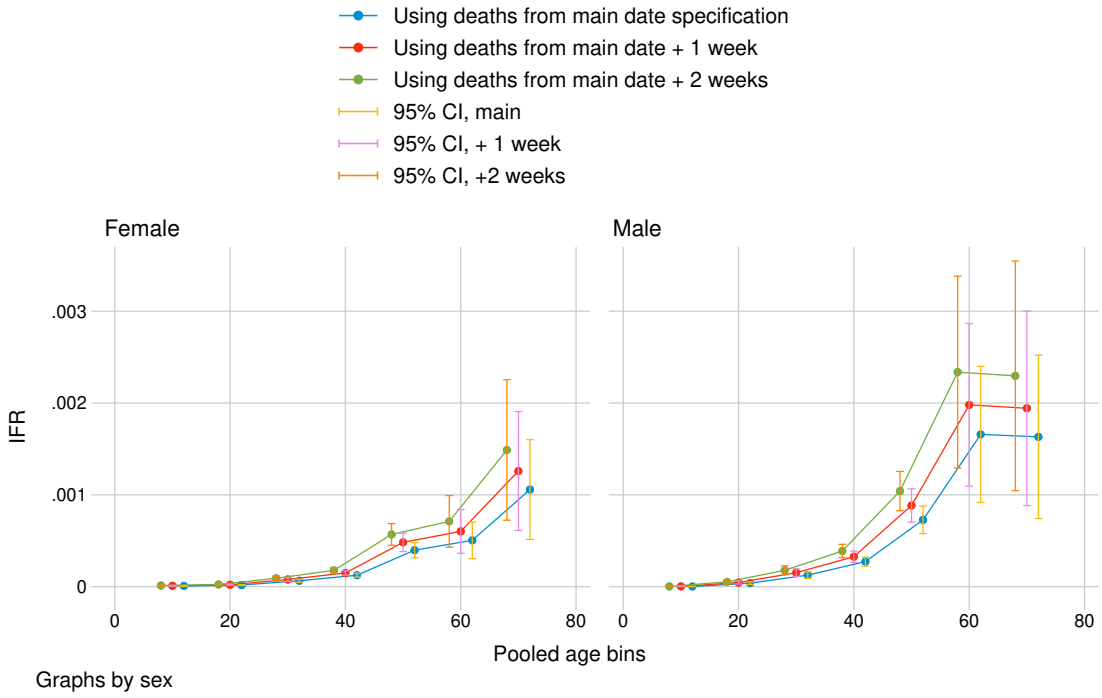
Mumbai: sensitivity analysis using number of COVID deaths from 1 and 2 weeks after date of deaths in main estimation



"Date of death" refers to the day on which we measured cumulative deaths as reported by the city government (BMC). The main date specification measured deaths two days after the end of seroprevalence sample collection. Graphs by sex with 95% confidence intervals. Standard errors reflect propagation of error from design-based uncertainty of seroprevalence estimates. IFRs are calculated in age bins 0-19, 20-29, ... 60-69, and 70+.

eFigure 2

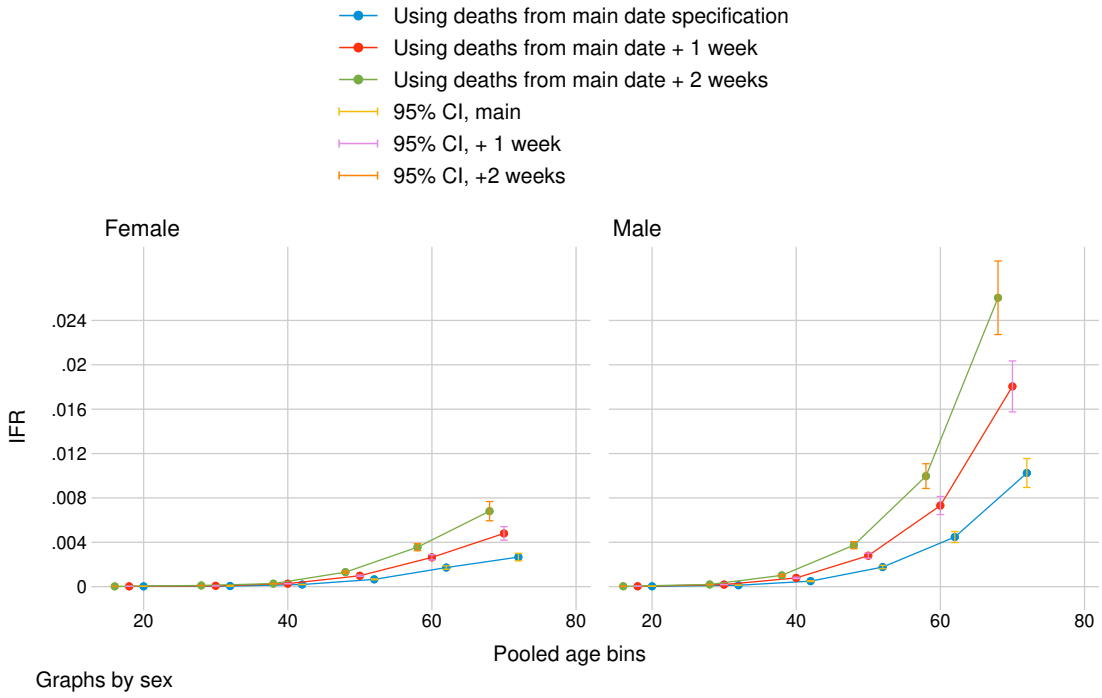
Karnataka: sensitivity analysis using number of COVID deaths from 1 and 2 weeks after date of deaths in main estimation



