Journal of Development Economics 159 (2022) 102988

Contents lists available at ScienceDirect

Journal of Development Economics

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

Regular article

Using household rosters from survey data to estimate all-cause excess death rates during the COVID pandemic in India

Anup Malani [a,b,*], Sabareesh Ramachandran [c]

[a] University of Chicago Law School, 1111 E. 60th Street, Chicago, 60637, IL, USA
[b] NBER, 1050 Massachusetts Ave., Cambridge, 02138, MA, USA
[c] University of California San Diego, 9500 Gilman Dr., La Jolla, 92093, CA, USA

ARTICLE INFO

JEL classification:
I10
I14
J10

Keywords:
SARS-Cov-2
COVID
Pandemic
India
Mortality rate
Excess deaths
Household survey

ABSTRACT

Official statistics on deaths from COVID undercount deaths due to lack of testing. In developed countries, death registries are used to estimate excess deaths due to COVID during the pandemic. However, few developing countries had complete death registries even before the pandemic and the pandemic further stressed administrative capacities. As a substitute, we estimate all-cause excess deaths in India using the member rosters of a large, representative household panel survey. We estimate roughly 4.2 million excess deaths during the pandemic through February 2022. We cannot demonstrate causality between COVID and deaths, but the timing and age structure of deaths is consistent with the COVID pandemic and excess deaths are positively correlated with reported infections. Finally, we find that excess deaths were higher among higher-income persons and were negatively associated with mobility. The methods in this paper can be used in countries with a household panel to measure health-related demographic indicators.

0. Introduction

COVID is the largest global pandemic since the 1918 flu. According to official reports, over 600 million people have been infected and 6.4 million have died (Johns Hopkins University and Medicine, 2022). Even these remarkable numbers, however, may be an undercount, especially in developing countries. Serological surveys suggest that 20–100 times more people have been infected and have antibodies than have been tested and counted in official reports (e.g., Malani et al., 2020; Mohanan et al., 2021). Likewise, reported deaths may be underreported due to low testing rates, low capacity to register deaths, and possible manipulation of death records (e.g., Rukmini, 2021).

It is important to accurately measure the number of people infected and dead due to the pandemic in order to estimate the real impact of the pandemic on people's lives. Accurate measurement is also necessary for estimating parameters, such as the infection mortality rate, the impact of non-pharmaceutical interventions, and efficacy of vaccine campaigns, that guide pandemic response.

Total excess deaths during the pandemic is one alternate measure of the mortality risk from the pandemic. In developed countries, data from accurate death registries have helped calculate excess deaths during COVID (e.g. Woolf et al., 2021). However, less than 65% of countries in Asia and 20% of countries in Africa have death rates from even partial registries available (United Nations, 2021). Even in countries with a partial registry, the death registration process itself may be affected by the lockdown during the pandemic, resulting in error in the excess deaths estimate (Sinha, 2022). Further, governments may have chosen to not report all COVID-related deaths, either to deflect blame away from them or to prevent people from panicking (e.g., Rukmini, 2021).[1]

India is one of the countries hardest hit by the pandemic. Officially, India is ranked second globally in number of infections with over 44 million and third in deaths with nearly 525,000 (Johns Hopkins University and Medicine, 2022). These official COVID-death numbers are widely thought to be undercounts (Gamio and Glanz, 2021). However, excess-death estimates based on India's death registries also likely undercount deaths: only 79% of all deaths were recorded by the registries in 2019 and India ranked 102nd out of 136 countries recently evaluated for completeness of death registries (Karlinsky, 2021).

* Corresponding author at: University of Chicago Law School, 1111 E. 60th Street, Chicago, 60637, IL, USA.
E-mail address: amalani@uchicago.edu (A. Malani).

[1] On top of having incomplete registries, developing countries also often fail to list cause of death. Developed countries' registries typically do include cause of death, although people debate whether COVID-attributed deaths in those registries are deaths where COVID was also present or deaths caused by COVID. The implication is that excess deaths estimates from developing countries are more likely to include deaths not caused by COVID.

https://doi.org/10.1016/j.jdeveco.2022.102988
Received 30 December 2021; Received in revised form 23 September 2022; Accepted 29 September 2022
Available online 7 October 2022
0304-3878/© 2022 Elsevier B.V. All rights reserved.

To address this data gap, we use an alternative source of deaths data – household roster of a large, panel data set – to estimate all-cause excess mortality in India during the pandemic. The data set is the Consumer Pyramids Household Survey (CPHS). Its nationally-representative sample includes roughly 174,000 households with roughly 870,000 current members. The survey is conducted on the same households every 4 months, with a representative quarter of the sample surveyed each month. The survey keeps a roster of all current and past household members and provides reasons for attrition, including death. We count these deaths before COVID to estimate a baseline death rate, and during COVID to calculate excess deaths during the pandemic. An important feature of our data is that it is private and measures death incidentally. This means it is immune to political censorship and is unlikely to have investigator-side bias with respect to death reporting.

In our preferred estimates, the COVID pandemic is associated with 4.2 million excess deaths, roughly 8 times the number of COVID deaths reported.[2] Excess deaths peak in the same months as infections peaked during the two waves that struck India (September 2020 for wave 1 and April–May 2021 for wave 2). Moreover, we find that the second wave experienced significantly greater deaths than the first wave. Although we cannot show causation between the pandemic and excess deaths, we do find that the age-pattern of deaths is COVID-like: deaths rise significantly relative to baseline for those over 60, but decline somewhat for those under 40. We also find a significant correlation between excess deaths in a district and confirmed cases in that district. Although we do not find statistically significant differences in mortality by sex or urban location, we do find that females and rural areas have lower mortality rates. We explore heterogeneity that is not obviously COVID-related and find that excess deaths are higher in families with a higher per-capita income and that this relationship is not entirely explained by the age-gradient on excess death rates.

An important caveat to our analysis is that excess death estimates vary by time interval. In our preferred estimates, the pandemic was associated with 2.2, 4.1, and −1.9 million excess deaths in waves 1 (original variant), 2 (delta), and 3 (omicron) of the pandemic, respectively, and 4.2 million excess deaths overall. This is despite the fact that the third wave was associated with a positive number of reported deaths: although the duration of wave 3 is much shorter than the other 2, at its apogee it claimed roughly 1100 lives daily. We believe – though it is hard to prove – that our negative estimate of excess deaths is due to a displacement effect: waves 1 and 2 claimed lives in 2020 and 2021 that would have died in 2022. Another way to put it is that the pandemic claimed the most vulnerable, who also had the shortest remaining lifespans. It is difficult to estimate displacement because it requires deconvolution, which in turn requires stronger assumptions about the distribution of remaining lifespans among those who died at each age or a lot more data to estimate, so we leave that to future research. A methodological implication of displacement is that estimates of the impact of the pandemic depend on whether one estimates it using data from just 2 waves, or from all 3. In theory, one should use all future data. But that will also tend to capture future events that may impact mortality, a catch-22.

Our use of household rosters from a survey to estimate health-related demographic parameters is possible because the data set we employ is representative, large, and repeatedly surveys the same households. The availability of covariates such as income, caste and occupation also allow us to explore correlates of mortality risk. However, the use of rosters in this manner has shortcomings that we must address. The main problem is that the survey measures whether a death occurred since the last time the household was surveyed, but does not measure exactly when the death occurred.[3] We primarily tackle this by restricting the sample to individuals who are observed in consecutive rounds and attribute deaths to the median month between the current and last completed survey. This interprets the death rate reported in month t as a moving average of death rates from months $t-3$ to t. We discuss other methodological issues with using household rosters to track demographics in Appendix.

We compare our estimates with estimates of excess deaths using the Civil Registry System (CRS) data, the official death registries, from 12 states (Banaji and Gupta, 2021) through 2 waves of the pandemic in India.[4] Our estimates of excess deaths are somewhat larger than the estimate from CRS data. This may be due to the incompleteness of the India's death registries (Deshmukh et al., 2021). We also compare our estimates to those from the US. Whereas we report a 14.16% increase in death rates during the pandemic, the US reports a roughly 22% (Woolf et al., 2021) increase in excess deaths. Our preferred estimates of deaths are lower than US estimates, though that could be because India has a relatively younger population[5] and COVID has greater infection fatality rate among the elderly.

Our main contribution is to provide novel estimates from India to a growing literature on excess deaths from COVID. Unlike studies from countries that have reliable death registries (Rossen et al., 2020; Woolf et al., 2021; Kontopantelis et al., 2021), it examines a country with unreliable registries.

Alternative estimates from India employ data on registered deaths but scale them up based on their degree of undercounting (Anand et al., 2021; Deshmukh et al., 2021). However, these estimates are only available for a third of India's 29 states. Deshmukh et al. (2021) also provides national estimates using other representative surveys. The main advantage of using CPHS over these other surveys is that CPHS has better temporal coverage and tremendous detail on the deceased, providing opportunities to explore whether excess deaths have "COVID-like" features and heterogeneity in death rates. A third approach is to apply estimates of infection fatality rates outside of India to estimates of infection rates in India (Anand et al., 2021). The problem with this approach is that India may not have the same infection fatality rates as other countries, just as it does not have the same rates of death from other diseases. Moreover, there are conflicting estimates of seroprevalence in the same place due to antibody decline and many locations lack any seroprevalence estimates. So infection rates have wide confidence bars. One other, contemporaneous paper employs CPHS to estimate excess deaths (Anand et al., 2021). We explore some of the data problems with CPHS a bit more than that paper. Moreover, we explore how death rates vary with a wider array of variables, such as incomes. Overall, our estimates are higher than, but not tremendously out of line with available excess-death estimates from other sources.

A second contribution is to show how best to use rosters from household surveys such as CPHS to measure items, like death, migration and marriage, that are implicitly measured by household rosters, in India. To some extent, the problems associated with using CPHS to measure roster-events, especially the timing of these events, are also a problem for measuring roster-events in surveys other than CPHS and outside India. Thus, our methods for addressing that may be relevant for counting roster-events from other surveys.

[2] The officially reported number of deaths till 28 February 2022 was 514,045.

[3] It does not measure why the death occurred either. Therefore, it only allows us to measure excess deaths. In a separate project, we are conducting verbal autopsies on all reported deaths in the survey during 2019–2021 to determine which deaths were plausibly due to COVID.

[4] Estimates from CRS are not available for wave 3 yet.

[5] The median age is 27.6 in India and 37.7 in the US. The percent of population above 64 is 9.8% in India, 25.6% in the US (Ritchie and Roser, 2019).

1. Background

According to the Global Burden of Disease (GBD) project (Fig. G.7) India had a death rate of roughly 0.7% (7 deaths per 1000 persons per year) in 2019, approximately 9.5 million deaths in a population of 1.4 billion (Vos et al., 2020). GBD shows an uptick in the death rate from 2018 to 2019, a pattern also evident – though more pronounced – in our CPHS data.

SARS-CoV-2 hit India in three waves (Fig. G.8A). The first cases were reported on 27 January 2020 (Andrews et al., 2020). The first wave peaked in September 2020, with almost 100,000 confirmed cases and 1,100 deaths daily. The second wave peaked in April 2021, with roughly 400,000 confirmed cases and 4000 deaths daily (www.covid19india.org, 2021). The third wave peaked in January 2022, with 300,000 confirmed cases and 1100 deaths daily (www.covid19bharat.org, 2022).

India imposed a national lockdown from 24 March to 1 June 2020, well before wave 1 (Fig. G.8B). Google mobility statistics show mobility fell 40% relative to January 2020 levels during that lockdown. After that, lockdowns were local and driven by states. But by the peak of wave 1, mobility had returned to about 15% below January 2020 levels. There were local lockdowns and a reduction of mobility during wave 2, but the decline was not as severe as during wave 1.

Official numbers on cases and deaths should be taken with a grain of salt. Confirmed cases undercount actual infections, at different rates over time. Perhaps 90% of cases were asymptomatic and unlikely to be tested (Waghmare et al., 2021). Per-capita testing rates in India were low relative to developed countries (Ritchie et al., 2020). Testing rates increased dramatically from wave 1 to 3, so the higher case counts may partly be due to testing not cases (www.covid19bharat.org, 2022).

Reported COVID death rates may also be lower than true death rates (e.g., Rukmini, 2021). First, many deaths in India, especially those outside the hospital setting, are not officially recorded (Gettleman et al., 2021). Second, not all dying individuals are tested for COVID (Bedi, 2022). Further, even individuals dying after a positive COVID test are sometimes recorded as a non-COVID death because they have co-morbidities that could have been the cause of death (Prasad, 2021). While some such deaths are not causally COVID deaths, some of them may be but are missed.

Because many COVID-attributable deaths are not labeled COVID deaths, researchers have examined all-cause mortality to gauge the impact of the pandemic. Typically, the level or projected trend of deaths pre-pandemic is compared to the level of deaths during the pandemic. In India, all-cause mortality is recorded by each state's Civil Registry System (CRS). The main alternative is the Sample Registration System (SRS), which calculates death rates based on a representative, 1% sample of the population.

Each source has its problems. The CRS has three problems. One, not all deaths are registered. For example, in 2017, among big states, the reporting rate was 63.5% in Jammu & Kashmir and 76.4% in Bihar (Rao and Gupta, 2020). Two, while reporting rates are improving over time, this trend complicates estimation. An increase in death rates could be due to better reporting or to an actual increase. Three, CRS reports with delay (Ravi, 2021). For example, only 14 (of 28) states have CRS data currently available (Rukmini, 2021).

