NBER WORKING PAPER SERIES

USING HOUSEHOLD ROSTERS FROM SURVEY DATA TO ESTIMATE ALL-CAUSE MORTALITY DURING COVID IN INDIA

Anup Malani
Sabareesh Ramachandran

Working Paper 29192
http://www.nber.org/papers/w29192

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
August 2021

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. We thank Mahesh Vyas, Kaushik Krishnan, Chinmay Tumbe, Shamika Ravi, Rukmini S, Prabhat Jha, Arvind Subramanian, Justin Sandefur, Abhshek Anand, Anmol Somanchi and seminar participants at the CMIE weekly webinar for helpful comments. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Using Household Rosters from Survey Data to Estimate All-cause Mortality during COVID in India

Anup Malani and Sabareesh Ramachandran

NBER Working Paper No. 29192
August 2021
JEL No. C80,I10,I14

ABSTRACT

Official statistics on deaths in India during the COVID pandemic are either incomplete or are reported with a delay. To overcome this shortcoming, we estimate excess deaths in India using the household roster from a large panel data set, the Consumer Pyramids Household Survey, which reports attrition from death. We address the problem that the exact timing of death is not reported in two ways, via a moving average and differencing monthly deaths. We estimate roughly 4.5 million (95% CI: 2.8M to 6.2M) excess deaths over 16 months during the pandemic in India. While we cannot demonstrate causality between COVID and excess deaths, the pattern of excess deaths is consistent with COVID-associated mortality. Excess deaths peaked roughly during the two COVID waves in India; the age structure of excess deaths is right skewed relative to baseline, consistent with COVID infection fatality rates; and excess deaths are positively correlated with reported infections. Finally, we find that the incidence of excess deaths was disproportionately among the highest tercile of income-earners and was negatively associated with district-level mobility.

Anup Malani
University of Chicago Law School
1111 E. 60th Street
Chicago, IL 60637
and NBER
amalani@uchicago.edu

Sabareesh Ramachandran
Department of Economics
UC San Diego
saramach@ucsd.edu

COVID is the largest pandemic since the 1918 flu. According to official reports, over 195 million people have been infected and 4 million have died. India is one of the hardest hit countries. Officially, India is ranked second globally in number of infections with over 30 million and third in deaths with over 400 thousand.

However, even these remarkable numbers may be an undercount. 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, there are concerns that deaths may be underreported. There have been reports of deaths being underreported due to lack of capacity to test and register the dead and to avoid blame directed at government officials (e.g., S, 2021). In order to both determine the full impact of the pandemic and to accurately determine the efficacy of policy interventions such as lockdowns and medical interventions such as vaccines, we need to more accurately estimate outcome such as infections and deaths.

We use data from the household rosters of a large, panel data set to estimate all-cause mortality in India during the pandemic. The data set is the Centre for Monitoring Indian Economy's 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 4.5 million excess deaths, roughly 13 times the number of COVID deaths reported[1]. Excess deaths peak in the same months as infections peaked during the two waves that struck India (September 2020 for wave I and April-May 2021 for wave II). Moreover, we find that the second wave experienced significantly greater deaths than the first wave. Although we do not find statistically significant differences in mortality by sex or urban location, we do find that the age-pattern of deaths is COVID-like: deaths rise significant 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.

[1]The officially reported number of deaths till May 31 was 3,31,911. Following that there was a reconciliation in the number of deaths and additional deaths were added on June 9-13, on July 12, and on July 20. These total to about 13,000 additional deaths that we add to the tally as of May 31 to get 3,44,911 deaths.

Incidentally, the excess deaths are higher in families with a higher per capita income.

This use of a household roster in 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. 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.[2] We address this in two ways. Primarily, we restrict 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 death rate reported in month t as a moving average of death rates from months $t-3$ to t. Second, as a robustness test, we estimate death rate in month t by asking, what would the change in actual death rate would have to be to generate the change in reported death rate that we observe?

We compare our estimates with estimates of excess deaths using the Civil Registry System (CRS) data from 14 states that such data is available for. Our estimates of excess deaths is about twice the estimate from CRS data. This may be due to limitations in capacity of states to register all deaths (Deshmukh et al., 2021). We also compare our estimates to those from the US. Whereas we report a 24% increase in death rates during the pandemic, the US reports a roughly 22% increase in excess deaths. Even our higher end estimates of deaths are in the range of US estimates.

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. To address the problem of incomplete registries, we employ a large, representative survey (CPHS) that is independent of political influence and allows estimation of hetergeneity in death rates.

Alternatives 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 14 of 28 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

[2]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.

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 estimate 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. To be fair, our estimates are not out of line with available excess death estimates from other sources.

A second contribution is to show how best to use the CPHS roster 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 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.

Our analysis has limitations. First, it does not estimate deaths from COVID: it estimates excess deaths during COVID relative to the number of deaths during a control period (e.g., 2019), before COVID. The deaths could be due to policies cause by COVID, such as lockdowns, or behavior responses to COVID. Second, not only do we not provide deaths for which COVID is a proximate cause, it is unclear that we provide estimates that are even a but-for cause, i.e., deaths that would not happen without COVID. To provide causal estimates, we need to get the control group right. With a stable population, this requires a valid control group. There are no groups untouched by COVID in 2020-21, so we use 2019 as a control period. But death rates may have a trend even without COVID, and we do not estimate that. Moreover, there was a jump in deaths in 2019. We do not believe this is a pre-trend because the age profile of deaths in 2020 is right-skewed in age relative to pre-2020 deaths. However, using a longer baseline from 2015-2019 lead to implausibly large estimates of death.

Section 1 provides some background on mortality rates in India and India's COVID pandemic. Section 2 presents the CPHS data, discusses our approach to addressing the challenges of using that data, and describes how we estimate both excess deaths and heterogeneous death rates. Section 3 presents our estimates of excess deaths during COVID from CPHS, shows how those estimates vary by time, demographics and infections, and compares our results to estimates from other sources. Finally, section 4 examines the policy implications of our findings and some of the limitations of

our analysis.

1 Background

According to data from the Global Burden of Disease project, reproduced in Figure 1, India had a death rate of roughly 0.7%, or 7 deaths per 1000 persons per year, in 2019. This implies a loss of approximately 9.5 million persons our of a population of almost 1.4 billion (Vos et al., 2020). Note that there was a slight uptick in the death rate in 2019, a pattern we will revisit in the CPHS data examined in this paper.