"Date of death" refers to the date on which we measured cumulative COVID-19 deaths. Main date of specification was determined separately for each sampled region as two days after the median date of sample collection. Graphs by sex with 95% confidence intervals. Standard errors reflect propagation of error from design-based uncertainty of seroprevalence estimates. IFRs are calculated in age bins 0-9, ... 60-69, and 70+.

eFigure 3

Tamil Nadu: sensitivity analysis using number of COVID deaths from 1 and 2 weeks after date of deaths in main estimation

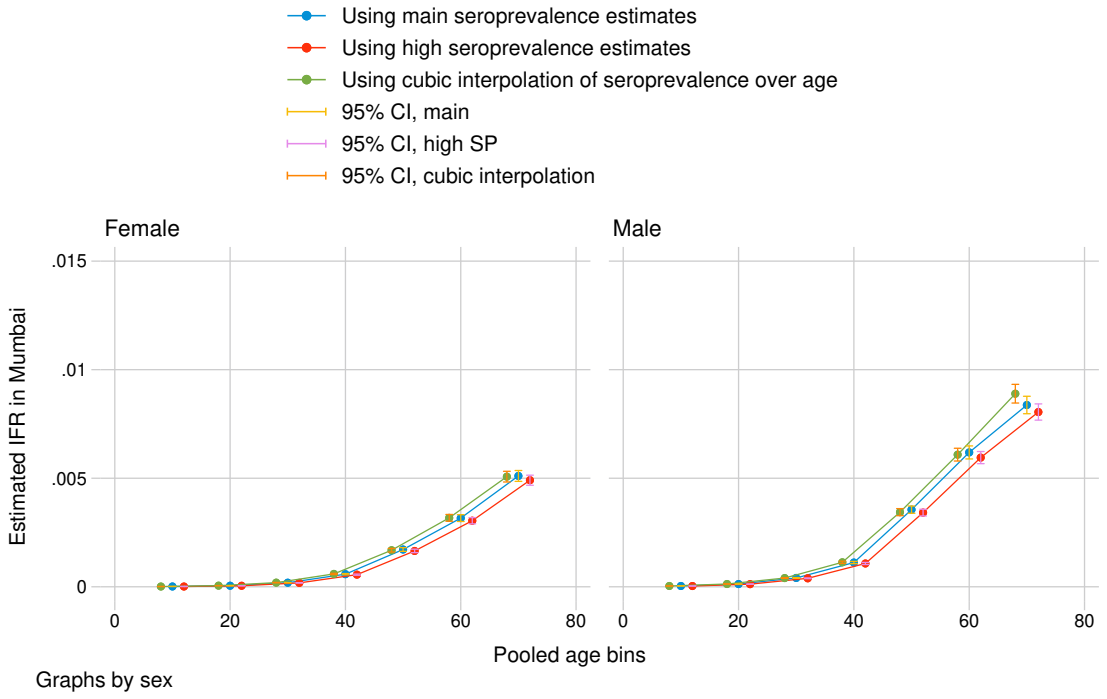


"Date of death" refers to the date on which we measured cumulative COVID-19 deaths. Main date of specification was determined separately for each sampled district (N = 37) as two days after the last date of sample collection. Graphs by sex with 95% confidence intervals. Standard errors reflect propagation of error from uncertainty in estimating positive test rate by HUD-age-sex group. IFRs are calculated in age bins 18-29, 30-39 ... 60-69, and 70+.

eFigure 4

Mumbai:

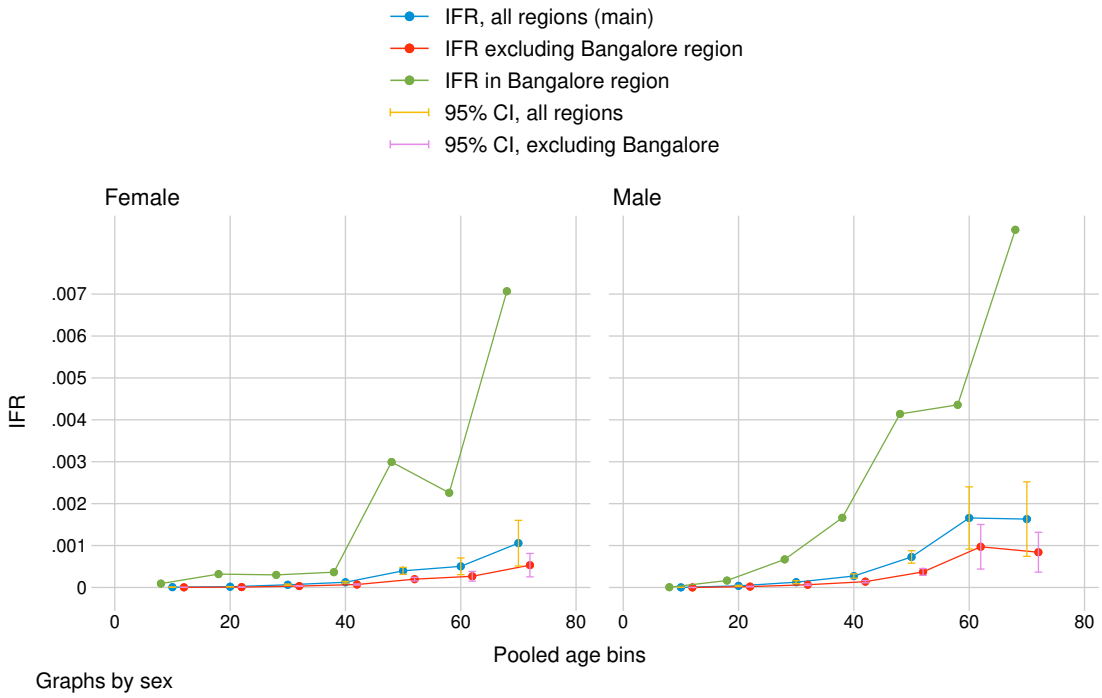
sensitivity analysis, using alternative estimate of seroprevalence and different interpolation method