The SRS also has problems. First, while the CRS is delayed a few months, the SRS is typically delayed 2 years. We may not get SRS estimates of COVID-period deaths until 2023. Second, even the SRS is misses about 12% of deaths (Gerland, 2014). Third, CRS and SRS can diverge. In 2017, the ratio of CRS to SRS deaths range from 38% in Uttar Pradesh to 124% in Tamil Nadu (Rao and Gupta, 2020). The CRS number can be higher than the SRS number not only because SRS may be an underestimate, but because CRS will report the death of a resident of one state in another state if they went to that other state for medical care and died there (Ravi, 2021).

Even if one can measure excess deaths, not all are linked to COVID. The death rate due to COVID is typically captured by two epidemiological parameters. The case fatality rate (CFR) is the number of confirmed deaths divided by the number of confirmed COVID cases. This is not a very useful statistic. Both numerator and denominator are undercounted. Moreover, CFR may reflect testing rates and selection into testing as much as harm from the disease. A better alternative is the infection fatality rate (IFR), i.e., COVID deaths divided by all COVID infections (rather than just confirmed infections).

Initial efforts to calculate the IFR used serological studies to estimate the denominator. However, they used official death counts as the numerator (Malani et al., 2020; Mohanan et al., 2021; Malani et al., 2021). Because those counts are also underestimates, correcting only the denominator likely led to an underestimate of the IFR. We will not be able to correct that in this paper, as all-cause excess deaths may include deaths not directly related to COVID. However, it does provide some insight into how off prior IFR estimates might be.

2. Methods

2.1. Data

2.1.1. Consumer Pyramids Household Survey

Our primary data source is the Centre for Monitoring the Indian Economy's Consumer Pyramids Household Survey (CPHS), a large, representative,[6] panel survey of Indian households. The sample is based off the 2011 Indian Census and representative at the level of strata defined as homogeneous regions × urban status and at the national level. A homogeneous region is a cluster of similar districts within a state.[7] Sample households are visited every 4 months, with each 4-month period called a round. However, a nationally-representative subsample of households are sampled each month. CPHS started in January 2014 and the latest data we could access are from April 2022.[8] We will start our analysis with data from 2015 because it took a year for the sample in CPHS to stabilize. As we shall explain, we attributed deaths reported in month τ to month $t = \tau - 2$; thus our analysis will focus on deaths attributed to the months between January 2015 and February 2022.

Although the purpose of the CPHS is to measure household economic characteristics, it also maintains a meticulous household roster. The roster records whether there is a death in the household since the last time a household was surveyed, typically 4 months earlier. CPHS also provides data on the demographics and income of each household member, location at the district level, and whether the household resides in a rural area, defined as a village in the 2011 Indian Census. We use these data to explore heterogeneity of death rates.

2.1.2. COVID cases and mobility data

We obtain district-by-day level data on confirmed cases from www.covid19bharat.org. We obtain estimates of daily infections by scaling up confirmed-case curves with estimates from a serological survey, as we explain in Appendix C.

We obtain mobility data from Google's Community Mobility Reports (Google, 2021). The units are percent relative to a baseline that is the median value for the corresponding day of week during the 5-week period Jan 3–Feb 6, 2020. Google reports 6 measures of mobility based on location; we take an average of the 5 measures other than mobility at home because home mobility rises during the pandemic.

[6] CPHS has been criticized for not being representative of poor populations (Dreze and Somanchi, 2021). We address this issue in Appendix.

[7] The sampling method is explained in Appendix.

[8] Whereas data through February 2022 are in the People of India file of the CPHS, the March and April 2022 roster is obtained from December 2021 income file, which is released in April 2022.

Because our death data are reported monthly, we average daily cases, infections and mobility over each month. We do not have case and mobility data prior to February 2020. Therefore, we assume cases and infections are 0 and that mobility is 100% before that date.

2.2. Data issues with CPHS

Using survey data rather than death registration to measure mortality rates raises a number of data cleaning problems. First, survey response rates fell during the pandemic, particularly during India's lockdown. Second, there may be selection bias in non-response. Specifically, non-response may be a function of whether a household experienced a death. Third, there appears to be a level jump in the death rate in 2019, prior to the pandemic. Fourth, CPHS has been criticized for not being representative of poor populations (Dreze and Somanchi, 2021). Finally, the CPHS does not report the precise date death occurred. We address the first 4 concerns in Appendix and address the timing issue here.

The date on which deaths are "observed" in CPHS is not necessarily the date the death occurred. Households do not report deaths to the CPHS *when* those deaths occur, but some time later when the household is surveyed. Nor does CPHS ask when deaths occurred. So, if a household answered two surveys in a row 4 months apart and reported a death in the second survey, all we know is the death occurred during the intervening 4 months. If the household skipped n surveys between surveys they answered, then a death reported in the last survey occurred sometime in the $k = 4(n + 1)$ months in between responses.

Our preferred solution is to reallocate death to the midpoint between answered surveys.[9] So if there is a gap of k months between surveys the household answered, then a death reported on month τ is re-allocated to month $t = \tau - (k/2)$. This solution to the timing problem is simple, but it gets the timing of deaths a bit off in the way a moving average would because it smooths out the jump in rates, in part to periods before and in part to periods after the jump.

In order to reduce error in the month to which a reported death is assigned, our analysis will focus on the subsample of households in each round that responded to the CPHS in the previous round, which was 4 months earlier.[10] Therefore, $k = 4$ and deaths reported at in month τ are reallocated to month $t = \tau - 2$, i.e., 2 months earlier.[11] As a result, though our CPHS sample runs through April 2022, our last month of attributed deaths is February 2022. A problem with restricting the sample to those who also respond to the previous round of surveys is that we have a slightly smaller sample on which we estimate excess deaths and there is a risk of selection bias if a household's decision to respond to the CPHS is not independent of whether there was a death in that household. Therefore, in Appendix, we show results that include a sample irrespective of duration between responses, and the results are similar.

[9] We explore a second solution in Appendix E: estimating the death rate by asking, how much would the true death rate have to have changed for the observed death rate to have changed as much as it did since the last month. We also explain that that solution is highly sensitive to measurement error, so we use the simpler solution in the main text.

[10] To put it another way, our primary analysis will not use reports in round from households that did not respond in the previous round, but may have responded 2 or more rounds ago.

[11] Of the households that respond in a given round, on average 82% responded in the last round. This rate falls during the lockdown, a topic we discuss in Appendix B.1. We address this by examining how excess death rate estimates vary across the number of rounds a household has skipped (Table B.6). More on this in Section 3.

2.3. Definition of pandemic period

We define the pandemic period as February 2020 to the last date of data (February 2022) because India's first confirmed cases are on 27 January 2020 (Andrews et al., 2020); the first wave as February 2020 to February 2021; and the second wave as March 2021 to November 2021; and the third wave as December 2021 to February 2022. The CPHS data are coded at the monthly level, so we cannot more finely define the pandemic or waves. In robustness exercises, we vary the start and stop dates +/− 2 months for our preferred estimation strategy.

2.4. Estimating excess deaths

We estimate excess deaths in two steps. First, we predict monthly death rates ($\hat{y}_{it}$) in the absence of the pandemic using data from before the pandemic. These predicted death rates are either the average level of deaths during a pre-pandemic baseline period or a trend in those deaths during the baseline period. Second, we regress the difference between an individual indicator for death (y_{it}) and predicted individual death rate, $y_{it} - \hat{y}_{it}$, during the pandemic on a month, pandemic-wave or entire pandemic time fixed effects. The coefficients on the time fixed effects give us estimates of excess-death rates for various time periods. We explain both steps below.

2.4.1. Predicting death rates

We consider 4 possible counterfactual death rates: 2 parameterizations of the baseline × 2 baseline periods.

Baseline parameterization. The baseline can be parameterized as either a level or trend. The level baseline is the mean death rate during the baseline period: $\hat{y}_{it} = (1/M_B)\sum_{s\in B}[(1/N_s)\sum_i y_{is}]$, where t and s index months, B is the baseline period, M_B is the number of months in the baseline period, and N_s is the number of people in the subsample in month s. In other words, we calculate the death rate in each month during the baseline and then take the average of monthly death rates through the baseline.

The trend baseline is computed in two steps. First, we estimate the trend during the baseline period with the following regression:

$$y_{it} = \alpha + \beta(t - t_0) + e_{it}, \tag{1}$$

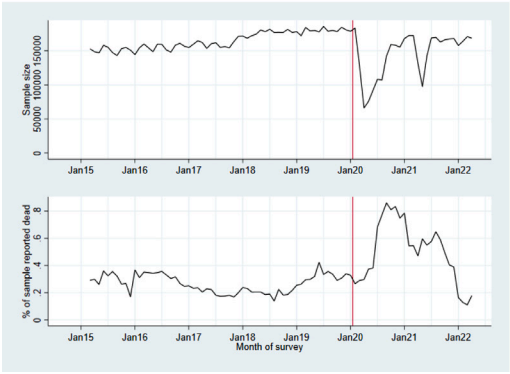
where the sample includes $t \in B$, i.e., months during the baseline period. Second, we predict y_{it} for each month t after the baseline period.

Although there are advantages to using a trend as the baseline, our preferred specification will use levels. The demographic literature typically calculates excess deaths using a baseline trend, for two possible reasons. One is to account for population growth. The literature uses data from death registries, which have information on deaths, but not population. Our data, however, include information on births and deaths. We look at a fixed sample of households and add persons when they are born and lose them when they die.[12] The other reason is to adjust for changes in death rates. Because death rates change slowly (in the absence of disasters), this argument is only important when making long-term projections. Here we focus on projections for a period less than 2 years, so changes in baseline death rates are unlikely to be material. Although we will present results for both baseline parameterizations, our preferred specification will use levels because neither argument for trends in strong in our application and using levels is simpler.

Baseline period. The baseline period can either be 2015–2019 or 2019 alone. The former uses more data, but the latter is more recent. Although we will report results with both periods, our preferred specification is using 2019 alone. First, when using a level baseline,

[12] In theory, there could be addition of household and attrition of households. However, those changes are orthogonal to population growth. Moreover, our results stand even if we hold the sample of households constant.

Panel A: Sample size and households reporting a death from March 2015 - April 2022.



Panel B: Time series of death rates from month fixed effects, January 2019 - February 2022.

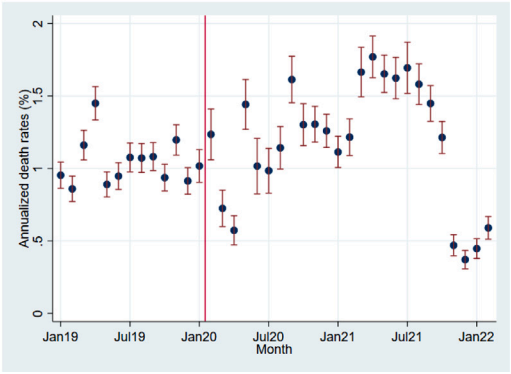


Fig. 1. Sample size, deaths, and death rates.
Notes. Panel A: The sample includes all responding households regardless of how frequently they respond to a survey. Deaths reported in month t are not allocated to a prior month. The y-axis in the lower graph is the proportion of individuals responding in that month who are reported to be dead. The data are not weighted to be representative. Panel B: The sample includes households that respond in consecutive rounds. Observations on individuals are weighted to be nationally representative even with non-response. Each point is the weighted mean death rate in a month and each whisker is the 95% confidence interval on that mean. Both panels: The red line demarcates the start of the pandemic in February 2020. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

using the latest year's data mitigates some of the error from not using a trend baseline. Second, and more importantly, there is a jump in death rates from 2017 to 2019 in the CPHS. Other data sources (e.g., the CRS and the Global Burden of Disease) also report a jump in 2019, though the magnitude of the jump is larger in CPHS. The jump is surprising because age-conditional mortality rates typically trend down in the long run and there is no known reason for the jump in 2019 across data sources. Using only 2019 as a baseline side-steps this issue.

2.4.2. Excess-death estimates

With predicted death rates in absence of a pandemic in hand, we obtain a monthly or by-wave time-series of excess estimates a regression of the form:

$$y_{it} - \hat{y}_{it} = \sum_k \gamma_k p_{k,t} + u_{it} \tag{2}$$

using data from the baseline period and the pandemic period (the combined periods denoted by P). The sample includes individuals from households that responded at both month $t-2$ and $t+2$. Here y_{it} is an indicator for whether individual i who was reported alive in month $t-2$ was reported dead in $t+2$. Our treatment variables are indicators $p_{k,t}$ for whether month t is in period k, where k can index months or waves.

Our main specification sets the treatment period to be the whole pandemic period ($S = \{\text{pandemic}\}$). Some specifications will replace that with an indicator for each month or wave during the pandemic. Other specifications will, e.g., add as regressors age or income indicators and interactions between time period indicators and age or income indicators. Our estimates of excess deaths come from the coefficients on these time and characteristic indicators (e.g., $\{\gamma_k\}$). Standard errors are clustered at the village/ward × month level to account for correlation in reporting of deaths within a locality.