SARS-CoV-2 hit India in two waves (Figure 2). The first confirmed cases were reported on January 27, 2020, in Kerala state amongst students returning from Wuhan University (Andrews et al., 2020). The first wave peaked in late September 2020, with almost 100,000 daily confirmed cases and 1,000 daily deaths. The second wave peaked in April 2021, with roughly 400,000 daily confirmed cases and 4,000 daily deaths (www.covid19india.org, 2021).

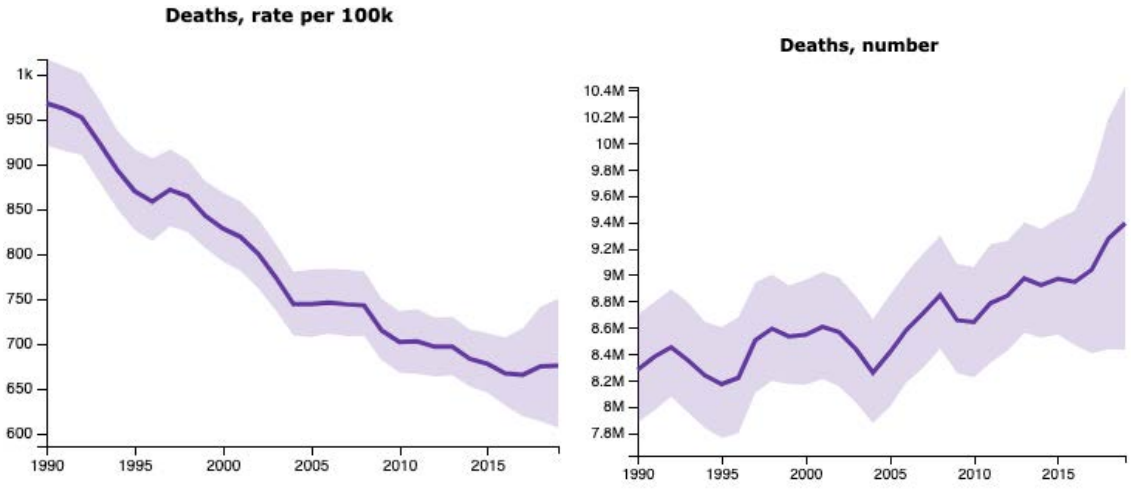
One should be cautious in interpreting these official numbers. Confirmed cases undercount actual infections, and at different rates over time. One reason is that perhaps 90% of cases, at least in 2020 with the original variant, were asymptomatic and asymptomatic cases are less likely to lead to testing. Another reason is that per capita testing rates in India were low relative to developed countries. Moreover, they increased dramatically from wave I to II, meaning that some of the higher case counts may be due to testing rather than cases.

As mentioned earlier, serological data suggests that confirmed cases may undercount infections by up to 2 orders of magnitude. Early studies suggested that perhaps 45% of the population was infected during the first wave. A serological survey in Mumbai found that, by July 2020, COVID seroprevalence[3] was 58% in slums (versus 17% in non-slums) (Malani et al., 2020) and slums account for roughly 40% of Mumbai's population. By end of August another study estimated that seroprevalence was 46% of the entire state of Karnataka (Mohanan et al., 2021). A nationwide survey by ICMR found that 67.6% of sampled individuals were seropositive after India's second wave (Sharma, 2021). With a population of roughly 1.35 billion, this implies almost 600 million infections after wave 1 and 900 million after wave 2. Contrast that to roughly 10 million cases confirmed after wave 1 and 30 million after wave 2.

Reported COVID death rates may also be lower than true death rates. First, not all deaths in

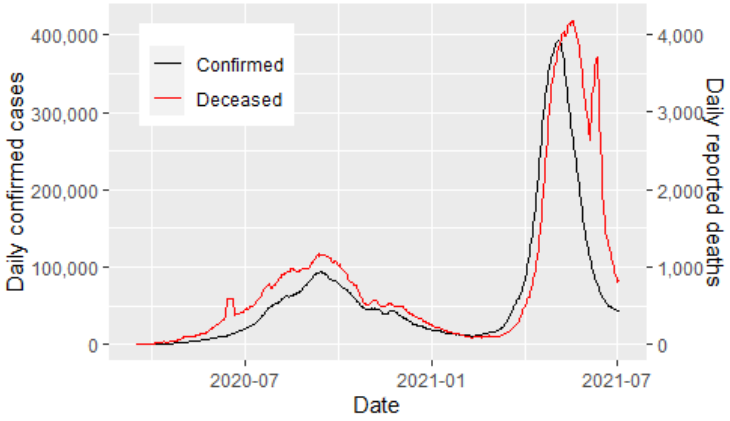
[3]COVID seroprevalence is the fraction of individuals with antibodies to COVID.

Figure 1: Death rate and total deaths in India over time.



Source: `ghdx.healthdata.org/gbd-results-tool`.

Figure 2: Confirmed COVID cases and deaths over time.



Source: `www.covid19india.org`.

India are officially recorded. Deaths that do not occur in hospitals may not be recorded in official systems. Second, not all individuals dying of COVID are tested for COVID and such deaths are missed in the official counts. Further, even individuals dying after a positive COVID death are sometimes recorded as a non-COVID death because they have some co-morbidities that could have been the cause of death. While all such deaths are not recorded as COVID deaths, some of them may have to be attributed to COVID and are missed.

Because many deaths attributable to COVID may, therefore, not be labeled as COVID deaths, researchers have examined all-cause mortality to gauge the impact of the pandemic. Usually this is done by comparing the level of deaths prior to the pandemic against the level of deaths during the pandemic. In India, all-cause mortality is recorded by each state via the Civil Registry System (CRS). The main alternative is the Sample Registration System (SRS), which calculates death rates based on a representive, 1% sample of the population.

However, 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 in death rates. Three, CRS reports with some delay Ravi (2021a). For example, only 14 (of 28) states have CRS data currently available (S, 2021).

The SRS also has problems. While the CRS is delayed a few months, the SRS is typically delayed 2 years. So, we may not get estimates of COVID-period deaths until 2023. Moreover, even the SRS is estimated to miss about 12% of deaths Gerland (2014). Indeed, 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 (2021a).

Even if one can measure excess deaths, however, they may not all be linked to COVID. The death rate due to COVID is typically captured in two epidemiological parameters. One is the case fatality rate (CFR), defined as the number of confirmed deaths divided by the number of confirmed COVID cases. This is not a tremendously useful statistic because both numerator and denominator are undercounted. In any case, CFR may reflect testing rates and selection into testing as much as the harm from the disease. A better alternative is the infection fatality rate (IFR), defined

as death 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. Because those counts are also underestimates, correcting only the denominator likely led to an underestimate of the IFR Malani et al. (2020); Mohanan et al. (2021); Malani et al. (2021). 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.