"Main seroprevalence estimates" use midpoint sensitivity estimate of the Abbott antibody test to calculate seroprevalence from seropositivity in sampled wards, then interpolates seroprevalence to finer age bins with inverse distance weighting (IDW). "High seroprevalence estimates" use minimum sensitivity of the Abbott test to calculate seroprevalence from seropositivity and IDW interpolation. The final sensitivity analysis uses midpoint sensitivity, but piecewise cubic Hermite interpolation to estimate seroprevalence in finer bins. Graphs by sex with 95% confidence intervals. Standard errors reflect propagation of error from design-based uncertainty of seroprevalence estimates. IFRs are calculated in age bins 0-19, 20-29, ... 60-69, and 70+.

eFigure 5

Karnataka: sensitivity analysis isolating Bangalore from other sampled regions

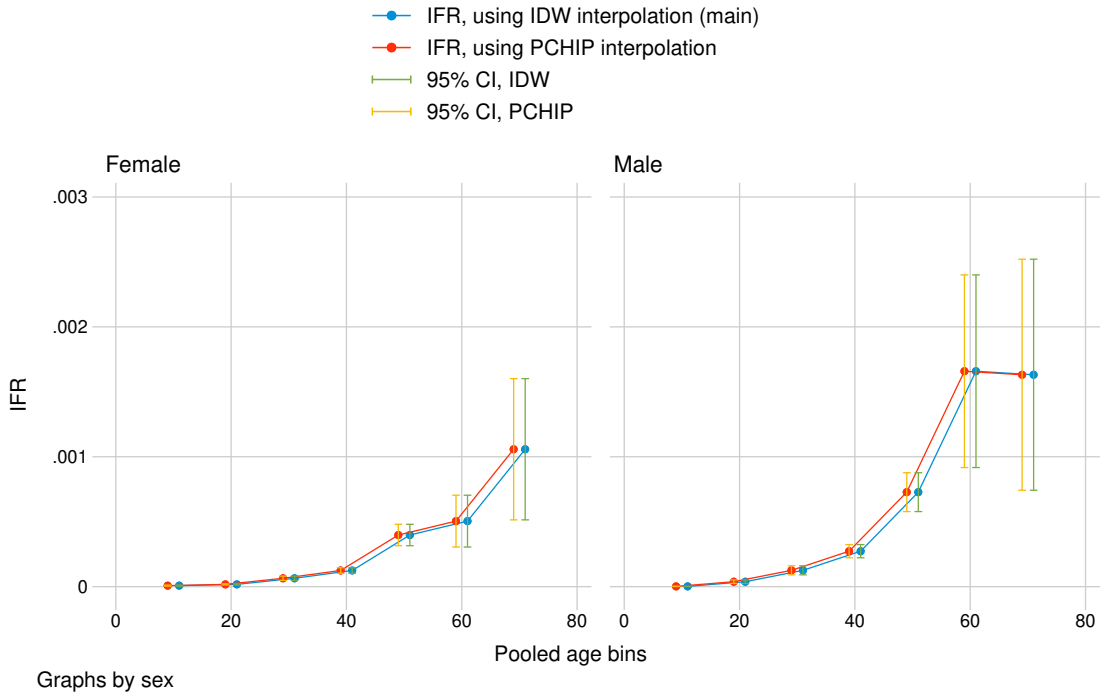


IFRs in main specification are calculated by pooling seroprevalence and death estimates from all five sampled regions of Karnataka. IFRs excluding Bangalore pool from the four remaining regions. Graphs by sex with 95% confidence intervals. Standard errors reflect propagation of error from design-based uncertainty of seroprevalence estimates. Confidence intervals are not reported for Bangalore due to small sample size, and age-specific estimated IFRs in Bangalore should not be interpreted as conclusive. IFRs are calculated in age bins 0-9, ... 60-69, and 70+.

eFigure 6

Karnataka:

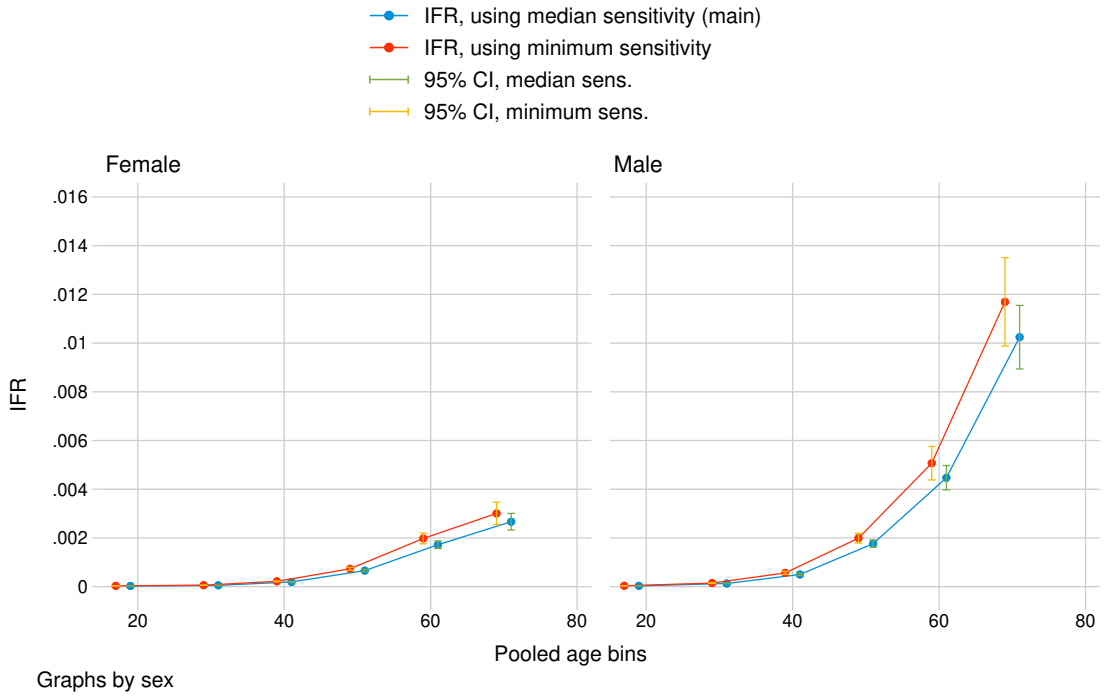
sensitivity analysis using piecewise cubic Hermite interpolation to estimate age bin share of deaths



Government reports provide age-shares of deaths in age bins of the form 11-20, 21-30, etc. To match seroprevalence estimates, we interpolate age-shares of deaths in the form 10-19, 20-29, etc. Main specification uses the inverse distance weighted average (IDW) to interpolate age shares. sensitivity analysis uses piecewise cubic Hermite interpolation. Interpolation was done with Stata package mipolate. Graphs by sex with 95% confidence intervals. Standard errors reflect propagation of error from design-based uncertainty of seroprevalence estimates. Confidence intervals are not reported for Bangalore due to small sample size, and age-specific estimated IFRs in Bangalore should not be interpreted as conclusive. IFRs are calculated in age bins 0-9, ... 60-69, and 70+.

eFigure 7

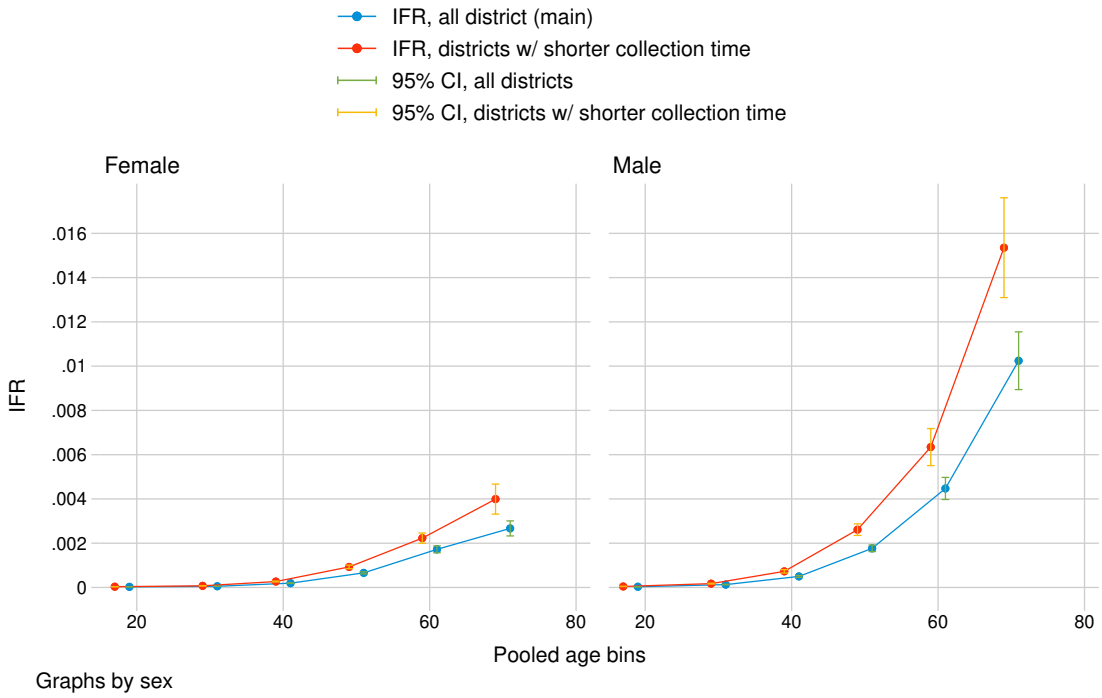
Tamil Nadu: sensitivity analysis using minimum sensitivity and corresponding specificity of immunoassays.