3. Results

Raw data from CPHS aggregated to the national level and without adjusting for the timing-of-death suggests a jump in the death rate during the COVID pandemic. Fig. 1A shows the sample size between March 2015 and April 2022 and presents death rate as of the date the deaths are reported (not the date they occurred). There is a drop in response rates during India's lockdown, which forced the CPHS to sample only about half its households—an issue we address in Appendix. There is a large rise in death rates during wave 1, a smaller rise in wave 2, and a very small rise in wave 3. The large rise in Wave 1, however, is partly due to delayed reporting of deaths by households that were omitted from the survey during the lockdown. Actual deaths during wave 1 were likely more spread out. The small wave 3 rise is consistent with the small relative spike in deaths in official death statistics (Fig. G.8A).

When we address the timing-of-death problem, we obtain a time series that shows a more moderate increase in death rates during COVID (Fig. 1B). Specifically, we focus on the sample that includes only households that respond in consecutive rounds, to reduce imprecision from reassigning the date of death (i.e., keep the moving average at 4 months rather than 8 or longer). We time shift observed deaths back to month $t-(k/2) = t-2$ for $k = 2$. Finally, we estimated weighted mean death rates by month, where the weights make responding households nationally representative.[13]

We find that death rates drop in April and May 2020, around the time of the national lockdown, but are at or above the 2019 average in other months of the pandemic. There are four spikes during the pandemic. One spike is June 2020, when lockdown is released, another is September 2020, when wave 1 peaks, the third is in March–May 2021, when wave 2 peaked, and the last is in February 2022, when wave 3 peaks. These spikes are significantly greater than adjacent months and, except for the February 2022 peak, above all but one 2019 month. The wave 1 and 2 spikes are greater than even the highest 2019 peak. The February 2022 rate is actually below any 2019 monthly rate, perhaps due to displacement of 2022 deaths into waves 1 and 2, a topic we will address later.

3.1. Excess-death estimates

The estimated excess-death rate during the pandemic depends on our baseline specification (Table 1). When we define the baseline as the mean death rate during the period 2015–2019 (column 1), our baseline death rate is 0.787%, not far off from the Global Burden of disease estimate. But excess-death rates during the pandemic are 0.399%, which represents a hard-to-believe increase of over 50%. When we use only 2019 level as the baseline (column 2), the baseline death rate rises to 1.038%, which is implausibly high. However, our estimate of the excess-death rate during the pandemic becomes 0.147%, a relative

[13] This means we use both the weight that makes the sample representative and the non-response factor that makes responding households representative of the sample.

Table 1
COVID death rate prior to the pandemic and the excess-death rate during the pandemic.

	Excess deaths			
	(1)	(2)	(3)	(4)
Pandemic period:				
Annualized rate (%)	0.399***	0.147***	0.235***	0.0891**
	(0.0304)	(0.0400)	(0.0281)	(0.0281)
Deaths (millions)	11.3	4.2	6.7	2.5
	(0.9)	(1.1)	(0.8)	(0.8)
By wave during pandemic period:				
Wave 1 deaths (millions)	5.9	2.2	3.8	1.6
	(0.6)	(0.8)	(0.6)	(0.6)
Wave 2 deaths (millions)	6.7	4.1	4.9	3.4
	(0.5)	(0.6)	(0.5)	(0.5)
Wave 3 deaths (millions)	−1.1	−1.9	−1.8	−2.2
	(0.1)	(0.2)	(0.1)	(0.1)
Baseline period:				
Annualized rate (%)	0.787	1.038	0.790	1.044
Specification:				
Baseline period	2015–19	2019	2015–19	2019
Trends	No	No	Yes	Yes

Notes. The sample includes households that respond in consecutive rounds. Observations on individuals are weighted to be nationally representative even with non-response. The regression model is (2). In the first two columns we assume a constant counterfactual death rate and in the last two columns we assume the counterfactual death rate has a linear trend. Excess deaths are calculated for the 25-month period from Feb 2020 to Feb 2022. In the first and third columns the baseline period include Jan 2015–Jan 2020. In the second and fourth columns the baseline period includes Jan 2019–Jan 2020. Excess deaths are calculated for all of India assuming the projected 2020 population of 1.361 billion people.

increase that is somewhat higher but in line with the relative increase in excess deaths from the US (Woolf et al., 2021).[14]

If we switch from a level baseline to a trend baseline, we see similar dynamics with baseline and excess death estimates: a baseline trend using 2015–2019 (just 2019) data yields a reasonable (implausibly-high) baseline but implausibly-high (reasonable) excess death estimate. The reason is that there is very little difference in trend in 2015–2019 versus just 2019 data.

Breaking the pandemic down into waves we see a common pattern across baselines: excess deaths in wave 2 are nearly double the number in wave 1 and negative in wave 3. The wave 2 result is unsurprising: wave 2 with the delta variant hit India harder than wave 1 with the original variant. While wave 1 seemed to last longer, wave 2 had a sufficiently high peak that it overcame the shorter duration. The wave 3 result is surprising, but can be explained with displacement. The additional people who died in waves 1 and 2 may have died in 2022 anyway. Thus, the excess deaths during wave 3 reflect deaths due to COVID cases during wave 3 and the absence of deaths from people who would have contributed to the baseline number of deaths in the absence of COVID.

Our estimate of excess deaths through the entire pandemic does not equal the sum of our estimates by wave. The former is calculated from estimates with a fixed effect for the entire pandemic period, while the latter from estimates in a separate regression with fixed effects for each wave. The two calculations are typically off by 0.2 million deaths. The reason is that wave-wise deaths are calculated by taking the wave-specific excess death rate estimates and multiplying by the fraction of a year in each wave and the population of India just before the pandemic, so small discrepancies between estimates for duration-weighted wave-wise rates and the pandemic-wise rate can generate non-trivial gaps in projections of number of excess deaths.

[14] Our estimate of the excess-death rate falls a bit to 0.174% or 0.178% (columns 2 and 3 of Table G.11) if we include fixed effects for homogeneous region or district, respectively.

Our preferred estimates use 2019 levels as a baseline. While 2019 levels are higher than 2017 and 2018, they are not unheard of: 2015 to 2016 death rates are similar to 2019. While the 2019 upswing in death rates in CPHS is larger, there is an upswing in death rates in even Global Burden of Disease estimates (Fig. G.7). Our preferred estimates suggest that there were 4.2 million deaths over the entire pandemic, with 2.2 million deaths in wave 1, 4.1 in wave 2, and −1.1 in wave 3.

Our main specification includes only households that respond in consecutive rounds of the survey (roughly 82% of the sample) to mitigate error from our solution to the timing-of-death problem. If we include in our estimation households that skip rounds, we obtain somewhat higher estimates of excess mortality during COVID: 0.181 with responders that skip up to 1 round, 0.216 if up to 2 rounds (Table B.6).[15]

Our estimate of excess deaths falls if we move forward our estimated start date for the pandemic, possibly because it is counting months before any confirmed cases as pandemic months (Table E.7). It rises as we move the start date back 2 months, in part because deaths fall during the lockdown, which occurs in April and May 2020, and pushing back the start date moves the low death rate months out of the pandemic period.

3.2. Heterogeneity of death rates

We explore heterogeneity in excess deaths, first, along lines that would help gauge whether our estimates are credibly picking up the effect of COVID. We then look at other factors that are policy-relevant.

3.2.1. COVID-related factors

Age. Excess deaths follow a COVID-like pattern with respect to age (Fig. 2 below and Table F.10 in Appendix). Excess deaths rates are positive for ages 50+ and significantly so for ages 60+, and insignificant and close to zero for lower ones. This right-skew is somewhat greater in wave 2 than in wave 1. Consistent with displacement, the negative excess deaths in wave 3 also skew towards higher ages.

Gender and location. While we see nominally COVID-like patterns with respect to sex and urban v rural location, these differences are not always statistically significant. Estimated CFR and IFR is greater among males (Green et al., 2021; Pastor-Barriuso et al., 2020; Nguyen et al., 2021) and we find roughly 10% higher excess death rates for males. However, the difference is not statistically significant (Table G.12A). Likewise, prior studies have shown greater infection rates in urban areas (Mohanan et al., 2020; Malani et al., 2021; Rader et al., 2020; Stier et al., 2020) and we find a higher excess death rates in urban areas. However, the difference is not significant except during wave 2, which coincides with a much larger excess rate than other waves (Table G.12B).

Infections. Excess deaths are positively correlated with confirmed cases and with infection. Table 2 reports the results of an individual-level regression based on (2), except that we replace the pandemic indicator with confirmed cases or infections. Cases and infections are reported as monthly averages at the district level. Infections are the same as cases, but scaled by seroprevalence estimates. We find that deaths are significantly correlated with cases or infections, even when we add controls for monthly average mobility at the district level. This finding increases the credibility of the claim that excess deaths during the pandemic picks up the effect of COVID. However, one should not interpret the coefficients as case fatality rates (CFR) or infection fatality rates (IFRs) as they may capture COVID deaths that were not included in case or infection counts and include non-COVID deaths.

[15] This seems inconsistent with the nature of non-response bias we discussed in Appendix B.2. Recall, however, that the sample in which responders had higher death rates was restricted to those who responded in round t and t+8, about 64% of the sample. If we include those that did not respond at both those times, the consecutive responders actually have lower death rates.

Table 2
Correlation of death rates with cases, infections, and mobility.

	Annualized death rates (%)				
	(1)	(2)	(3)	(4)	(5)
Infections (sero scaled)	0.0325***	0.0320***			
	(0.00493)	(0.00511)			
Cases			7.987***	7.066***	
			(1.720)	(1.791)	
Mobility		−0.00000655		−0.0000139*	−0.0000270***
		(0.00000622)		(0.00000551)	(0.00000580)
2019 mean	1.063***	1.054***	1.072***	1.061***	1.100***
	(0.0256)	(0.0267)	(0.0231)	(0.0231)	(0.0213)
N	2540551	2496679	3485020	3439372	3439372

Notes. Estimates in columns 1 and 2 are from a regression of death rates against sero-survey scaled infections from an SIR model. The independent variable is the proportion of individuals estimated to have been infected in the months considered. Estimates in columns 3 and 4 are from a regression of death rates against officially reported COVID cases. The independent variable is the proportion of people who tested positive for the virus in the months considered. We control for mobility in columns 2 and 4. We use the mobility data provided by Google measured in percentage points change from baseline. Estimates in column 5 are from a regression of death rates against Google mobility. Standard errors clustered at the village/ward × month level are reported in parentheses. */**/*** indicate $p < 0.05/0.01/0.001$.

Table 3
Annualized death rates by income groups.

	Annualized death rates (%)				
	(1)	(2)	(3)	(4)	(5)
Pandemic	0.0406				
	(0.0603)				
Pandemic × 2nd tercile	0.0918		0.112		0.148
	(0.0731)		(0.0731)		(0.0803)
Pandemic × 3rd tercile	0.287***		0.289***		0.244**
	(0.0819)		(0.0826)		(0.0919)
Wave 1		−0.0222			
		(0.0711)			
Wave 1 × 2nd tercile		0.180		0.190*	
		(0.0919)		(0.0921)	
Wave 1 × 3rd tercile		0.422***		0.322**	
		(0.107)		(0.107)	
Wave 2		0.373***			
		(0.0837)			
Wave 2 × 2nd tercile		−0.0532		−0.0272	
		(0.0987)		(0.0976)	
Wave 2 × 3rd tercile		0.203		0.247*	
		(0.118)		(0.117)	
Wave 3		−0.756***			
		(0.0715)			
Wave 3 × 2nd tercile		0.269**		0.322**	
		(0.0959)		(0.0998)	
Wave 3 × 3rd tercile		0.334***		0.522***	
		(0.101)		(0.105)	
2019 mean	1.160***	1.160***			
	(0.0463)	(0.0463)			
2nd tercile	−0.172**	−0.172**	0.000932	0.000932	−0.0284
	(0.0540)	(0.0540)	(0.0540)	(0.0540)	(0.0548)
3rd tercile	−0.232***	−0.232***	−0.372***	−0.372***	−0.403***
	(0.0567)	(0.0567)	(0.0581)	(0.0581)	(0.0597)
Age controls	No	No	Yes	Yes	Yes
Occupation and caste controls	No	No	No	No	Yes
N	3713943	3713943	3713943	3713943	3155137

Notes. Estimates are from a regression model based on Eq. (2), with the addition of an income tercile indicator and income tercile indicator interacted with the pandemic or wave indicator. For each individual we calculate the income per capita in 2018. We compute the household's income percentile in their homogeneous region and region type (urban/rural). Households between 33 and 67 percentiles are in income tercile 2 and households between 67 and 100 percentiles are in income tercile 3. Columns 1 and 2 include no controls. Columns 3 and 4 include as controls indicators for each age category. Column 5 includes as control indicators for each age category, caste category, and occupation category. The sample includes only consecutive observations and is weighted to be nationally representative. Standard errors clustered at the village/ward × month level are reported in parentheses. */**/*** indicate $p < 0.05/0.01/0.001$.