One final background fact to keep in mind when interpreting the monthly time series of excess deaths is that India imposed a national lockdown towards the beginning of the pandemic in 2020. That lockdown because March 24 and ended June 1. After that date, lockdowns were local and driven by states. But by the peak of wave 1, mobility had returned to about 15% below 2019 levels (from an apogee of 40%) below those levels. Thus India's most severe lockdown occurred well before the peak of the first wave. There was a decline in mobility and local lockdowns during wave 2, but it was not as severe as during the first wave.

2 Methods

2.1 Data

2.1.1 Consumer Pyramids Household Survey

Our primary source of data is the Centre for Monitoring the Indian Economy's Consumer Pyramids Household Survey (CPHS). This is a large, representative, panel survey of Indian households. The survey sample presently includes roughly 174,000 households with 870,000 million current members. The sample, which is based off the 2011 Indian Census, is representative at the level of homogeneous regions, defined as a cluster of similar districts within a state, and at the national level.[4] Sample households are visited every 4 months, with each 4 month period called a round. However, a nationally representative sample of the households are sampled each month, so the study can be nationally representative at the monthly level as well. CPHS started in January 2014 and the

[4]The country is divided into 99 of these regions. Within each region, rural and urban strata are defined. In the rural strata, 2011 Census 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 strata, towns are divided into strata based on population. Within each size strata towns are randomly selected. Within towns, census enumeration blocks are randomly selected. In each CEB 16 households are selected.

latest data we are able to access are from July 2021.[5] The CPHS is available to the public for a subscription fee.

Although the purpose of the CPHS is to measure income and consumption of a household and its members, it also maintains a meticulous household roster. This roster tracks attrition of members and the reason for attrition. A common reason for attrition is emigration away, e.g., for marriage or work. But another reason may be death, our primary outcome. The roster measures 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. It also provides location at the district level and, within district, 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 State Civil Registration System Data

A secondary data source is counts of deaths from States' Civil Registration System (CRS). We obtain these data from the Development Data Lab and a newspaper article (Ramani, 2021). As of this writing, CRS data are available for 14 states (See Figure 9). We will use these data to benchmark our estimates from CPHS.

2.1.3 COVID cases and mobility data

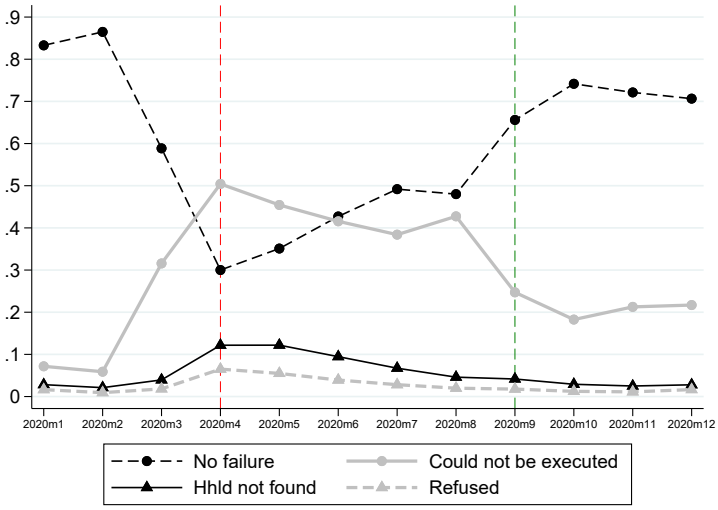
We obtain district x day level data on confirmed cases from `www.covid19india.org`. We obtain estimates of daily infections by scaling up confirmed case curves with estimates from a serological survey, as we explain in Appendix Section A.

We obtain mobility data from Google's Community Mobility Reports Google (2021). The units for Google's measure of mobility 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.

Because our death data are reported at the monthly level, we also average daily cases, infections and mobility over each month. We do not have case and mobility data prior to February 2020.

[5] CPHS started with roughly 165,000 households in 2014 and built up the sample over time. Whereas data through April 2021 are in the People of India file of the CPHS, the May and June 2021 roster is obtained from January and February 2021 income file, which is released in May and June 2021.

Figure 3: CPHS non-execution and non-response rates during 2020.



Notes: Red line indicates first month of phone surveys. Green 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.

Therefore, we assume cases and infections are 0 and that mobility is 100% before that date.

2.2 Issues with CPHS mortality data

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 concerns in turn.

2.2.1 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 households/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. First, they switched to a phone survey. Second, surveyors' managers, rather than surveyors, conducted the survey to keep the quality of phone surveys high. Because there are fewer managers than surveyors, CPHS decided not to call half the households in each strata (defined above). As a result, response rates fell. Figure 3 shows that the response rate of the subset of households that were contacted fell to roughly 60% and responding households constituted roughly 35% of the full sample at the height of the lockdown in April and May 2020. When CPHS finished its second round in August 202, 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, but there was no switch to telephone interviews and the dip was not as severe as during the lockdown.

The low response rates are partly addressed by using the responding sample as the denominator in our estimate of teh excess death rate. (We address non-random response in the next section) Moreover, one can obtain obtain information about deaths in non-responding households in later rounds. Many households who do not respond in a current round respond in later rounds, during which they report deaths since the last survey they answered. (We will address the timing of of those deaths in a few sections.)

2.2.2 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.[6] 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. Of course it could be that survey response is a function of unobservables even conditioned on observables. But given how many observables we have, this seems unlikely. If we make the assumption in, e.g.,

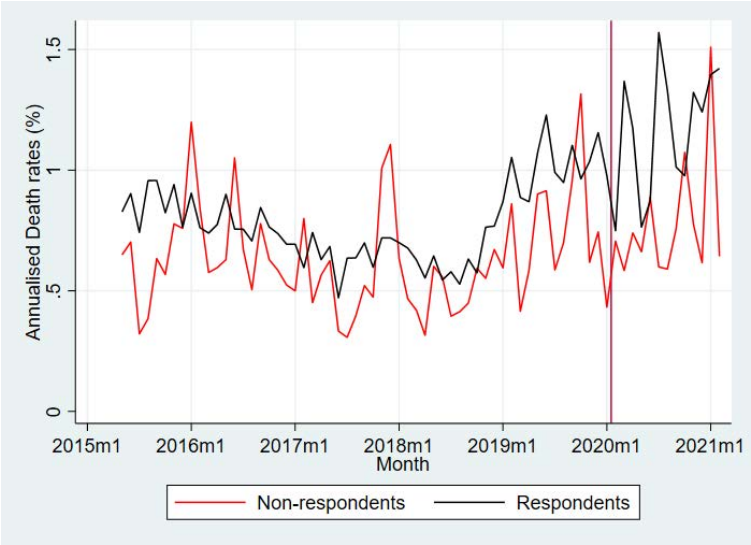
[6]The former may not affect our estimate of excess deaths during COVID if non-random response is such that the trend of deaths in responding households is similar to the trend in non-responding households, though this is an optimistic assumption.