Two kits were used to evaluate seropositivity. Seroprevalence rate was calculated from the seropositivity rate using the Rogan-Gladen correction for imperfect test sensitivity and specificity. Main estimation used the manufacturer-provided sensitivity and corresponding specificity of the kits. The robustness check uses the lowest estimated sensitivity of both kits, which was the manufacturer-provided estimate for the Ortho-Clinical kit. Graphs by sex with 95% confidence intervals. Standard errors reflect propagation of error from uncertainty in estimating positive test rate by HUD-age-sex group. IFRs are calculated in age bins 18-29, 30-39 ... 60-69, and 70+.

eFigure 8

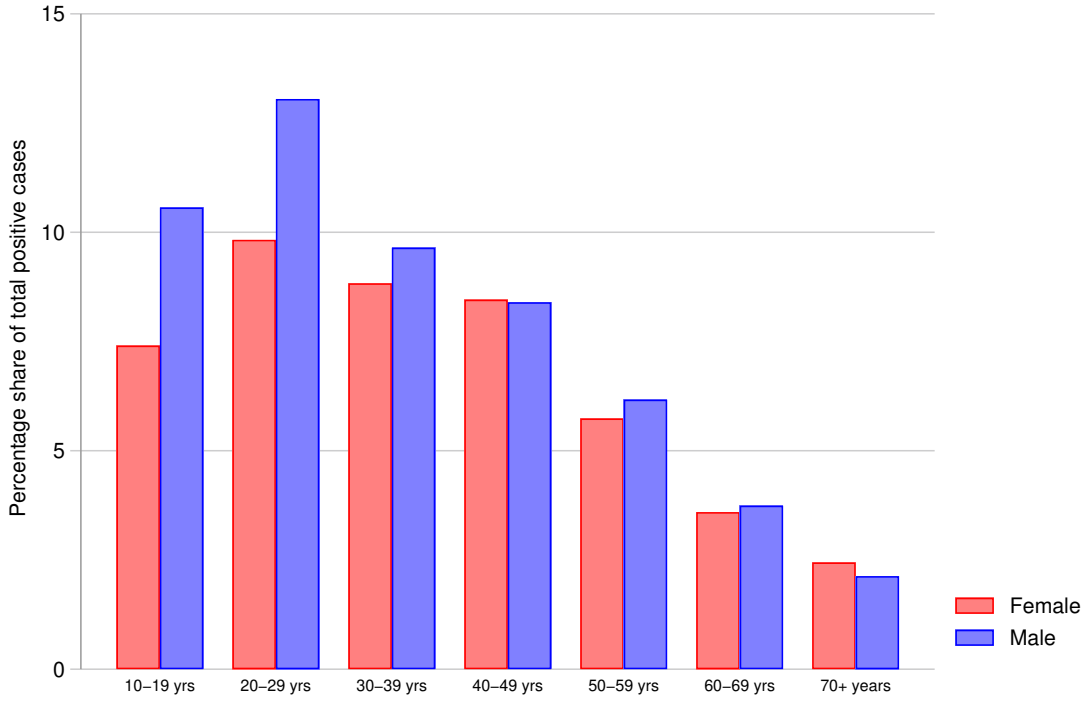
Tamil Nadu: sensitivity analysis excluding districts where sample collection duration exceeded 3 weeks



IFRs in main specification are calculated by pooling seroprevalence and death estimates from all 37 districts of Tamil Nadu. IFRs estimated from districts with shorter collection time exclude 6 districts where seroprevalence surveying lasted longer than three weeks. Graphs by sex with 95% confidence intervals. Standard errors reflect propagation of error from uncertainty in estimating positive test rate by HUD-age-sex group. IFRs are calculated in age bins 18-29, 30-39 ... 60-69, and 70+.

eFigure 9

Age-sex cohorts' share of positive cases from sampled wards in Mumbai seroprevalence survey



"Total positive cases" refers to the estimated number of total infections in Mumbai, multiplying age- and sex-specific seroprevalence rate by group population, summed across age-sex groups and wards. The age- and sex-share of total cases refers to estimated number of infections in age-sex group *ag*, divided by estimated total infections. Age bins are 0-19, 20-29, ...60-69, and 70+.