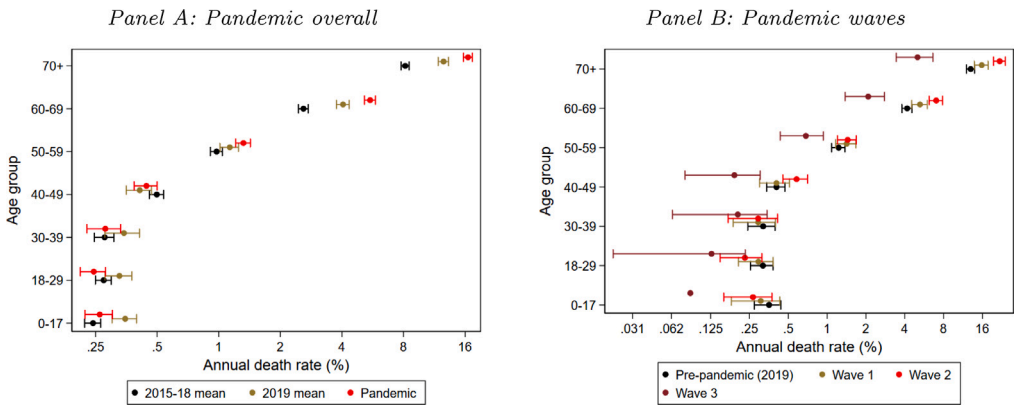


Fig. 2. Age pattern of death rates during the pandemic.
Notes. Estimates are from a regression of an indicator for individual deaths on time period indicators: $y_{it} = \sum_s \beta_s \mathbb{I}(t \in s) + w_{it}$. Here y_{it} is an indicator for whether individual i who responded in month $t-2$ was reported dead in month $t+2$, and s are indicators for different periods. We only include individuals who reported in the prior survey attempt and the current one. We run separate regressions for respondents in each age group listed in the y-axis. Coefficients on period indicators estimate death rates during those periods for the relevant sample population. In Panel A, we use data from 2015 onwards and report coefficients from an indicator for 2019 and for the pandemic period. In Panel B, we use data from 2019 onwards and include indicators for 2019 and the three waves. in Panel B, for age category 0–17, the lower limit of the confidence interval is −0.015. Since the x-axis is log-scaled, this confidence interval is omitted from the graph.

3.2.2. Policy-relevant factors

Mobility. One of the concerns with using excess mortality to measure the harm from the pandemic is that it picks up both the direct effect of COVID infections and the indirect effect of any associated behavioral or policy response. For example, it is possible that the pandemic deterred people from hospitals for fear of getting infected and triggered a lockdown that reduced traffic accidents. We find mixed evidence on the indirect effects of the pandemic.

On the one hand, Fig. 1, which shows a sharp drop in death rates in April and May 2020, suggests that India's national lockdown (March 24–June 1, 2020) was associated with a sharp reduction in deaths. One should be cautious, however, in interpreting this figure because of the timing-of-death problem. In our first solution to this problem, the death rate attributed to, say, April are actually reported in June.

On the other hand, deaths are negatively correlated with Google mobility. Table 2 also reports the results of an individual-level regression based on (2), except that we replace the pandemic indicator with Google's mobility index. Our estimated coefficient in column 5 is that a 10 percentage point reduction in mobility (relative to a baseline of 100 in February 2020) was associated with a 0.00027 percentage point increase in the annualized death rate.

Income. Because CPHS has information on income, we can also compare excess deaths by income. Serological surveys suggest that, in cities, slums were more affected in wave 1 (Malani et al., 2020). News reports suggest that wave 2 disproportionately affected resident that did not live in slums (Khandekar, 2021). To validate these claims, we add indicators of income terciles and the interaction of those tercile indicators and pandemic or wave indicators to the regression in Eq. (2). This evidence suggests that the pandemic had a bigger mortality impact on the top tercile (Table 3 column 1) overall and in waves 1 and 2. In wave 3, the middle and top terciles were equally hit. Interestingly, whereas before the pandemic, mortality rates were higher for the lowest tercile, after the pandemic the three terciles have almost equal mortality rates.

The higher excess death in the top tercile of income is not driven by the fact that richer households have older members. To check this, we control for age by adding indicators for 7 different age categories and the interaction of those with pandemic indicators to the regression in the last paragraph. We find that higher income continues to be associated with greater excess death rates during the pandemic even with these age controls.

Caste and occupation. We explore whether excess deaths were higher among lower castes by adding indicators for 5 caste groups (scheduled castes, scheduled tribes, other backward castes, intermediate castes, and upper castes) and interaction between caste indicators and pandemic or wave indicators to the regression in Eq. (2). We find no significant differences in excess death rates by caste categories (Table G.13). We also explore whether excess death during the pandemic was a function of occupation by adding occupation indicators (student, home-based work, farm work, non-agricultural labor, office work, retired) and the interaction of occupation indicators with pandemic or wave indicators. We find home- and farm-based workers had higher excess mortality during the pandemic (Table G.14). However, to determine whether these findings are driven by age, income or caste, we estimated a regression that added to Eq. (2) simultaneously income, age, caste, and occupation indicators and the income, age, caste and occupation indicators each interacted with indicators for the pandemic. We find that heterogeneity in pandemic-period excess deaths across occupation is muted and insignificant (except for farm workers) when controlling for age or for age, caste and income. (However, age and income gradients persist even when controlling for caste and occupation.)

4. Discussion

Our preferred estimates imply that there were 4.2 million excess deaths in India during the pandemic, 2.2 million during wave 1 and 4.1 million during wave 2. The lower total is obtained because deaths were lower than baseline during wave 3, perhaps because earlier waves accelerated deaths that would otherwise occur in early 2022. These estimates only include consecutive responders, but are similar to estimates if we include households that skip up to one round of survey.

Our preferred estimate of death is roughly 8x as large as the official number of COVID deaths. This does not prove that COVID caused 8x more deaths, but it does suggest official numbers may be a substantial undercount. It is true that all-cause deaths include non-COVID deaths. But these are excess deaths during the pandemic, so it is likely that COVID directly or indirectly (via policy or behavioral change) is related to these deaths. Of course, we cannot demonstrate causation as we do not have a strictly exogenous introduction of COVID or variation in infections.

We benchmark our findings against estimates from the CRS, the official registry of deaths, in the literature. Our preferred estimate is somewhat higher than the estimates of excess deaths in papers that employ CRS data (Anand et al., 2021; Deshmukh et al., 2021). For example, Banaji and Gupta (2021) extrapolate excess deaths from the 12 of the 14 states for which CRS data are presently available. They

estimate excess deaths in all of India to lie between 2.8 and 5.2 million excess deaths from April 2020–June 2021 depending on how one addresses undercounting by the CRS. Our estimates for this same time period and using the same period of data are at 4.5 million in the middle of the range of CRS estimates.[16]

If our measure of excess deaths is assumed to be due to COVID, that disease easily becomes the leading cause of death in India. Prior to the pandemic, the leading causes of death were non-communicable diseases: cardiovascular disease (2.57 million death annually), chronic respiratory diseases (1.16 million annually) and neoplasms or cancers (0.93 million annually). The leading cause of death from communicable disease was respirator illness and tuberculosis (0.86 million annually).

There are three reasons to believe our estimates are in part picking up the direct effect of COVID. First, excess deaths follow a COVID-like age pattern. Second, excess deaths peak when India's two waves peak. Third, pandemic period deaths are correlated with the amount of infection in a district.

Our analysis certainly has limitations. First, we do not know the cause of death. Therefore, it is difficult to provide whether official COVID death counts are correct or not. We will attempt to address this in follow-on work that will conduct verbal autopsies on the deaths in CPHS households since 2019.

Second, CPHS data show a big jump in death rates in 2019. This might cast doubt on the validity of CPHS data or suggest that pre-trend that accounts for the mortality increase we observe. The fact that our estimates of excess pandemic-related deaths are consistent with those from other sources in India, suggests that our use of 2019 as a benchmark is valid for estimating that excess deaths. The fact that there was a change in the age pattern of deaths from 2019 to 2020, but not from 2018 to 2019, suggests that changes in 2020 are not a pre-trend.

Third, we have to impute the timing of death because deaths are sometimes reported months after they occur. We offer a solution to the problem that yields death rates that are at the higher end of CRS estimates after the latter are adjusted for undercounting. Moreover, alternatives such as the CRS have their own problems. CRS and SRS undercounted deaths, requiring estimates or assumptions about undercounting rates. Moreover, CRS is not available for all states and SRS will not be reported for several years.

Fourth, one might be concerned about non-random response by households. However, unless one includes households that did not respond for 12 months, our estimates of excess deaths do not change substantially even though our restricted sample comprise 95% of our sample.

Finally, we estimate – surprisingly – that death rates were lower than baseline during wave 3. We believe this is driven by displacement: individuals who would have died in early 2022 died in waves 1 and 2 in 2020 and 2021, respectively. The fact that the displacement drove deaths below baseline is not entirely surprising. Official data suggests that, even though the omicron variant generated a tremendous number of infections in wave 3, the ratio of confirmed deaths to reported cases was much lower in wave 3 than waves 1 and 2 (Fig. G.8A). The reason deaths rates were low is likely that a high portion of the population was immune, either through prior infection or vaccination, by wave 3.[17]

The extent and implications of displacement are difficult to estimate, but warrant further research. To estimate displacement, we need to know the distribution of conditional life expectancy amongst those who died in each month of the pandemic. The fact that displacement drove deaths below our crude baseline by wave 3 suggests that those with the lowest remaining lifespans already died in the pandemic. However, this logic implies that we cannot be sure if our wave 2 excess deaths estimates include displacement from wave 1. Nor can we be sure that those who died earlier in the pandemic do not include those with somewhat longer lifespans until we get more years of data: we may see their displacement in the next few years. A methodological implication of this logic is that the standard method of estimating excess deaths – the method we employ – yields estimates that change as one gets more periods of data. This complicates attribution of deaths to particular events such as waves or even the pandemic. One could wait a very long time to capture all displacement. But waiting also means the mortality rate will reflect not just displacement of COVID-related deaths but also changes in death rates due to factors other than COVID, be it medical innovation, development, a bad monsoon, or climate change. Nevertheless, understanding displacement is critical to estimating the harm from the pandemic. If COVID picked off the more vulnerable people at each age, then its impact may be smaller than a simple sum of ages of victims weighted by age-specific average longevity (see Table 3).

CRediT authorship contribution statement

Anup Malani: Conceptualization, Funding acquisition, Methodology, Project administration, Writing – original draft, Writing – review & editing. **Sabareesh Ramachandran:** Conceptualization, Data curation, Formal analysis, Methodology, Visualization, Writing – review & editing.

Data availability

We will share our code. We cannot furnish the data, but it is available to the public for a subscription fee by the producer of the data.

Acknowledgments

We thank Mahesh Vyas, Kaushik Krishnan, Chinmay Tumbe, Shamika Ravi, Rukmini S, Prabhat Jha, Arvind Subramanian, Justin Sandefur, Abhishek Anand, Anmol Somanchi and seminar participants at the CMIE weekly webinar for helpful comments. We thank Philip Mogan, Bartek Woda, and Satej Soman for help with COVID case and deaths data and the anonymous data scientists at www.covid19india.org and www.covid19bharat.org for scraping data on COVID cases and deaths.

Funding

Malani acknowledges funding from the Becker Friedman Institute at the University of Chicago to purchase a subscription to the Consumer Pyramids Household Survey and the support of the Barbara J. and B. Mark Fried Fund at the University of Chicago Law School.

Appendix A. Sampling method in CPHS

The country is divided into 99 of these regions. Rural areas are defined as 2011 Census villages. Urban areas are towns and cities. In the rural areas, villages are randomly selected. Within each village 16 households are selected by randomly picking a cluster of homes and then conducting systematic sampling. In the urban areas, towns are divided into substrata based on population. Within each size substrata, towns are randomly selected. Within towns, census enumeration blocks are randomly selected. In each CEB 16 households are selected.

[16] Our apples-to-apples estimates are smaller than our wave 1 plus wave 2 estimates (6.3 million) in Table 1 because our wave definitions include more months than, e.g., (Banaji and Gupta, 2021). Our apples-to-apples estimates smaller than our overall pandemic period definition because the latter includes wave 3 and the negative excess deaths estimates during that wave.

[17] The Indian Council for Medical Research estimated 90% seroprevalence in Delhi by September 2021 (Sharma et al., 2021).

Appendix B. Additional issues with CPHS mortality data

Recall that using survey data rather than death registration to measure mortality rates raises a number of data cleaning problems. First, survey response rates fell during the pandemic, particularly during India's lockdown. Second, there may be selection bias in non-response. Specifically, non-response may be a function of whether a household experienced a death. Third, there appears to be a level jump in the death rate in 2019, prior to the pandemic. Fourth, there CPHS has been criticized for not being representative of poor populations (Dreze and Somanchi, 2021). We address these here.

B.1. Low response rate during lockdown

The CPHS experienced a sharp decline in response rates during the lockdown in India. CPHS is ordinarily an in-person survey and the typical per-round, household response rate (responding/sample households) prior to the pandemic was roughly 85%. However, when India's central government declared a lockdown on March 24, 2020, in person surveys had to cease. CPHS made two changes: they switched to a phone survey and surveyors' managers, rather than surveyors, conducted the survey to keep up the quality of surveys. Because there are fewer managers than surveyors, CPHS decided only to call a quasi-random, representative subsample of households. The asked managers to pick household phone numbers with only information on strata (defined above) of households and required that the ratio of urban-to-rural households in each homogeneous remain the same as intended pre-pandemic. As a result, response rates fell. Fig. B.3 shows that the fraction of households that were not contacted rose to roughly 50% of the full sample from April–August 2020 and responding households constituted just 30% of the full sample at the height of the lockdown in April–May 2020, implying a response rate of roughly 60% in April 2020. When CPHS finished its second round in August 2020, it returned to in-person surveys. However, the response rate only rose to 75%. There was also a drop in response rates during wave 2, during which local lockdowns forced local use of telephonic surveys as before.