Table 1: 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.

Figure 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.

Altonji et al. (2005) that unobservables have the same explanatory power as observables, then our low bound on the 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. Figure 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 groups death rates almost always appear to be above non-respondent group rates. 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 0.276 percentage points higher per annum before the pandemic, and rise 0.182% percentage points further above non-respondents during the pandemic, with each difference being statistically significant (Appendix Table A1). 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, the bias is different for the remainder of CPHS that does not repond at t and $t+8$. So the important take-away is that there could be non-random survey response.

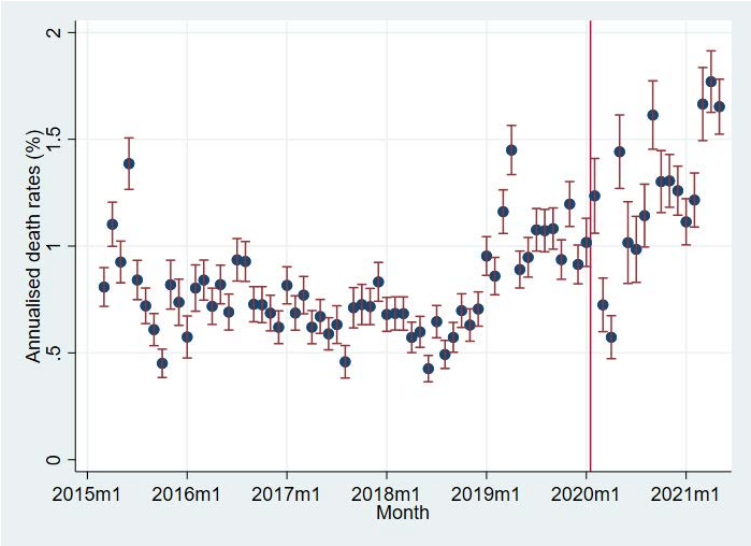
There is a solution to non-random response, but it has separate problems. Non-random response can be addressed by using both respondent and non-respondent households in estimate excess deaths. After all, non-respondents at $t+1$ typically respond at $t+2$ and this fills in holes in the death data. However, 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.

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

2.2.3 Rise in deaths in 2019

The annualized death rate calculated from the CPHS rises in 2019, a year before the pandemic (Figure 5). 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?

Figure 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. The red line demarcates the start of the pandemic in February 2020.

We do not think that the jump in 2019 is a pre-trend unrelated to COVID that continues into 2020. As Figure 8 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 the pandemic. The argument for using 2019 is that the implied excess deaths seem less outlandish and thus more plausible, though the excess deaths estimates remains 8 times the reported number of COVID deaths. The argument against using 2019 as a baseline is that it implies a baseline death rate of 1.07%, 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 (7.9%) 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, out 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.

2.2.4 Timing of deaths

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

The problem is illustrated with the example in Table 2. 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.

We offer 2 solutions, each of which has pros and cons. Our preferred solution is to reallocate death to the midpoint between answered surveys. So if there is a gap of k months between surveys the household answered, then a death reported on month t is re-allocated to month $t-(k/2)$. This solution, which is illustrated in row 3, is akin to treating the reported number at t as a moving average of the true death rate for $t-(k/2)$ (row 4). We implement this solution with a regression of the form,

$$d_{i,t,t-k} = \delta + \beta \cdot \mathbb{I}(t-(k/2) \in \text{pandemic}) + \epsilon_{it} \tag{1}$$

where $d_{i,t,t-k}$ is an indicator for whether an individual was last reported alive in month $t-k$ and reported dead in month t and $\mathbb{I}(t-(k/2) \in \text{pandemic})$ is an indicator for whether the month to which the death is reassigned was during the pandemic. Standard errors are clustered at the village/ward $\times$ month level to account for correlation in reporting of deaths within a locality.

The advantage of the first solution is that it is simple. The disadvantage is that 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.

Table 2: 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

The second solution is 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. We implement this solution with a regression of the form

$$d_{i,t,t-k} = \sum_{k=4,8,\ldots} \delta_k \cdot \mathbb{I}(k) + \sum_{m \in \text{pandemic}} \beta_m \cdot \mathbb{I}(t-k \leq m \leq t) + \epsilon_{ist} \tag{2}$$

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 $\mathbb{I}(t-k \leq m \leq t)$ is an indicator whether the intervening period between surveys was during the pandemic. Standard errors are clustered at the village/ward $\times$ month level to account for correlation in reporting of deaths within a locality.

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.

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 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, but we also report the second measure in Appredix D.

2.2.5 Representativeness of CPHS

Dreze and Somanchi (2021) argues that CPHS undersamples 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 overweights 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).

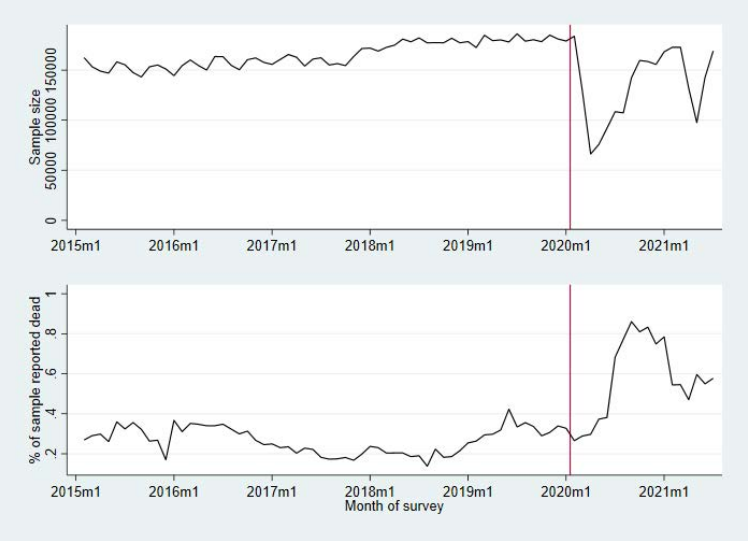
Of course, the fact of difference between CPHS and government surveys of literacy is not dispositive of whether CPHS is biased since it is possible that the government surveys are the ones that are off. After all, the government has taken steps to suppress data (e.g., the 2017-18 consumption survey by 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.[7] 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).