1 Supplementary Tables

eTable 1

Zone-wise case multipliers for main and higher seroprevalence estimates based on sampled wards

		No. infections			γ_z	
Ward	*Zone*	*BMC report*	*main SP*	*high SP*	*main SP*	*high SP*
(1)	(2)	(3)	(4)	(5)	(6)	(7)
F North	City	4,017	190,652	211,835	47.6	52.3
M West	Eastern	2,965	139,791	155,322	47.5	52.9
R North	Western	2,421	145,413	161,569	60.6	66.4

"Number of infections, main SP" refers to the estimated seroprevalence (using the midpoint estimated sensitivity of the antibody test) multiplied by population in each sampled ward. "Number of infections, high SP" uses lowest bound sensitivity of the antibody test. Case multiplier "γ_z, main SP" (Column 6) was calculated by dividing Column 4 by Column 3. "γ_z, high SP" (Column 7) was calculated by dividing Column 5 by Column 3. Main SP indicates seroprevalence estimated from midpoint of two published estimates of sensitivity of the antibody test. High SP indicates seroprevalence was estimated using the minimum sensitivity and maximum specificity of the antibody test, generating a upper-bound estimate.

eTable 2

Karnataka: duration of sample collection by region

Region	Duration of sample collection (days)	Dates of sample collection
Bangalore	73	June 17 – August 29
Mysore	18	August 3 – August 21
Kannada	16	August 6 – August 21
Belgaum	17	July 8 – July 25
Gulbarga	10	July 21 – July 31

eTable 3

Mumbai: summary statistics used in calculating IFR

	Seroprev. sample size		Seroprev. rate		No. deaths	
Age	Male	Female	Male	Female	Male	Female
(1)	(2)	(3)	(4)	(5)	(6)	(7)
10–19	210	166	0.339	0.573	14	7
20–29	642	561	0.323	0.555	59	32
30–39	946	897	0.344	0.476	168	90
40–49	929	811	0.428	0.502	454	245
50–59	750	605	0.403	0.531	983	530
60–69	424	307	0.441	0.494	1094	589
70–89	153	88	0.503	0.397	1000	539

Columns 2 and 3 are the number of participants in the seroprevalence survey. Columns 4 and 5 reflect the city-wide seroprevalence rate, adjusted for antibody test sensitivity and specificity. Because we allowed seroprevalence rate to vary across wards, seroprevalence was calculated as number of estimated infections divided by population. Columns 6 and 7 are the number of deaths reported by the city government, split by gender with the assumption that deaths were 65% male, 35% female. See Materials and Methods for details.

eTable 4

Karnataka: summary statistics used in calculating IFR

	Seroprev. sample size		Seroprev. rate		No. deaths	
Age	Male	Female	Male	Female	Male	Female
(1)	(2)	(3)	(4)	(5)	(6)	(7)
10–19	28	30	0.503	0.430	5	14
20–29	84	86	0.342	0.539	59	41
30–39	84	137	0.443	0.388	191	85
40–49	178	175	0.533	0.516	406	168
50–59	129	127	0.516	0.513	687	359
60–69	59	41	0.367	0.474	801	326
70–89	19	19	0.468	0.485	549	321

Columns 2 and 3 are the number of participants in the seroprevalence survey. Columns 4 and 5 reflect the state-wide seroprevalence rate, adjusted for antibody test sensitivity and specificity. Columns 6 and 7 are the number of deaths reported by the state government. See Materials and Methods for details.

eTable 5

Tamil Nadu: summary statistics used in calculating IFR

	Seroprev. sample size		Seroprev. rate		No. deaths	
Age	Male	Female	Male	Female	Male	Female
(1)	(2)	(3)	(4)	(5)	(6)	(7)
18–29	2267	3025	0.292	0.242	56	41
30–39	1794	3576	0.268	0.306	184	90
40–49	1719	3041	0.276	0.295	629	259
50–59	1475	2340	0.284	0.296	1619	614
60–69	1229	1488	0.261	0.249	2510	923
70–89	742	659	0.232	0.258	3029	882

Columns 2 and 3 are the number of participants in the seroprevalence survey. Columns 4 and 5 reflect the state-wide seroprevalence rate, adjusted for antibody test sensitivity and specificity. Because we calculated seroprevalence and corresponding deaths using different dates for different districts, seroprevalence was calculated as total infections across district at time of seroprevalence data collection, divided by population. Columns 6 and 7 are the number of deaths reported by the state government, summed across districts at the time of seroprevalence data collection. See Materials and Methods for details.

eTable 6

Bihar male migrants: summary statistics used in calculating IFR

Age	No. infected	% successfully tracked	No. deaths
(1)	(2)	(3)	(4)
10–19	568	0.674	0
20–29	1472	0.628	6
30–39	989	0.670	12
40–49	543	0.602	5
50–59	189	0.667	3
60–69	69	0.681	2
70–89	13	0.615	1

The table summarizes the group used for analysis: randomly sampled male migrants. Column 2 is the number of men in the sample, who were all infected. Column 3 is the percentage of infected men for whom trackers successfully confirmed an outcome: either recovery or death. Column 4 is the number of confirmed deaths.

References

Bryan, Andrew, Gregory Pepper, Mark H. Wener, Susan L. Fink, Chihiro Morishima, Anu Chaudhary, Keith R. Jerome, Patrick C. Mathias, and Alexander L. Greninger, "Performance Characteristics of the Abbott Architect SARS-CoV-2 IgG Assay and Seroprevalence in Boise, Idaho," *Journal of Clinical Microbiology*, 2020, *58* (8).

Chaudhuri, Susmita, Ramachandran Thiruvengadam, Souvick Chattopadhyay, Farha Mehdi, Pallavi Kshetrapal, Tripti Shrivastava, Bapu Koundinya Desiraju, Gaurav Batra, Gagandeep Kang, and Shinjini Bhatnagar, "Comparative evaluation of SARS-CoV-2 IgG assays in India," *Journal of Clinical Virology*, October 2020, *131*.