Low response rates are themselves not an issue because we only look at the proportion of *responding* people who are dead. Also, even if a household does not respond in one round, they may respond in subsequent rounds. In a robustness check we do not restrict to households that respond in consecutive rounds. All reported deaths are accounted for in this estimation.

B.2. Non-random response during COVID

It is possible that households that respond to the CPHS are not representative of the CPHS sample or that the nature of non-random response changed during COVID. The former affects our estimates of baseline death rates unless CPHS's weights make non-responding households representative even with non-random response.[18] Even if this the former is not true, the latter affects our estimate of the excess-death rate during COVID.

We have mixed evidence about the representativeness of the responding sample. The optimistic view comes from a simple exercise in the vein of Altonji et al. (2005). CPHS has hundreds of variables on each household, including current and lagged responses to income, time use and consumption questions. We estimated a regression of survey response on covariates (other than death) selected via LASSO prior in 2019 and then in 2020. Our estimated R^2 was < 0.01 (Table B.4). Of course, it could be that survey response is a function of unobservables

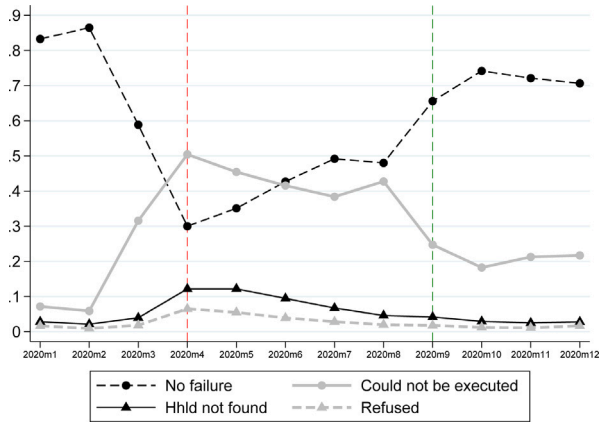


Fig. B.3. CPHS non-execution and non-response rates during 2020.
Notes: Red vertical dashed line indicates first month of phone surveys. Green dashed line indicates month that in-person surveys resumed. The sample includes all households. "Could not be executed" (non-execution) includes both CMIE's decision not to contact a household and it's inability to speak to a household member because, e.g., no one answered the door ("door-lock"). "Household not found" means CMIE attempted to contact the household but surveyors were unable to locate the household. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

even conditioned on observables. But given how many observables we have, this seems unlikely. If we make the assumption in, e.g., Altonji et al. (2005) that unobservables have the same explanatory power as observables, then our low bound on the R^2 from observables implied low R^2 for unobservables.

On the other hand, we do have some evidence that the fact of death affects response rates. This comes from the following exercise. First, we took the set of households that responded in round t and $t+2$, about 64% of the sample. (Since rounds are 4 months long, this means 8 months apart.) Some of these households also responded in $t+1$ (the $t+1$ "respondent" group) and some did not (the "non-respondent" group). (Respondents are 83% of the CPHS subsample that responds at t and $t+8$.) Second, we compare the number of deaths that occurred between t and $t+2$ in the respondent group and the non-respondent group. Recall that, even if a household does not respond at $t+1$, they eventually report their deaths in $t+2$, so we see all deaths in this period for both groups. Fig. B.4 plots the annualized death rate in the respondent and non-respondent groups by month of the $t+1$ survey. Respondent households have slightly more deaths prior to COVID, though in some months non-respondent death rates rise above response ones. But in 2020, the gap widens and the respondent group's death rates almost always appear to be above non-respondent group's rates.

Table B.5 provides estimates of the sort of bias one would get if one focused only on households responding in consecutive rounds as opposed to on households that at least responded in round t and t+2. The latter are about 64% of the sample. A regression of death rates on a pandemic indicator, a respondent household indicator and the interaction of the two indicators reveals that respondent death rates are 0.281 percentage points higher per annum before the pandemic, and rise 0.175 percentage points further above non-respondents during the pandemic, with each difference being statistically significant (Table B.5). Our finding that, of households that respond at t and $t+8$, households with a death are more likely to respond to a survey does not imply that is true for the full sample. Indeed, as we explain the in the next paragraph, the bias is different for the remainder of CPHS that does not respond at t and $t+8$. So the important take-away is that there could be non-random survey response, not that we can sign it.

There is a solution to non-random response, but it has separate problems. Non-random response can be addressed by using both respondent and non-respondent households when estimate excess deaths.

[18] The former may not affect our estimate of excess deaths during COVID if non-random response is such that the trend of deaths in responding households is similar to the trend in non-responding households, though this is an optimistic assumption.

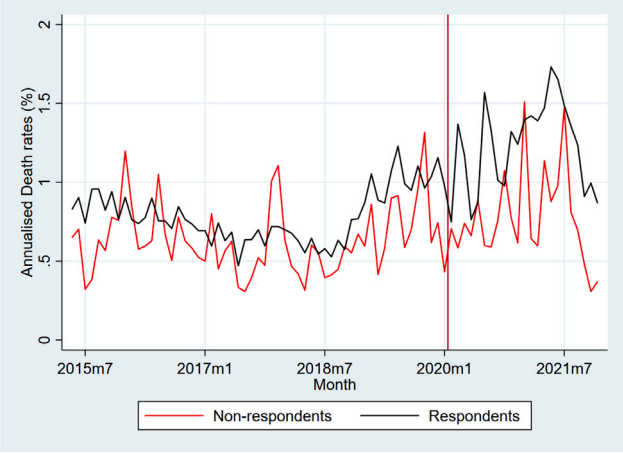


Fig. B.4. Annualized death rates for households that responded in round $t+1$ and those that did not.
Notes: The sample includes households that responded in round t and $t+2$. Some of these households also responded in $t+1$ (the $t+1$ "respondent" group) and some did not (the "non-respondent" group). This figure plots the annualized death rate in the respondent and non-respondent groups by month of the $t+1$ survey.

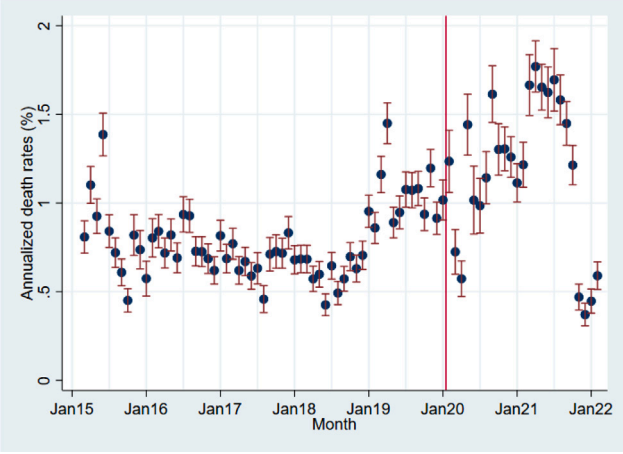


Fig. B.5. Annualized death rates by month from January 2015–June 2021.
Notes: The sample includes t+1 respondent and non-respondent households. Deaths reported in month t are allocated to month $t-2$, for reasons explained later. Each point is the weighted mean death rate in a month and each whisker is the 95% confidence interval on that mean. The red vertical line demarcates the start of the pandemic in February 2020. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table B.4
Fit (R^2) from LASSO-generated prediction model for CPHS non-execution and/or non-response in 2019 and 2020 using observables (other than death).

	March–June		Sept–Dec	
	2019	2020	2019	2020
Non-execution & non-response	.004	.008	.007	.007
Non-execution	.005	.009	.010	.008
Non-execution or response	.003	.008	.003	.003

Notes. The sample includes all households in the time period indicated in the column header. The table reports the fit (R^2) from a regression of an indicator for the action in the row label on (a) strata fixed effects (homogeneous region x community type) and (b) covariates selected by LASSO from the set of all variables (excluding death) on the household and its members for the same time period. Observations are at the household level. No weights are included.

After all, non-respondents at $t+1$ typically respond at $t+2$ and this fills in holes in the death data. However, there is loss of precision about the timing of deaths when $t+1$ non-respondents are included in the analysis: their deaths have to be allocated over 8 months rather than just 4 months. Table B.6 presents estimates of baseline death rate in 2019 and excess deaths during COVID as we vary the sample to include more or less non-consecutive respondents. The first column is our main sample of consecutive responders. The second and third allow into the sample households that skip at most 1 and at most 2 rounds of the survey, respectively. The last column includes all households. Adding non-responders increases our estimates of baseline mortality and excess deaths during COVID. The estimates rise rather than fall because none of the columns in Table B.6 have the same sample as that in Table B.5.

Because we believe that the timing problem is greater than the non-random response problem, we highlight results using consecutive observations in the main text, and report estimates using even non-consecutive observations here in Appendix.

Table B.5
Death rate by consecutive survey response status.

	Annualized death rates (%) (1)
Pandemic	0.0570
	(0.0570)
Respondent	0.281***
	(0.0490)
Pandemic × Respondent	0.175**
	(0.0638)
2019 Non-respondent mean	0.732***
	(0.0456)
N	3176467

Notes. This table reports the results from regressing an indicator for whether an individual died against an indicator for the duration of the pandemic, an indicator for the response status and the interaction of these two. The sample includes individuals who are observed in month t and month t+8, about 64% of the entire sample. The dependent variable for an individual in month t is whether their death status was reported by month t+8. Respondent is an indicator for whether the individual also responded in month t+4, whether or not they were alive that month. Observations on individuals are weighted to be nationally representative even with non-response. Standard errors clustered at the village/ward × month level are reported in parentheses. */**/*** indicates $p < 0.05/0.01/0.001$.

Table B.6
Excess-death rates for different samples of responders.

	Annualized death rates (%)			
	(1)	(2)	(3)	(4)
Pandemic	0.147***	0.181***	0.216***	0.264***
	(0.0400)	(0.0387)	(0.0383)	(0.0386)
2019 mean	1.038***	1.075***	1.083***	1.097***
	(0.0286)	(0.0282)	(0.0281)	(0.0282)
Sample composition	Responses 4 mo apart	Responses 4/8 mo apart	Responses 4/8/12 mo apart	All
% of full sample	82%	95%	98%	100%
N	3713943	4271836	4428134	4515543

Notes. This table reports the results of the main regression where we regress deaths against a pandemic indicator. All deaths are assigned to the month in the middle of the period when the individual was last surveyed and when they were reported to be dead. The pandemic indicator is set to 1 iff the middle month is after Jan 2020. In the first column we only include household responses in those months for which they responded again in the next round. In the second and third columns we include response if the household responded in at least one of the next two and three rounds respectively. In the fourth column, we include the entire sample. Standard errors clustered at the village/ward × month level are reported in parentheses. */**/*** indicates $p < 0.05/0.01/0.001$.

B.3. Rise in deaths in 2019

The annualized death rate calculated from the CPHS rises in 2019, a year before the pandemic (see Fig. B.5, which simply adds data from 2015–2018 to Fig. 1B). This is consistent with what is observed in the Global Burden of Disease data, though the magnitude is larger in the

CPHS. This raises two questions. One, is there a pre-trend that begins in 2019? Perhaps the death rate during COVID is actually caused by something else started in 2019. Two, what years are the appropriate benchmarks against which excess deaths during the pandemic should be calculated?

We do not think that the jump in 2019 is a pre-trend unrelated to COVID that continues into 2020 for several reasons. First, there is not a change in sample composition in 2019 leading to the rise in death rates. Of all households who respond in 2018, 98.5% households also respond in 2019. Also, of all households who respond in 2019, 99% households had also responded in 2018.

Second, as Fig. 2 will show, the age-wise death rate in 2019 is significantly higher than the death rate that for 2015–2018 for both the elderly (age 60+) and for youth (0–17). By contrast, the age-wise death rate during the pandemic is significantly higher than during 2019 for only the elderly (60+). The jump in 2019 is not consistent with the age profile of COVID deaths, while the jump in 2020 is.

We use both 2019 and 2015–2019 as a baseline for the purposes of calculating excess deaths during COVID. The argument for using 2019 is that the implied excess deaths is more in line with estimates based on CRS data after correcting for undercounting (e.g., Banaji and Gupta, 2021). The argument against using 2019 as a baseline is that it implies a baseline death rate of 1.038%, which is much higher than the death rate reported in the Global Burden of Disease. By contrast, the death rate implied by the 2015–2019 baseline (0.787%) is closer to the GBD baseline. Because the purpose of this paper is to estimate excess deaths and not the baseline death rate, we prefer employing the 2019 baseline, even though we report results from both baselines.

A natural question is whether the CPHS is to be believed given how high the baseline is in 2019. Our main answer is that, just because the baseline is high does not mean the change from 2019 to 2020–21 is incorrect. Indeed, our estimates of excess deaths from CPHS will be in line with the median estimate of excess deaths from the 10 states that have reported CRS data thus far.

B.4. Representativeness of CPHS

Dreze and Somanchi (2021) argues that CPHS under-samples the poor based on evidence that it yields both higher levels of literacy and faster improvement in literacy than government surveys. Dreze and Somanchi's criticism is a problem for us if the poor have a different death rate during COVID. The problem gets worse if the CPHS sample becomes less representative over time. It is possible that the poor have a lower death rate, as we show in Section 3.2.2; this would to lower our estimate of excess deaths from COVID if the overall number over-weights the rich. However, we do not believe that the problem gets worse over time as the gape between our control period (2019) and treatment period (2020–May 2021) is quite short. Moreover, experience from a serological survey conducted by one of us in Mumbai suggests that sampling the wealthy is more difficult than sampling the poor during lockdown (Malani et al., 2020).