2.3 Definition of pandemic period

We define the pandemic period as starting in February 2020 and persisting to the present. India's first confirmed cases are on January 31, 2020. We define the first wave as lasting from February - December 2020 and the second wave as from January 2021 - the present. The CPHS data are coded at the monthly level, so we cannot more finely define the start of the pandemic or waves. However, in robustness exercises, we vary the start and stop dates +/- 2 months for our preferred estimation strategy.

[7]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.

Figure 6: Sample size and households reporting a death from January 2015 - June 2021.



Notes: The sample includes 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 is Panel B is the death rate over 4 months (i.e., since the prior survey of that household). The data are not weighted to be representative. The red line demarcates the start of the pandemic in February 2020.

3 Results

Raw data from CPHS at the national level suggests a jump in the death rate during the COVID pandemic. Panel A of Figure 6 shows the sample size between January 2015 and June 2021 and Panel B presents death rate as of the date the deaths are reported (not the date they occurred). There is a large rise during Wave 1 and the beginning of a rise in Wave 2. This is consistent with a pandemic-associated rise in deaths, but is implausibly high: 4-month death rates death rates from 0.4% to 0.8% of households reporting a death.

3.1 Main estimates

When we clean the data based on our first (and preferred) solution to the timing-of-death problem and re-weight the data, we obtain a time series that shows a more moderate increase in death rates duing COVID. Specifically, we focus on the sample that includes only households that respond in consecutive rounds, so as 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 $k = 2$. We estimate a version of the regression in (1) where we replace the

pandemic indicator with month fixed effects. Finally, we weight observations to make responding households nationally representative.[8]

A plot of the resulting monthly fixed effects (Figure 7) shows 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 three spikes during the pandemic. One spike is June 2020, when lockdown is released, another is September 2020, which wave 1 peaks, and the last is in March - May 2021, when wave 2 peaked. These spikes are significantly greater than adjacent months and all but one 2019 month. The Wave 1 and 2 spikes are greater than even the highest 2019 peak.

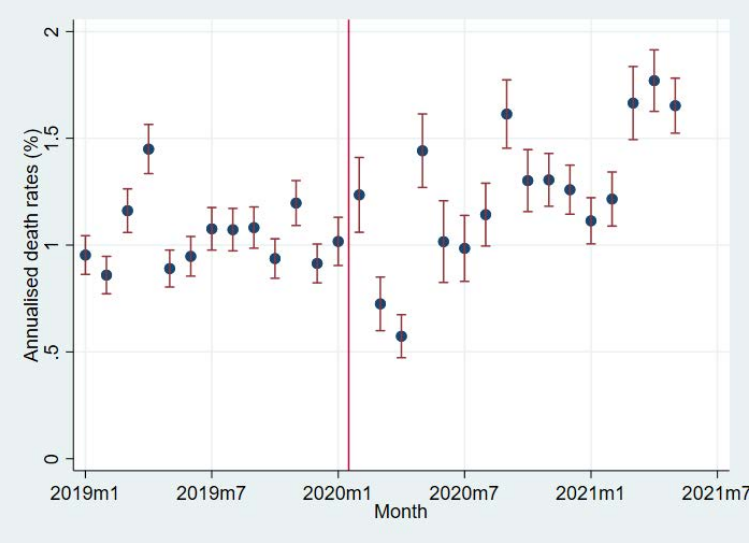
The estimate excess death rate during the pandemic depends on our baseline. To estimate a single annualised excess death rate for the pandemic, we replace the month fixed effects in our last analysis with a baseline indicator and a pandemic indicator (Table 3). When we define the baseline as the period 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 0.501%, which represents a hard-to-believe increase of over 50%. When we use only 2019 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 is 0.250%, a relative increase that is in line with the relative increase in excess deaths from the US. Our estimate of the excess death rate falls a bit to 0.229% or 0.235% (columns 3 and 4) if we include fixed effects for homogeneous region or district, respectively. Critically (from a policy perspective), the death rate is significantly greater during the second wave than the first wave (column 5).

If we include in our estimation of (1) not just households that respond in consecutive rounds of the survey (roughly 82% of the sample), but also ones that skip rounds, we obtain somewhat higher estimates of excess mortality during COVID (0.251 with responders that skip up to 1 round, 0.301 if up to 2 rounds, Appendix Table A2). This seems inconsistent with the nature of non-response bias we discussed in Section 2.2.2. Recall, however, 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.

We obtain similar estimates of excess deaths in the pandemic when we use our second solution to

[8]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.

Figure 7: Time series of death rates from month fixed effects, January 2019 - April 2021.



Notes: 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 coefficient and each whisker is the 95% confidence interval on month fixed effects in the regression in (1), where we replace the pandemic indicator with month fixed effects. The red line demarcates the start of the pandemic in February 2020.

Table 3: COVID death rate prior to the pandemic and the excess death rate during the pandemic

	Annualised death rates (%)				
	(1)	(2)	(3)	(4)	(5)
During Pandemic	0.501***	0.250***	0.229***	0.235***	
	(0.0400)	(0.0478)	(0.0467)	(0.0455)	
Baseline	0.787***	1.038***	2.527***	2.182**	1.038***
	(0.0118)	(0.0286)	(0.385)	(0.751)	(0.0286)
Wave 1					0.153**
					(0.0554)
Wave 2					0.427***
					(0.0699)
Controls	None	None	HR FE	District FE	None
Data used	2015-21	2019-21	2019-21	2019-21	2019-21
N	7837848	2816368	2816368	2816368	2816368

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 (1). 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 0.078 and the effect remains significant. */**/** indicates $p < 0.05/0.01/0.001$.

the timing-of-death problem, though our monthly estimates of excess deaths is highly variable and is sensitive to the definition of when the pandemic starts. We explore this less-preferred solution in Appendix Section D.