CMO-PRC, "Pre-release cm-243: Directions for random testing," May 2020.

Debroy, Sumitra, "Men account for 65% of Mumbai's 6,000+ Covid deaths, CFR high at 6.1%: BMC," *The Times of India*, August 2020.

Deshingkar, Priya, Rajiv Khandelwal, and John Farrington, "Support for migrant workers: The missing link in India's development," Technical Report, Overseas Development Institute September 2008.

Ibarrondo, F. Javier, Jennifer A. Fulcher, David Goodman-Meza, Julie Elliott, Christian Hofmann, Mary A. Hausner, Kathie G. Ferbas, Nicole H. Tobin, Grace M. Aldrovandi, and Otto O. Yang, "Rapid Decay of Anti–SARS-CoV-2 Antibodies in Persons with Mild Covid-19," *New England Journal of Medicine*, September 2020, *383* (11). Publisher: Massachusetts Medical Society pages = 1085–1087.

Laxminarayan, Ramanan, Brian Wahl, Shankar Reddy Dudala, K. Gopal, Chandra Mohan B, S. Neelima, K. S. Jawahar Reddy, J. Radhakrishnan, and Joseph A. Lewnard, "Epidemiology and transmission dynamics of COVID-19 in two Indian states," *Science*, November 2020, *370* (6517), 691–697.

Lewnard, Joseph A., Vincent X. Liu, Michael L. Jackson, Mark A. Schmidt, Britta L. Jewell, Jean P. Flores, Chris Jentz, Graham R. Northrup, Ayesha Mahmud, Arthur L. Reingold, Maya Petersen, Nicholas P. Jewell, Scott Young, and Jim Bellows, "Incidence, clinical outcomes, and transmission dynamics of severe coronavirus disease 2019 in California and Washington: prospective cohort study," *BMJ*, May 2020, *369*.

Linton, Natalie M., Tetsuro Kobayashi, Yichi Yang, Katsuma Hayashi, Andrei R. Akhmetzhanov, Sung-Mok Jung, Baoyin Yuan, Ryo Kinoshita, and Hiroshi Nishiura, "Incubation Period and Other Epidemiological Characteristics of 2019 Novel Coronavirus Infections with Right Truncation: A Statistical Analysis of Publicly Available Case Data," *Journal of Clinical Medicine*, February 2020, *9* (2).

Long, Quan-Xin, Xiao-Jun Tang, Qiu-Lin Shi, Qin Li, Hai-Jun Deng, Jun Yuan, Jie-Li Hu, Wei Xu, Yong Zhang, Fa-Jin Lv, Kun Su, Fan Zhang, Jiang Gong, Bo Wu, Xia-Mao Liu, Jin-Jing Li, Jing-Fu Qiu, Juan Chen, and Ai-Long Huang, "Clinical and immunological assessment of asymptomatic SARS-CoV-2 infections," *Nature Medicine*, August 2020, *26* (8), 1200–1204.

Malani, Anup, Daksha Shah, Gagandeep Kang, Gayatri Nair Lobo, Jayanthi Shastri, Manoj Mohanan, Rajesh Jain, Sachee Agrawal, Sandeep Juneja, Sofia Imad, and Ullas Kolthur-Seetharam, "Seroprevalence of SARS-CoV-2 in slums versus non-slums in Mumbai, India," *The Lancet Global Health*, November 2020, *0* (0).

Mohanan, Manoj, Anup Malani, Kaushik Krishnan, and Anu Acharya, "Prevalence of SARS-CoV-2 in Karnataka, India," *JAMA*, February 2021.

Plebani, Mario, Andrea Padoan, Davide Negrini, Benedetta Carpinteri, and Laura Sciacovelli, "Diagnostic performances and thresholds: The key to harmonization in serological SARS-CoV-2 assays?," *Clinica Chimica Acta; International Journal of Clinical Chemistry*, October 2020, *509*, 1–7.

Ripperger, Tyler J., Jennifer L. Uhrlaub, Makiko Watanabe, Rachel Wong, Yvonne Castaneda, Hannah A. Pizzato, Mallory R. Thompson, Christine Bradshaw, Craig C. Weinkauf, Christian Bime, Heidi L. Erickson, Kenneth Knox, Billie Bixby, Sairam Parthasarathy, Sachin Chaudhary, Bhupinder Natt, Elaine Cristan, Tammer El Aini, Franz Rischard, Janet Campion, Madhav Chopra, Michael Insel, Afshin Sam, James L. Knepler, Andrew P. Capaldi, Catherine M. Spier, Michael D. Dake, Taylor Edwards, Matthew E. Kaplan, Serena Jain Scott, Cameron Hypes, Jarrod Mosier, David T. Harris, Bonnie J. LaFleur, Ryan Sprissler, Janko Nikolich-Žugich, and Deepta Bhattacharya, "Orthogonal SARS-CoV-2 Serological Assays Enable Surveillance of Low-Prevalence Communities and Reveal Durable Humoral Immunity," *Immunity*, November 2020, *53* (5), 925–933.e4.

Rogan, W. J. and B. Gladen, "Estimating prevalence from the results of a screening test," *American Journal of Epidemiology*, January 1978, *107* (1), 71–76.