Of course, the fact of difference between CPHS and government surveys of literacy is not dispositive of whether CPHS is biased since it is possible that the government surveys are the ones that are off. After all, the government has taken steps to suppress data (e.g., the 2017–18 consumption survey by the National Statistical Survey Office) that it finds unflattering. Moreover, government surveys are known to give different estimates of items like slums populations, with the differences driven by the policy aim of the survey.[19] Dreze and Somanchi suggest that CPHS is the one likely to be wrong because its frame samples more from the main streets of villages than from outskirts, where the poor tend to live. CMIE has responded that its sampling does get to outskirts and that the bias has not changed over time because that sampling frame is largely fixed and that it method for selection (of new households) has been constant (Vyas, 2021).

[19] For example, public health officials in Mumbai in private conversations have noted that surveys of slum populations by the public health department tend to generate higher estimates of slum population because higher numbers in slums are more likely to generate large appropriations for the department.

Table E.7
Robustness of excess-death rates to pandemic definition.

	Nov 2018	Dec 2018	Jan 2019	Feb 2019	Mar 2019
Dec 2019	0.17	0.14	0.12	0.11	0.08
	0.99	1.02	1.05	1.06	1.08
Jan 2020	0.19	0.16	0.14	0.13	0.11
	0.99	1.02	1.04	1.05	1.07
Feb 2020	0.19	0.17	0.15	0.14	0.12
	0.99	1.02	1.04	1.05	1.06
Mar 2020	0.19	0.16	0.14	0.13	0.11
	1.00	1.02	1.04	1.05	1.07
Apr 2020	0.21	0.18	0.16	0.16	0.14
	0.99	1.01	1.03	1.04	1.06

Notes. This table presents a matrix of estimates of the baseline and excess-death rate as we vary the definition of the baseline period (columns) and the pandemic period (rows). The columns refer to the start date of the baseline. The rows refer to the start date of the pandemic. Each cell contains an estimate of the annualized excess-death rate (top) and baseline death rate (bottom) based on (2).

Appendix C. Estimating infections

We estimate infections in three steps. First, we obtain data from a population-representative seroprevalence survey by the state of Tamil Nadu. The survey was conducted on 26,140 persons from 15 October–30 November 2020. The sample included individuals aged 18 and above who provided informed consent. Details on the study and its estimate of seroprevalence are available from (Malani et al., 2021). The survey was designed to be representative for urban and rural areas of districts and for demographic groups defined by age and gender.

Second, we extrapolate seroprevalence rates from this study to districts on 30 November 2020 based on the urban versus rural share of districts because urban share predicts more variability in seroprevalence than demographics and demographics do not vary much across districts.

Third, we take the curve that describes each district's new confirmed case over time (from www.covid19india.org) and scale it vertically up so that the total sum of scaled cases until 30 November equals the district's population times its estimated seroprevalence on that date. The resulting curve estimates the number of new individuals in a district infected with COVID on each day.

We had lots of choices for serological studies to use for our scaling. We could not use them all because they would yield inconsistent curves. Nor was there a clear way to combine them in a meta-analysis sense. We chose to use only the Tamil Nadu study for several reasons. First, it has a large sample size relative to other studies, e.g., the Karnataka study by Mohanan et al. (2021). It was one of a few state-wide or bigger serological surveys. Second, the study provided district-wise estimates unlike the national studies by the Indian Council for Medical Research (Sharma, 2021). While the Tamil Nadu study generates estimates of infection that are larger than the ICMR's first 3 rounds of serological surveys, it generates estimates that are in line with ICMR's fourth survey, which was conducted after wave 2. Third, we worked on the Tamil Nadu study, we knew the design and could vouch for its quality. We also had access to the data from that study so could better extrapolate from it to other districts.

Appendix D. Comparing different pandemic definitions

Table E.7 presents a matrix of estimates of the baseline and excess-death rate as we vary the definition of the baseline period (columns) and the pandemic period (rows).

Table E.8
Timing of CPHS observation of deaths and correction for that timing.

		Annualized death rate (deaths/1000) in each month											
	Formula	1	2	3	4	5	6	7	8	9	10	11	12
1. Truth (d)		7	7	7	7	9	7	7	7	7	7	7	7
2. Observed (y)	$y_t = (d_t + d_{t-1} + d_{t-2} + d_{t-3})/4$	7	7	7	7	7.5	7.5	7.5	7.5	7	7	7	7
3. Solution 1 (z)	$z_t = y_{t-2}$	7	7	7.5	7.5	7.5	7.5	7	7	7	7	7	7
4. Moving average of d	$m_t = (d_{t-1} + \ldots + y_{t+2})/4$	7	7	7.5	7.5	7.5	7.5	7	7	7	7	7	7
5. Solution 2	$r_t = (4y_t) - (r_{t-1} + r_{t-2} + r_{t-3})$	7	7	7	7	9	7	7	7	7	7	7	7

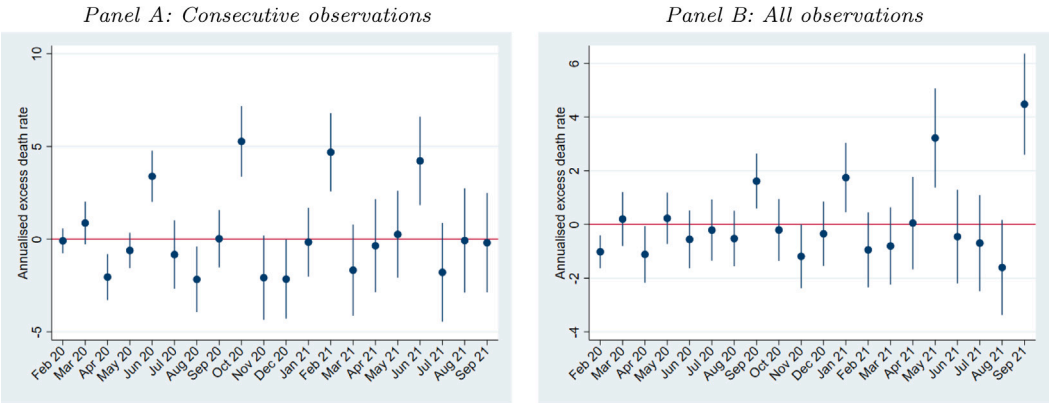


Fig. F.6. Monthly death rates from second solution.
Notes. This figure plots monthly death rates for each month. The coefficients from regression (E.1) are plotted here. The indicator for month t is 1 for an observation if the month t is between the month in which the individual was surveyed (including month of survey) and the month in which the individual is next surveyed in (excluding the month of survey). The regression is weighted using the individual's weights. Results in Panel A use only responses from households that answer consecutive surveys. Results in Panel B use responses from all households.

Table F.9
Excess death rates under the second timing solution for different pandemic definitions.

Panel A: Consecutive observations

Pandemic start month	Excess death rates (Annualized rate (%))	Standard error
Dec 2019	2.229	(0.572)
Jan 2020	2.145	(0.594)
Feb 2020	2.712	(0.578)
Mar 2020	2.871	(0.606)
Apr 2020	2.952	(0.639)
May 2020	3.291	(0.675)

Panel B: All observations

Pandemic start month	Excess death rates (Annualized rate (%))	Standard error
Dec 2019	1.003	(0.501)
Jan 2020	0.885	(0.521)
Feb 2020	1.154	(0.495)
Mar 2020	1.456	(0.519)
Apr 2020	1.662	(0.547)
May 2020	2.125	(0.577)

Notes. This table reports the excess-death rates computed using the specification in Eq. (E.1). The different rows represent different start dates for the pandemic period. The excess-death rates are the mean of the monthly death rates multiplied by 12. The table on the left only uses observations from households that answer consecutive surveys. The table on the right uses all observations.

Appendix E. Second solution to the timing-of-death problem

The timing-of-death problem is illustrated with the example in Table E.8. Suppose the true death rates over a 10-month period are 7 per 1000 for each month except month 4, where it jumps to 9 (row 1). Moreover, assume one quarter of households are surveyed each month and all households respond to surveys, which are 4 months apart. Because only a quarter of households are interviewed each month, those extra 2 deaths will, statistically, be distributed over 4 months after the death (row 2). The problem we face is how to back out the jump to 9 in row 1.

Our preferred solution from the main text, which is illustrated in row 3 of Table E.8, is akin to treating the reported number at t as a moving average of the true death rate for $t - (k/2)$ (row 4).

Here we explore a second solution: to estimate the death rate by asking the question, how much would the true death rate have to have changed for the observed death rate to have changed as much as it did since the last month. The formula that provides the answer is in row 5.

The advantage of the second solution is that it can, in some cases, back out the true death rate. But there are two problems. First, death rates need to be stable for a period otherwise one cannot solve the formula because it uses prior value of estimated rates to measure current rates. Second, if observed deaths have some error unrelated to timing, this solution magnifies those errors. The solution recognizes that the actual changes have to be larger than observed changes because the survey process smooths out changes over a few months (row 2). But if there are errors in observed data, then the errors are also magnified. This can increase variability of results from solution 2, which we will demonstrate. Because we think there could be errors in CPHS rosters, our preferred measure is the first one.

We implement our second solution to the timing-of-death problem with a regression of the form

$$d_{i,t,t-k} = \sum_{k=4,8,\ldots} \delta_k \cdot \mathbb{I}(k) + \sum_{m \in \text{pandemic}} \beta_m \cdot \mathbb{I}(t-k \le m \le t) + \epsilon_{ist} \qquad \text{(E.1)}$$

where $\mathbb{I}(k)$ is an indicator for whether the gap between the current survey and the one to which she last responded is k months, and $\mathbb{I}(t-k \le m \le t)$ is an indicator whether the intervening period between surveys was during the pandemic. Standard errors are clustered at the village/ward × month level to account for correlation in reporting of deaths within a locality.

Table F.10
Excess-death rate during the pandemic and its waves in different age groups.

Panel A: Pandemic overall							
	(1) 0-17	(2) 18-29	(3) 30-39	(4) 40-49	(5) 50-59	(6) 60-69	(7) 70+
During Pandemic	−0.0932	−0.0730	−0.0391	0.0378	0.0901	1.318***	3.551***
	(0.0501)	(0.0375)	(0.0472)	(0.0448)	(0.0980)	(0.287)	(0.768)
2019 mean	0.355***	0.318***	0.319***	0.405***	1.229***	4.175***	12.98***
	(0.0420)	(0.0323)	(0.0388)	(0.0334)	(0.0742)	(0.189)	(0.488)
N	774895	779978	462487	626187	480082	226455	101663

Panel B: By waves							
	(1) 0-17	(2) 18-29	(3) 30-39	(4) 40-49	(5) 50-59	(6) 60-69	(7) 70+
Wave 1	−0.0496	−0.0248	−0.0270	−0.0000725	0.189	1.077**	2.831***
	(0.0635)	(0.0449)	(0.0536)	(0.0539)	(0.129)	(0.371)	(0.752)
Wave 2	−0.0888	−0.0877*	−0.0269	0.175**	0.215	2.842***	8.814***
	(0.0551)	(0.0420)	(0.0618)	(0.0642)	(0.123)	(0.422)	(0.837)
Wave 3	−0.268***	−0.191***	−0.116	−0.214***	−0.545***	−2.093***	−7.954***
	(0.0524)	(0.0536)	(0.0713)	(0.0571)	(0.128)	(0.358)	(0.714)
2019 mean	0.355***	0.318***	0.319***	0.405***	1.229***	4.175***	12.98***
	(0.0420)	(0.0323)	(0.0388)	(0.0334)	(0.0742)	(0.189)	(0.393)
N	774895	779978	462487	626187	480082	226455	101663

Notes. The sample includes households that respond in consecutive rounds. Observations on individuals are weighted to be nationally representative even with non-response. The regression model in Panel A is (2); the model in Panel B replaces the pandemic indicator in (2) with wave indicators. Each column reports estimates from a different regression. Regressions for each category are run by restricting the sample to those in that age category alone. Standard errors clustered at the village/ward × month level are reported in parentheses. */**/*** indicates $p < 0.05/0.01/0.001$.

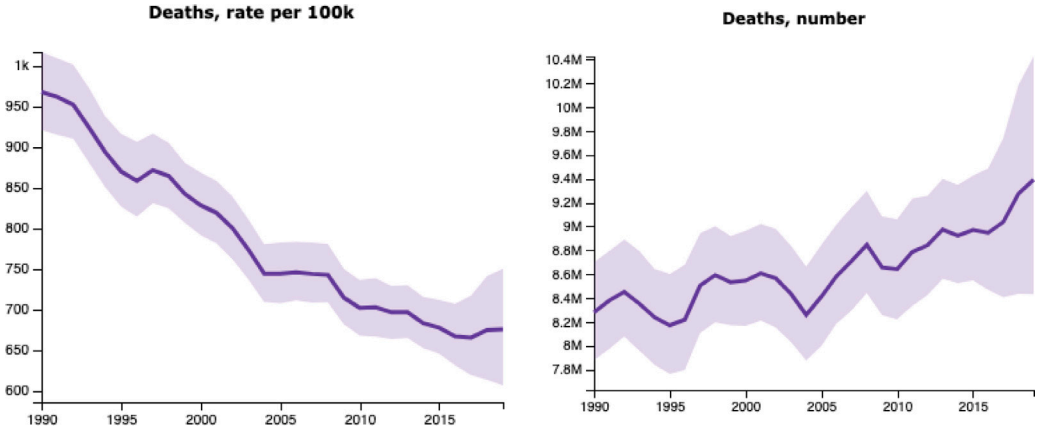


Fig. G.7. Death rate and total deaths in India over time.
Source: ghdx.healthdata.org/gbd-results-tool.