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 A3). 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 first explore heterogeneity in excess deaths 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

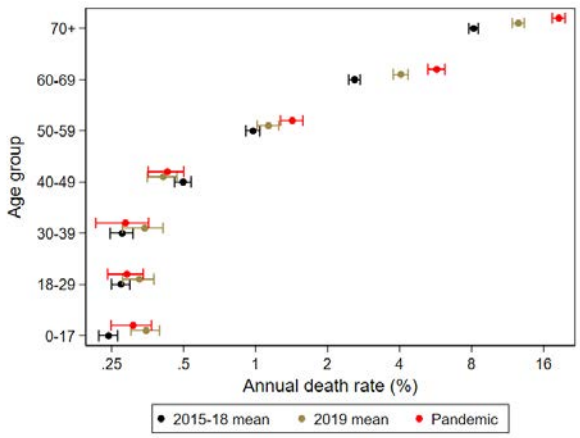
Age. Excess deaths follow a COVID-like pattern with respect to age (Figure 8 below and Table A5 in the Appendix). Excess deaths are significantly positive for higher ages, and insignificant and close to zero for lower ones. This right-skew is somewhat greater in wave 2 than in wave 1.

Gender and location. We do not see COVID-like patterns, however, with respect to gender or location. Estimated CFR and IFR is greater among males (Green et al., 2021; Pastor-Barriuso et al., 2020), but we do not find significantly greater excess deaths among males. Likewise, prior studies have shown greater infection rates in urban areas Mohanan et al. (2020); Malani et al. (2021), but we do not find significantly greater excess death in cities during the pandemic.

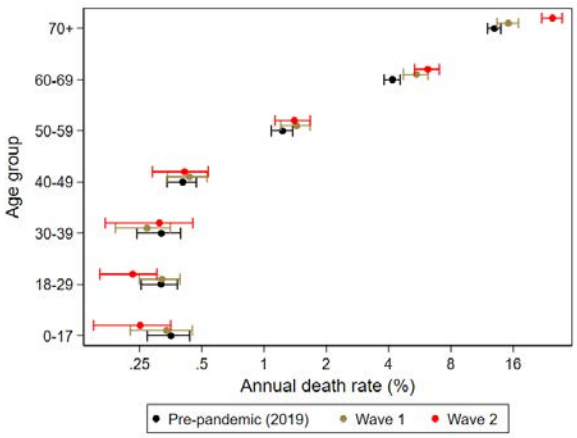
Infections. Excess deaths are positively correlated with confirmed cases and with infection. Table 5 reports the results of an individual-level regression based on (1), except that we replace the pandemic indicator with confirmed cases or infections. Cases and infections are reported as monthly averages 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.

Figure 8: Age pattern of death rates during the pandemic

Panel A: Pandemic overall



Panel B: Pandemic waves



Notes. Estimates are from a regression model based on 1, with some modification. In Panel A, we use data from 2015 onwards and add a indicator for 2019. In Panel B, we use data from 2019 onwards. In each row of a figure we present coefficients from a regression that includes data on members in the indicated age group only.

Table 4: Annualised death rates by gender and urban status

Panel A: Gender

	Annualised death rates(%) (1)	(2)
Pandemic	0.244***	
	(0.0557)	
Wave 1		0.142*
		(0.0638)
Wave 2		0.432***
		(0.0830)
Pandemic × Female	0.0111	
	(0.0664)	
Wave 1 × Female		0.0235
		(0.0782)
Wave 2 × Female		-0.0123
		(0.0978)
Female	-0.000111	-0.000111
	(0.000126)	(0.000126)
2019 mean	1.054***	1.054***
	(0.0340)	(0.0340)
N	2816368	2816368

Panel B: Urban/rural

	Annualised death rates(%) (1)	(2)
Pandemic	0.212***	
	(0.0593)	
Wave 1		0.176*
		(0.0717)
Wave 2		0.278***
		(0.0779)
Pandemic × Urban	0.118	
	(0.0996)	
Wave 1 × Urban		-0.0691
		(0.109)
Wave 2 × Urban		0.453**
		(0.158)
Urban	-0.000105	-0.000105
	(0.000194)	(0.000194)
2019 mean	1.048***	1.048***
	(0.0362)	(0.0362)
N	2816368	2816368

Notes. Estimates are from a regression model based on 1, with the addition of a gender (urban) indicator and gender (urban) indicator interacted with the pandemic or wave indicator. Sample includes only consecutive observations and excludes emigrants. Standard errors clustered at the village/ward × month level are reported in parentheses. $p < 0.05/0.01/0.001$.

Table 5: Correlation of death rates with cases, infections, and mobility

	Annualised death rates (%)				
	(1)	(2)	(3)	(4)	(5)
Infections (sero scaled)	0.0286***	0.0276***			
	(0.00421)	(0.00426)			
Cases			0.00000118***	0.000000969**	
			(0.000000351)	(0.000000362)	
Mobility		-0.000264		-0.000387*	-0.000599***
		(0.000158)		(0.000161)	(0.000153)
2019 mean	1.074***	1.058***	1.134***	1.104***	1.112***
	(0.0260)	(0.0267)	(0.0253)	(0.0260)	(0.0258)
N	2648617	2610715	2638509	2638509	2638509

Notes. Estimates in columns 1 and 2 are from a regression of death rates against sero scaled infections from an SIR model. Estimates in columns 3 and 4 are from a regression of death rates against officially reported COVID cases. We control for mobility in columns 2 and 4. Estimates in column 5 are from a regression of death rates against google mobility. Standard errors clustered at the village/ward $\times$ month level are reported in parentheses. $p < 0.05/0.01/0.001$.

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 as well as indirect effect of association 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, Figure 7, 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. Because we employ 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 5 also reports the results of an individual-level regression based on (1), except that we replace the pandemic indicator with Google's mobility index. Our estimated coefficient in column 5 is that a 10% reduction in mobility was associated with a 0.5% increase in the 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 income terciles and pandemic or wave indicators to the regression in (1). This evidence suggests that pandemic had a bigger mortality impact on the top tercile (column 1), but that this imbalance was more pronounced in wave 2 (column 2). The first wave affected the middle and highest tercile more than the lowest. The second wave affected only the top tercile more than the bottom tercile.