Shenzhen YHLO Biotech No. Ltd., "Customer Notification: Sensitivity and Specificity of iFlash-SARS-CoV-2 and IgM kits from Clinical Trials.," Technical Report 2020.

Stringhini, Silvia, Ania Wisniak, Giovanni Piumatti, Andrew S. Azman, Stephen A. Lauer, Hélène Baysson, David De Ridder, Dusan Petrovic, Stephanie Schrempft, Kailing Marcus, Sabine Yerly, Isabelle Arm Vernez, Olivia Keiser, Samia Hurst, Klara M. Posfay-Barbe, Didier Trono, Didier Pittet, Laurent Gétaz, François Chappuis, Isabella Eckerle, Nicolas Vuilleumier, Benjamin Meyer, Antoine Flahault,

Laurent Kaiser, and Idris Guessous, "Seroprevalence of anti-SARS-CoV-2 IgG antibodies in Geneva, Switzerland (SEROCoV-POP): a population-based study," *The Lancet*, August 2020, *396* (10247), 313–319.

Theel, Elitza S., Julie Harring, Heather Hilgart, and Dane Granger, "Performance Characteristics of Four High-Throughput Immunoassays for Detection of IgG Antibodies against SARS-CoV-2," *Journal of Clinical Microbiology*, July 2020, *58* (8).

USFDA, U.S. Food and Drug Administration, "Serology Test Evaluation Report for "Architect i1000 SARS-CoV-2 IgG" from Abbott," Technical Report 2020.

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

	Item No	Recommendation
Title and abstract	1	(*a*) Indicate the study's design with a commonly used term in the title or the abstract **pg. 2**
		(*b*) Provide in the abstract an informative and balanced summary of what was done and what was found **pg. 2, 3**
Introduction		
Background/rationale	2	Explain the scientific background and rationale for the investigation being reported **pg. 5**
Objectives	3	State specific objectives, including any prespecified hypotheses **pg. 6**
Methods		
Study design	4	Present key elements of study design early in the paper **pg. 6**
Setting	5	Describe the setting, locations, and relevant dates, including periods of recruitment, exposure, follow-up, and data collection **pg. 6-8**
Participants	6	(*a*) Give the eligibility criteria, and the sources and methods of selection of participants **pg. 7-8, supplement**
Variables	7	Clearly define all outcomes, exposures, predictors, potential confounders, and effect modifiers. Give diagnostic criteria, if applicable **pg. 2, 6-9**
Data sources/ measurement	8*	For each variable of interest, give sources of data and details of methods of assessment (measurement). Describe comparability of assessment methods if there is more than one group **pg. 7-8, supplement**
Bias	9	Describe any efforts to address potential sources of bias **pg. 8-11, supplement**
Study size	10	Explain how the study size was arrived at **pg. 7-8**
Quantitative variables	11	Explain how quantitative variables were handled in the analyses. If applicable, describe which groupings were chosen and why **pg. 8-11 supplement**
Statistical methods	12	(*a*) Describe all statistical methods, including those used to control for confounding **pg. 8-11, supplement**
		(*b*) Describe any methods used to examine subgroups and interactions **pg 8-11**
		(*c*) Explain how missing data were addressed **pg. 9, 10, supplement**
		(*d*) If applicable, describe analytical methods taking account of sampling strategy **pg. 9, 10 supplement**
		(*e*) Describe any sensitivity analyses **pg. 9, 10, supplement**
Results		
Participants	13*	(a) Report numbers of individuals at each stage of study—eg numbers potentially eligible, examined for eligibility, confirmed eligible, included in the study, completing follow-up, and analysed **pg. 7, 8**
		(b) Give reasons for non-participation at each stage **pg. 7, 8, 10**
		(c) Consider use of a flow diagram **NA**
Descriptive data	14*	(a) Give characteristics of study participants (eg demographic, clinical, social) and information on exposures and potential confounders **pg. 6-7, 15**
		(b) Indicate number of participants with missing data for each variable of interest **pg. 7, 8, 10**
Outcome data	15*	Report numbers of outcome events or summary measures **NA**
Main results	16	(*a*) Give unadjusted estimates and, if applicable, confounder-adjusted estimates and their precision (eg, 95% confidence interval). Make clear which confounders were adjusted for and why they were included **pg. 11, 12**

		(*b*) Report category boundaries when continuous variables were categorized **NA**
		(*c*) If relevant, consider translating estimates of relative risk into absolute risk for a meaningful time period **NA**
Other analyses	17	Report other analyses done—eg analyses of subgroups and interactions, and sensitivity analyses **pg. 9, 10, 12, figure 2, supplement**
Discussion		
Key results	18	Summarise key results with reference to study objectives **pg. 13, 15**
Limitations	19	Discuss limitations of the study, taking into account sources of potential bias or imprecision. Discuss both direction and magnitude of any potential bias **pg. 13, 14**
Interpretation	20	Give a cautious overall interpretation of results considering objectives, limitations, multiplicity of analyses, results from similar studies, and other relevant evidence **pg. 14, 15**
Generalisability	21	Discuss the generalisability (external validity) of the study results **pg. 10, 15**
Other information		
Funding	22	Give the source of funding and the role of the funders for the present study and, if applicable, for the original study on which the present article is based **pg. 19**

*Give information separately for exposed and unexposed groups.

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