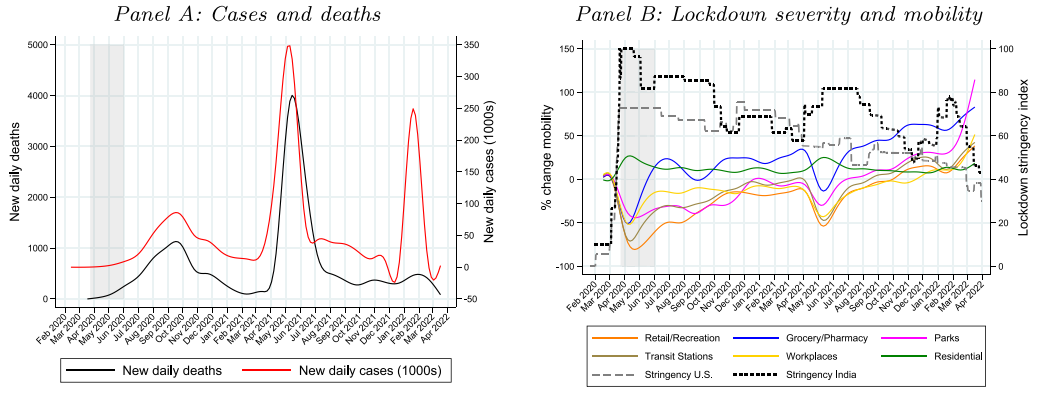


Fig. G.8. COVID Trajectory, Severity of Lockdown, and Mobility Changes.
Note. Case and death data are from Our World in Data (Ritchie et al., 2020). We show aggregated daily reported cases and deaths from the government. Shaded period marks the national lockdown. Lockdown severity data are from Oxford (Hale et al., 2020). Mobility data are from Google mobility reports (Google, 2021). Shaded period marks the national lockdown. Time periods cover February 2020–January 2021. We thank Philip Monagan for making this figure.

Table G.11
COVID death rate prior to the pandemic and the excess-death rate during the pandemic, with location fixed effects.

	Annualized death rates (%)		
	(1)	(2)	(3)
Pandemic	0.147***	0.136***	0.149***
	(0.0400)	(0.0399)	(0.0389)
Constant	1.038***	2.020***	1.966***
	(0.0286)	(0.300)	(0.574)
Controls	None	HR FE	District FE
N	3713943	3713943	3713943

Notes. The sample includes households that respond in consecutive rounds. Observations on individuals are weighted to be nationally representative even with non-response. The regression model is (2). HR FE means fixed effects for homogeneous region, a cluster of similar districts within a state; district FE means district fixed effects. Standard errors clustered at the village/ward × month level are reported in parentheses. In the column 2, if we instead cluster standard errors at the homogeneous region × month level, the standard error is 0.073 and the effect remains significant. */**/*** indicates $p < 0.05/0.01/0.001$.

The coefficient δ_k estimates a pre-COVID death rate for observations that are k months apart and our parameter of interest β_m captures the increment in true death rate implied by the increment in observed death rate in each month during COVID. The coefficient γ_i nets out seasonality in deaths.

Our second solution to the timing-of-death problem produces implausibly variable and/or high estimates of excess deaths from the pandemic.

Excess variability is illustrated in Fig. F.6A, which plots monthly excess-death rates (relative to 2019 rates) through September 2021 based on (a) estimating (E.1), but with month fixed effects instead of a pandemic period indicator, and (b) using, importantly, observations only from households that answer consecutive surveys. This solution produces a strong saw-tooth pattern with spikes every 4 months in the excess-death rate. The reason for this pattern is that the second solution is premised on the idea that a jump in month t is reallocated evenly across months t to t+3. The solution corrects that by reallocating deaths from months 2–4 back to month 1. But a jump on observed error may reflect a true increase in death rates, or an error in the roster. If the jump were a true jump in death rates, that correction would be correct. But if it is a positive error in month t, then months t+1 to t+3 are inappropriately suppressed. The suppression ends disappears in month t+4 so month t+4 is higher than t+3. If there happens to be another positive shock in t+4, the pattern repeats. We believe that is what was happening in Fig. F.6A.

Including observations from households that may not answer consecutive surveys mitigates the saw-tooth pattern (Fig. F.6B). Consecutive surveys are 4 months apart and, so, solution 2 reallocates positive jumps in deaths only over 4 months. When that is relaxed, some of the jump is allocated over 4 months, some over 8 months, and so on. This dampens the 4-month cycles our second solution generates from responders to consecutive surveys.

Estimates of excess deaths during the pandemic are higher when we use our second solution than in our first solution. Table F.9 reports our estimate of excess deaths through September 2021 using solution 2 for various definitions of the pandemic period. If we only use observations from households that answer consecutive observations, our estimate of the annualized pandemic period excess-death rate is 1.763 and significant if we use our preferred pandemic start date (February 2020). It falls to 0.497 and insignificant if we use all observations. The latter is partly due to the fact that deaths in households that do not answer the last round of 2020 or the first round of 2021 are included, but we have not captured deaths in those households yet; they may report these deaths in future rounds. As with our first solution, the excess-death rate under the second solution falls as we move up the start date and rises as we move it back.

Table G.12
Annualized death rates by gender and urban status.

Panel A: Gender

	Annualized death rates (%)	
	(1)	(2)
Pandemic	0.156***	
	(0.0471)	
Wave 1		0.120*
		(0.0584)
Wave 2		0.440***
		(0.0666)
Wave 3		−0.506***
		(0.0617)
Pandemic × Female	−0.0194	
	(0.0545)	
Wave 1 × Female		0.0581
		(0.0706)
Wave 2 × Female		−0.0757
		(0.0775)
Wave 3 × Female		−0.129
		(0.0663)
Female	−0.000111	−0.000111
	(0.000126)	(0.000126)
Constant	1.054***	1.054***
	(0.0340)	(0.0340)
N	3713943	3713943

Panel B: Urban/rural

	Annualized death rates (%)	
	(1)	(2)
Pandemic	0.117*	
	(0.0491)	
Wave 1		0.139*
		(0.0644)
Wave 2		0.321***
		(0.0632)
Wave 3		−0.528***
		(0.0646)
Pandemic × Urban	0.0919	
	(0.0846)	
Wave 1 × Urban		0.0270
		(0.105)
Wave 2 × Urban		0.249*
		(0.124)
Wave 3 × Urban		−0.107
		(0.0922)
Urban	−0.000105	−0.000105
	(0.000194)	(0.000194)
Constant	1.048***	1.048***
	(0.0362)	(0.0362)
N	3713943	3713943

Notes. Estimates are from a regression model based on (2), with the addition of a gender (urban) indicator and gender (urban) indicator interacted with the pandemic or wave indicator. The sample includes only those responding in consecutive rounds. Standard errors clustered at the village/ward × month level are reported in parentheses. */**/*** indicates $p < 0.05/0.01/0.001$.

Appendix F. Age-wise deaths

Table F.10 reports estimates of a regression based on (2), but on samples in different age bins. It shows that the excess-death rate during the pandemic is larger and significant in older ages. Moreover, this age skew was more pronounced in the second wave.

Table G.13
COVID death rate by nature of caste categories.

	Annualized death rates (%)		
	(1)	(2)	(3)
ST × Pandemic	0	0	0
	(.)	(.)	(.)
SC × Pandemic	−0.163	−0.193	−0.133
	(0.144)	(0.144)	(0.155)
OBC × Pandemic	−0.164	−0.223	−0.150
	(0.136)	(0.136)	(0.147)
Intermediate Caste × Pandemic	−0.0144	−0.104	0.0188
	(0.175)	(0.176)	(0.198)
Upper Caste × Pandemic	−0.0692	−0.156	−0.0837
	(0.152)	(0.152)	(0.164)
ST	0	0	0
	(.)	(.)	(.)
SC	0.0454	−0.0160	0.00400
	(0.101)	(0.101)	(0.101)
OBC	0.0565	−0.124	−0.104
	(0.0930)	(0.0927)	(0.0935)
Intermediate Caste	0.144	−0.260*	−0.217
	(0.112)	(0.113)	(0.113)
Upper Caste	0.114	−0.230*	−0.177
	(0.109)	(0.110)	(0.112)
N	3671763	3671763	3150097
Age control	No	Yes	Yes
Income and work place controls	No	No	Yes

Notes. Estimates are from a regression model based on (2), with the addition of a caste category indicator and caste category indicator interacted with the pandemic or wave indicator. The sample includes only those responding in consecutive rounds. CPHS records whether a member belongs to one of the 5 caste categories listed in the table. The category ST is the excluded category and the coefficients against other categories is the difference relative to ST. Standard errors clustered at the village/ward × month level are reported in parentheses. */**/*** indicates $p < 0.05/0.01/0.001$.

Appendix G. Appendix exhibits

G.1. Figures

See Figs. G.7 and G.8 and Tables G.11–G.14.

Table G.14
COVID death rate by occupation category.

	Annualized death rates (%)		
	(1)	(2)	(3)
Student × Pandemic	0	0	0
	(.)	(.)	(.)
Home-based work/unemployed × Pandemic	0.214***	0.0574	0.0711
	(0.0593)	(0.0666)	(0.0679)
Farm worker × Pandemic	0.373***	0.231*	0.250*
	(0.0875)	(0.0975)	(0.0992)
Non-agriculture labourer × Pandemic	−0.0641	−0.144	−0.134
	(0.103)	(0.109)	(0.112)
Office work × Pandemic	0.158**	0.0758	0.0279
	(0.0608)	(0.0733)	(0.0751)
Retired/aged × Pandemic	2.031***	0.431	0.511
	(0.435)	(0.424)	(0.420)
Student	0	0	0
	(.)	(.)	(.)
Home-based work/unemployed	0.493***	0.0176	−0.0144
	(0.0424)	(0.0521)	(0.0530)
Farm worker	0.455***	−0.0228	−0.0850
	(0.0623)	(0.0713)	(0.0731)
Non-agriculture labourer	0.505***	0.275**	0.210*
	(0.0801)	(0.0871)	(0.0889)
Office work	0.160***	−0.0691	−0.00597
	(0.0421)	(0.0549)	(0.0560)
Retired/aged	8.159***	2.821***	2.866***
	(0.284)	(0.280)	(0.275)
N	3713943	3713943	3660340
Age control	No	Yes	Yes
Income and caste controls	No	No	Yes

Notes. Estimates are from a regression model based on (2), with the addition of a occupation category indicator and occupation category indicator interacted with the pandemic or wave indicator. The sample includes only those responding in consecutive rounds. CPHS records nature of occupation for each individual. We further classify these occupations into six categories. (Student includes Student and Non-schooling Child. Home-based work/unemployed includes Home Maker, Home-based Worker, and Unoccupied. Farm worker includes Agricultural Labourer, Organized Farmer, and Small Farmer. Non-agricultural laborer includes Small Trader/Hawker/ Businessman without Fixed Premises, and Wage Labourer. Office work includes Businessman, Industrial Workers, Legislator/Social Worker/ Activists, Manager, Non-Industrial Technical Employee, Qualified Self Employed Professionals, Self Employed Entrepreneur, Self employed professional, Support Staff, White Collar Clerical Employees, White collar worker, and White-Collar Professional Employees and Other Employees. Retired/Aged is a category by itself.) The category student is the excluded category in the regression and the coefficients against other categories is the difference relative to student. Standard errors clustered at the village/ward × month level are reported in parentheses. */**/*** indicates $p < 0.05/0.01/0.001$.

References

Altonji, J.G., Elder, T.E., Taber, C.R., 2005. Selection on observed and unobserved variables: Assessing the effectiveness of Catholic schools. J. Polit. Econ. 113 (1), 151–184.

Anand, A., Sandefur, J., Subramanian, A., 2021. Three new estimates of India's all-cause excess mortality during the COVID-19 pandemic. Center Glob. Dev..

Andrews, M.A., Areekal, B., Rajesh, K.R., Krishnan, J., Suryakala, R., Krishnan, B., Muraly, C.P., Santhosh, P.V., 2020. First confirmed case of COVID-19 infection in India: A case report. Indian J. Med. Res. 151 (5), 490–492. http://dx.doi.org/10.4103/ijmr.IJMR_2131_20.

Banaji, M., Gupta, A., 2021. Estimates of pandemic excess mortality in India based on civil registration data. Cold Spring Harbor Laboratory Press, http://dx.doi.org/10.1101/2021.09.30.21264376, MedRxiv URL https://www.medrxiv.org/content/early/2021/10/01/2021.09.30.21264376.

Bedi, A., 2022. Delhi violating ICMR, WHO norms by not covid-testing the dead. But it's not the only state. The Print (May 20, 2020), URL https://theprint.in/health/delhi-violating-icmr-who-norms-by-not-covid-testing-the-dead-but-its-not-the-only-state/425549/.