Table 6: Annualised death rates by income groups

	Annualised death rates(%)					
	(1)	(2)	(3)	(4)	(5)	(6)
Pandemic	0.113		0.237		0.0823	
	(0.0689)		(0.122)		(0.0801)	
Pandemic × 2nd tercile	0.132		-0.0223		0.177	
	(0.0867)		(0.121)		(0.109)	
Pandemic × 3rd tercile	0.357***		0.235		0.387**	
	(0.0986)		(0.129)		(0.146)	
Wave 1		-0.0247		0.0416		-0.0411
		(0.0767)		(0.135)		(0.0893)
Wave 1 × 2nd tercile		0.239*		-0.00842		0.338**
		(0.102)		(0.141)		(0.129)
Wave 1 × 3rd tercile		0.367**		0.154		0.569**
		(0.114)		(0.149)		(0.181)
Wave 2		0.371***		0.591**		0.315**
		(0.102)		(0.190)		(0.118)
Wave 2 × 2nd tercile		-0.0682		-0.0459		-0.124
		(0.122)		(0.174)		(0.151)
Wave 2 × 3rd tercile		0.328*		0.372		0.0501
		(0.147)		(0.194)		(0.193)
2019 mean	1.160***	1.160***	1.170***	1.170***	1.158***	1.158***
	(0.0463)	(0.0463)	(0.0746)	(0.0746)	(0.0541)	(0.0541)
2nd tercile	-0.172**	-0.172**	-0.150*	-0.150*	-0.184**	-0.184**
	(0.0540)	(0.0540)	(0.0715)	(0.0715)	(0.0671)	(0.0671)
3rd tercile	-0.232***	-0.232***	-0.232**	-0.232**	-0.242**	-0.242**
	(0.0567)	(0.0567)	(0.0726)	(0.0726)	(0.0795)	(0.0795)
Sample	All	All	Urban	Urban	Rural	Rural
N	2816368	2816368	1852668	1852668	963700	963700

Notes. Estimates are from a regression model based on 1, with the addition of a 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 percentile are in income tercile 2 and households between 67 and 100 percentile are in income tercile 3. Columns 1 and 2 includes all data, columns 3 and 4 include only urban regions and columns 5 and 6 include only rural regions. 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. $p < 0.05/0.01/0.001$.

4 Discussion

Our preferred estimates imply that there were 4.5 million excess deaths in India during the pandemic, 1.9 million during wave 1 and 2.4 million during wave II. These estimates only include consecutive responders and are similar to estimates if we include households that skip up to one round of surveys.

Our preferred estimate of death is roughly 13x as large as the official number of COVID deaths. This does not prove that COVID caused 13x 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. Due to power considerations in the CPHS data, we do not compare deaths at the state-level amongst the 14 states for which CRS data is presently available. Instead we compare the sum of deaths across the 14 stats. CRS data imply 1.7 million excess deaths across the 14 states from February 2020 - May 2021. The estimate using CPHS data in the same states and period is 3.4 million. While our estimates are substantially higher, it is well known that CRS substantially undercounts deaths (Ravi, 2021a). Moreover, our estimates of excess deaths are closer to what estimates that correct for undercounting in the CRS or that look at surveys other than CPHS find (Anand et al., 2021; Deshmukh et al., 2021). Our preferred estimate is somewhat higher than the highlighted estimates in those papers, but are within their range of estimates from those papers. Moreover, our estimate implies an percentage increase in the death rate that is in line with the percentage increase found in the US.

Ravi (2021b) has criticized CPHS as being inappropriate for measuring. (She has also criticized the CRS for being incomplete.) She prefers that SRS be used to measure excess deaths because it checks each death twice and has a sample size 10 times larger than CPHS.[9] Checking deaths twice will tend to reduce sampling error, but if this error is mean zero, it should have small effects given our sample size of nearly 850,000 persons. Moreover, the precision benefits from sample size are diminishing with sample size. So 8.5 million observations may not yield a much better estimate than 850 thousand. An even better argument for considering CPHS is that SRS will not be made

[9]Ravi also notes that CPHS has an unexplaned spike in 2019. This spike is also in GPD estimates, but is smaller. We addressed the 2019 spike in Section 2.2.3.

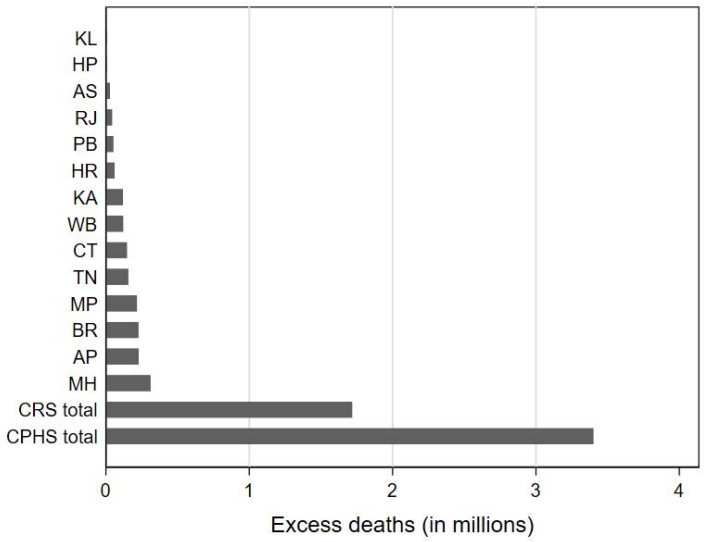


Figure 9: Excess Deaths from CRS data

Notes. The excess deaths from CRS is calculated by subtracting the number of deaths per month in or after Feb 2020 from the number of deaths per month in 2018 and 2019. The data for Maharashtra, Punjab, West Bengal, Haryana, Himachal Pradesh, and Kerala are obtained from Ramani (2021). The data for Karnataka, Madhya Pradesh, Assam, Rajasthan, Chhattisgarh, Tamil Nadu, Bihar, and Andhra Pradesh were accessed through Development Data Lab (`http://www.devdatalab.org/`).

available for 2 years. Moreover, because it is a government survey, it may be met with skepticism about whether it is manipulated. No data set is perfect for measuring excess deaths from COVID, but CPHS provides a reasonable, timely estimate.

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, our data show a big jump in death rates in 2019. This might cast doubt on the validity of our 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 two solutions to the problem and they produce roughly similar results. Moreover, alternatives such as the CRS have their own problems. CRS and SRS are undercounted. CRS has a measurement error problem: death a reported where they occur not where people live; so, in years past, death counts in some states were 100% what the SRS estimated.

Finally, 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 sample comprise 95% of our sample.

References

Altonji, J. G., T. E. Elder, and C. R. Taber (2005). Selection on observed and unobserved variables: Assessing the effectiveness of catholic schools. *Journal of political economy 113*(1), 151–184.

Anand, A., J. Sandefur, and A. Subramanian (2021). Three new estimates of india's all-cause excess mortality during the covid-19 pandemic.