Deshmukh, Y., Suraweera, W., Tumbe, C., Bhowmick, A., Sharma, S., Novosad, P., Fu, S.H., Newcombe, L., Gelband, H., Brown, P., Jha, P., 2021. Excess mortality in India from june 2020 to june 2021 during the covid pandemic: death registration, health facility deaths, and survey data. http://dx.doi.org/10.1101/2021.07.20.21260872, 2021.07.20.21260872, MedRxiv URL http://medrxiv.org/content/early/2021/07/23/2021.07.20.21260872.abstract.

Dreze, J., Somanchi, A., 2021. View: New barometer of India's economy fails to reflect deprivations of poor households. Econ. Times June 21 (June 21), bit.ly/35HtxH8 URL https://economictimes.indiatimes.com/opinion/et-commentary/view-the-new-barometer-of-indias-economy-fails-to-reflect-the-deprivations-of-poor-households/articleshow/83696115.cms.

Gamio, L., Glanz, J., 2021. Just how big could India's true covid toll be? N.Y. Times May 25, 2021.

Gerland, P., 2014. UN population division's methodology in preparing base population for projections: Case study for India. Asian Populat. Stud. 10 (3), 274–303. http://dx.doi.org/10.1080/17441730.2014.947059.

Gettleman, J., Yasir, S., Kumar, H., Raj, S., 2021. As Covid-19 devastates India, deaths go undercounted. In: New York Times. May 31, 2021 URL https://www.nytimes.com/2021/04/24/world/asia/india-coronavirus-deaths.html.

Google, 2021. COVID-19 Community Mobility Report - India. Report, Google, URL https://www.gstatic.com/covid19/mobility/2020-05-29_IN_Mobility_Report_en-GB.pdf.

Green, M.S., Nitzan, D., Schwartz, N., Niv, Y., Peer, V., 2021. Sex differences in the case-fatality rates for COVID-19—A comparison of the age-related differences and consistency over seven countries. PLOS ONE 16 (4), e0250523. http://dx.doi.org/10.1371/journal.pone.0250523.

Hale, T., Webster, S., Petherick, A., Phillips, T., Kira, B., 2020. Variation in government responses to COVID-19. Report, Blavatnik school of government working paper.

(continued on next page)

Johns Hopkins University and Medicine, 2022. COVID-19 Dashboard by the Center for Systems Science and Engineering at Johns Hopkins University. Report, Johns Hopkins University, URL https://coronavirus.jhu.edu/map.html.

Karlinsky, A., 2021. International completeness of death registration 2015–2019. http://dx.doi.org/10.1101/2021.08.12.21261978, 2021.08.12.21261978, MedRxiv URL http://medrxiv.org/content/early/2021/08/14/2021.08.12.21261978.abstract.

Khandekar, O., 2021. How India's slums are faring against the second covid19 wave. Livemint April 29 (April 29), URL https://lifestyle.livemint.com/news/big-story/how-india-s-slums-are-faring-against-the-second-covid19-wave-111619620504375.html.

Kontopantelis, E., Mamas, M.A., Deanfield, J., Asaria, M., Doran, T., 2021. Excess mortality in England and Wales during the first wave of the COVID-19 pandemic. J. Epidemiol. Commun. Health 75 (3), 213. http://dx.doi.org/10.1136/jech-2020-214764, URL http://jech.bmj.com/content/75/3/213.abstract.

Malani, A., Ramachandran, S., Tandel, V., Parasa, R., Sudharshini, S., Prakash, V., Yoganathan, Y., Raju, S., Selvavinayagam, T., 2021. Seroprevalence in Tamil Nadu in october-november 2020. MedRXiv.

Malani, A., Shah, D., Kang, G., Lobo, G.N., Shastri, J., Mohanan, M., Jain, R., Agrawal, S., Juneja, S., Imad, S., Kolthur-Seetharam, U., 2020. Seroprevalence of SARS-CoV-2 in slums versus non-slums in mumbai, India. Lancet Glob. Health http://dx.doi.org/10.1016/S2214-109X(20)30467-8, URL http://www.sciencedirect.com/science/article/pii/S2214109X20304678.

Mohanan, M., Malani, A., Krishnan, K., Acharya, A., 2020. Prevalence of COVID-19 in rural Versus Urban Areas in a low-income country: Findings from a statewide study in karnataka, India. http://dx.doi.org/10.1101/2020.11.02.20224782, 2020.11.02.20224782, MedRxiv URL https://www.medrxiv.org/content/medrxiv/early/2020/11/04/2020.11.02.20224782.full.pdf.

Mohanan, M., Malani, A., Krishnan, K., Acharya, A., 2021. Prevalence of SARS-CoV-2 in karnataka, India. JAMA 325 (10), 1001–1003. http://dx.doi.org/10.1001/jama.2021.0332.

Nguyen, N.T., Chinn, J., De Ferrante, M., Kirby, K.A., Hohmann, S.F., Amin, A., 2021. Male gender is a predictor of higher mortality in hospitalized adults with COVID-19. PLOS ONE 16 (7), e0254066. http://dx.doi.org/10.1371/journal.pone.0254066.

Pastor-Barriuso, R., Pérez-Gómez, B., Hernán, M.A., Pérez-Olmeda, M., Yotti, R., Oteo-Iglesias, J., Sanmartín, J.L., León-Gómez, I., Fernández-García, A., Fernández-Navarro, P., Cruz, I., Martín, M., Delgado-Sanz, C., Fernández de Larrea, N., León Paniagua, J., Muñoz-Montalvo, J.F., Blanco, F., Larrauri, A., Pollán, M., 2020. Infection fatality risk for SARS-CoV-2 in community dwelling population of Spain: nationwide seroepidemiological study. BMJ 371, m4509. http://dx.doi.org/10.1136/bmj.m4509, URL https://www.bmj.com/content/bmj/371/bmj.m4509.full.pdf.

Prasad, R., 2021. Coronavirus | do excess deaths suggest mortality crossed one million? The Hindu (June 20, 2021), URL https://www.thehindu.com/sci-tech/science/coronavirus-do-excess-deaths-suggest-mortality-crossed-one-million/article34860615.ece.

Rader, B., Scarpino, S.V., Nande, A., Hill, A.L., Adlam, B., Reiner, R.C., Pigott, D.M., Gutierrez, B., Zarebski, A.E., Shrestha, M., Brownstein, J.S., Castro, M.C., Dye, C., Tian, H., Pybus, O.G., Kraemer, M.U.G., 2020. Crowding and the shape of COVID-19 epidemics. Nat. Med. 26 (12), 1829–1834. http://dx.doi.org/10.1038/s41591-020-1104-0.

Rao, C., Gupta, M., 2020. The civil registration system is a potentially viable data source for reliable subnational mortality measurement in India. BMJ Glob. Health 5 (8), e002586. http://dx.doi.org/10.1136/bmjgh-2020-002586, https://pubmed.ncbi.nlm.nih.gov/32792407 https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7430426/.

Ravi, S., 2021. Counting deaths in India is difficult. Hindustan Times July 14, 2021 (July 14, 2021), URL https://www.hindustantimes.com/opinion/counting-deaths-in-india-is-difficult-101626273326958.html.

Ritchie, H., Mathieu, E., Rodés-Guirao, L., Appel, C., Giattino, C., Ortiz-Ospina, E., Hasell, J., Macdonald, B., Beltekian, D., Roser, M., 2020. Coronavirus pandemic (COVID-19). In: Our World in Data. https://ourworldindata.org/coronavirus.

Ritchie, H., Roser, M., 2019. Age structure. In: Our World in Data. https://ourworldindata.org/age-structure.

Rossen, L.M., Branum, A.M., Ahmad, F.B., Sutton, P., Anderson, R.N., 2020. Excess Deaths Associated with COVID-19, by Age and Race and Ethnicity - United States, January 26-October 3, 2020, Vol. 69. MMWR. Morbidity and Mortality Weekly Report, (42), pp. 1522–1527. http://dx.doi.org/10.15585/mmwr.mm6942e2, https://pubmed.ncbi.nlm.nih.gov/33090978 https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7583499/.

Rukmini, S., 2021. Gauging pandemic mortality with civil registration data. The Hindu URL https://www.thehindu.com/opinion/op-ed/gauging-pandemic-mortality-with-civil-registration-data/article35157185.ece.

Sharma, H., 2021. Two-thirds of Indians have covid antibodies, 40 crore still at risk: ICMR. Indian Express July 21, 2021 (July 21, 2021).

Sharma, P., Basu, S., Mishra, S., Gupta, E., Aggarwal, R., Kale, P., Mundeja, N., Charan, B.S., Singh, G.K., Singh, M.M., 2021. SARS-CoV-2 seroprevalence in Delhi, India - september-october 2021 – a population based seroepidemiological study. http://dx.doi.org/10.1101/2021.12.28.21268451, 2021.12.28.21268451, MedRxiv URL http://medrxiv.org/content/early/2021/12/29/2021.12.28.21268451.abstract.

Sinha, A., 2022. India numbers up, but UP death registrations fell in pandemic year. The Indian Express (May 11, 2022), URL https://indianexpress.com/article/india/india-numbers-up-but-up-death-registrations-fell-in-pandemic-year-7910483/.

Stier, A.J., Berman, M.G., Bettencourt, L.M.A., 2020. COVID-19 attack rate increases with city size. http://dx.doi.org/10.1101/2020.03.22.20041004, 2020.03.22.20041004, MedRxiv URL http://medrxiv.org/content/early/2020/04/03/2020.03.22.20041004.abstract.

United Nations, S.D., 2021. Coverage of birth and death registration. November 15, 2021 URL https://unstats.un.org/unsd/demographic-social/crvs/.

Vos, T., Lim, S.S., Abbafati, C., Abbas, K.M., Abbasi, M., Abbasifard, M., Abbasi-Kangevari, M., Abbastabar, H., Abd-Allah, F., Abdelalim, A., Abdollahi, M., Abdollahpour, I., Abolhassani, H., Aboyans, V., Abrams, E.M., Abreu, L.G., Abrigo, M.R.M., Abu-Raddad, L.J., Abushouk, A.I., Acebedo, A., Ackerman, I.N., Adabi, M., Adamu, A.A., Adebayo, O.M., Adekanmbi, V., Adelson, J.D., Adetokunboh, O.O., Adham, D., Afshari, M., Afshin, A., Agardh, E.E., Agarwal, G., Agesa, K.M., Aghaali, M., Aghamir, S.M.K., Agrawal, A., Ahmad, T., Ahmadi, A., Ahmadi, M., Ahmadieh, H., Ahmadpour, E., Akalu, T.Y., Akinyemi, R.O., Akinyemiju, T., Akombi, B., Al-Aly, Z., Alam, K., Alam, N., Alam, S., Alam, T., Alanzi, T.M., Albertson, S.B., Alcalde-Rabanal, J.E., Alema, N.M., Ali, M., Ali, S., Alicandro, G., Alijanzadeh, M., Alinia, C., Alipour, V., Aljunid, S.M., Alla, F., Allebeck, P., Almasi-Hashiani, A., Alonso, J., Al-Raddadi, R.M., Altirkawi, K.A., Alvis-Guzman, N., Alvis-Zakzuk, N.J., Amini, S., Amini-Rarani, M., Aminorroaya, A., Amiri, F., Amit, A.M.L., Amugsi, D.A., Amul, G.G.H., Anderlini, D., Andrei, C.L., Andrei, T., Anjomshoa, M., Ansari, F., Ansari, I., Ansari-Moghaddam, A., Antonio, C.A.T., Antony, C.M., Antriyandarti, E., Anvari, D., Anwer, R., Arabloo, J., Arab-Zozani, M., Aravkin, A.Y., Ariani, F., Ärnlöv, J., Aryal, K.K., Arzani, A., Asadi-Aliabadi, M., Asadi-Pooya, A.A., Asghari, B., Ashbaugh, C., Atnafu, D.D., et al., 2020. Global burden of 369 diseases and injuries in 204 countries and territories, 1990–2019: a systematic analysis for the global burden of disease study 2019. Lancet 396 (10258), 1204–1222. http://dx.doi.org/10.1016/S0140-6736(20)30925-9.

Vyas, M., 2021. View: There are practical limitations in CMIE's CPHS sampling, but no bias. Econ. Times June 23 (June 23), URL https://economictimes.indiatimes.com/opinion/et-commentary/view-there-are-practical-limitations-in-cmies-cphs-sampling-but-no-bias/articleshow/83788605.cms.

Waghmare, R., Gajbhiye, R., Mahajan, N.N., Modi, D., Mukherjee, S., Mahale, S.D., 2021. Universal screening identifies asymptomatic carriers of SARS-CoV-2 among pregnant women in India. Eur. J. Obstet. Gynecol. Reprod. Biol. 256, 503–505. http://dx.doi.org/10.1016/j.ejogrb.2020.09.030, URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7510530/.

Woolf, S.H., Chapman, D.A., Sabo, R.T., Zimmerman, E.B., 2021. Excess deaths from COVID-19 and other causes in the US, march 1, 2020, to january 2, 2021. JAMA 325 (17), 1786–1789. http://dx.doi.org/10.1001/jama.2021.5199.

www.covid19bharat.org, 2022. COVID19bharat.

www.covid19india.org, 2021. Coronavirus outbreak in India.