Andrews, M. A., B. Areekal, K. R. Rajesh, J. Krishnan, R. Suryakala, B. Krishnan, C. P. Muraly, and P. V. Santhosh (2020). First confirmed case of covid-19 infection in india: A case report. *The Indian journal of medical research 151*(5), 490–492.

Deshmukh, Y., W. Suraweera, C. Tumbe, A. Bhowmick, S. Sharma, P. Novosad, S. H. Fu, L. Newcombe, H. Gelband, P. Brown, and P. Jha (2021). Excess mortality in india from june 2020 to june 2021 during the covid pandemic: death registration, health facility deaths, and survey data. *medRxiv*, 2021.07.20.21260872.

Dreze, J. and A. Somanchi (2021). View: New barometer of india's economy fails to reflect deprivations of poor households. *The Economic Times June 21*(June 21).

Gerland, P. (2014). Un population division's methodology in preparing base population for projections: Case study for india. *Asian Population Studies 10*(3), 274–303.

Google (2021). Covid-19 community mobility report - india. Report.

Green, M. S., D. Nitzan, N. Schwartz, Y. Niv, and V. Peer (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.

Khandekar, O. (2021). How india's slums are faring against the second covid19 wave. *Livemint April 29*(April 29).

Kontopantelis, E., M. A. Mamas, J. Deanfield, M. Asaria, and T. Doran (2021). Excess mortality in england and wales during the first wave of the covid-19 pandemic. *Journal of Epidemiology and Community Health 75*(3), 213.

Malani, A., S. Ramachandran, V. Tandel, R. Parasa, S. Sudharshini, V. Prakash, Y. Yoganathan, S. Raju, and T. Selvavinayagam (2021). Seroprevalence in tamil nadu in october-november 2020. *MedRXiv*.

Malani, A., D. Shah, G. Kang, G. N. Lobo, J. Shastri, M. Mohanan, R. Jain, S. Agrawal, S. Juneja, S. Imad, and U. Kolthur-Seetharam (2020). Seroprevalence of sars-cov-2 in slums versus non-slums in mumbai, india. *The Lancet Global Health*.

Mohanan, M., A. Malani, K. Krishnan, and A. Acharya (2020). Prevalence of covid-19 in rural versus urban areas in a low-income country: Findings from a state-wide study in karnataka, india. *medRxiv*, 2020.11.02.20224782.

Mohanan, M., A. Malani, K. Krishnan, and A. Acharya (2021). Prevalence of sars-cov-2 in karnataka, india. *JAMA 325*(10), 1001–1003.

Pastor-Barriuso, R., B. Pérez-Gómez, M. A. Hernán, M. Pérez-Olmeda, R. Yotti, J. Oteo-Iglesias, J. L. Sanmartín, I. León-Gómez, A. Fernández-García, P. Fernández-Navarro, I. Cruz, M. Martín, C. Delgado-Sanz, N. Fernández de Larrea, J. León Paniagua, J. F. Muñoz-Montalvo, F. Blanco, A. Larrauri, and M. Pollán (2020). Infection fatality risk for sars-cov-2 in community dwelling population of spain: nationwide seroepidemiological study. *BMJ 371*, m4509.

Ramani, S. (2021). Excess deaths in maharashtra were at least 3 times the official covid toll. *The Hindu August 4*(August 4).

Rao, C. and M. Gupta (2020). The civil registration system is a potentially viable data source for reliable subnational mortality measurement in india. *BMJ global health 5*(8), e002586.

Ravi, S. (2021a). Counting deaths in india is difficult. *Hindustan Times July 14, 2021*(July 14, 2021).

Ravi, S. (2021b). Covid deaths: Collective indifference to vital statistics. *The Indian Express August 9*(August 9).

Rossen, L. M., A. M. Branum, F. B. Ahmad, P. Sutton, and R. N. Anderson (2020). Excess deaths associated with covid-19, by age and race and ethnicity - united states, january 26-october 3, 2020. *MMWR. Morbidity and mortality weekly report 69*(42), 1522–1527.

S, R. (2021, July 6, 2021). Gauging pandemic mortality with civil registration data. *The Hindu*.

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

Vos, T., S. S. Lim, C. Abbafati, K. M. Abbas, M. Abbasi, M. Abbasifard, M. Abbasi-Kangevari, H. Abbastabar, F. Abd-Allah, A. Abdelalim, M. Abdollahi, I. Abdollahpour, H. Abolhassani, V. Aboyans, E. M. Abrams, L. G. Abreu, M. R. M. Abrigo, L. J. Abu-Raddad, A. I. Abushouk, A. Acebedo, I. N. Ackerman, M. Adabi, A. A. Adamu, O. M. Adebayo, V. Adekanmbi, J. D. Adelson, O. O. Adetokunboh, D. Adham, M. Afshari, A. Afshin, E. E. Agardh, G. Agarwal, K. M. Agesa, M. Aghaali, S. M. K. Aghamir, A. Agrawal, T. Ahmad, A. Ahmadi, M. Ahmadi, H. Ahmadieh, E. Ahmadpour, T. Y. Akalu, R. O. Akinyemi, T. Akinyemiju, B. Akombi, Z. Al-Aly, K. Alam, N. Alam, S. Alam, T. Alam, T. M. Alanzi, S. B. Albertson, J. E. Alcalde-Rabanal, N. M. Alema, M. Ali, S. Ali, G. Alicandro, M. Alijanzadeh, C. Alinia, V. Alipour, S. M. Aljunid, F. Alla, P. Allebeck, A. Almasi-Hashiani, J. Alonso, R. M. Al-Raddadi, K. A. Altirkawi, N. Alvis-Guzman, N. J. Alvis-Zakzuk, S. Amini, M. Amini-Rarani, A. Aminorroaya, F. Amiri, A. M. L. Amit, D. A. Amugsi, G. G. H. Amul, D. Anderlini, C. L. Andrei, T. Andrei, M. Anjomshoa, F. Ansari, I. Ansari, A. Ansari-Moghaddam, C. A. T. Antonio, C. M. Antony, E. Antriyandarti, D. Anvari, R. Anwer, J. Arabloo, M. Arab-Zozani, A. Y. Aravkin, F. Ariani,

J. Ärnlöv, K. K. Aryal, A. Arzani, M. Asadi-Aliabadi, A. A. Asadi-Pooya, B. Asghari, C. Ashbaugh, D. D. Atnafu, et al. (2020). Global burden of 369 diseases and injuries in 204 countries and territories, 1990x2013;2019: a systematic analysis for the global burden of disease study 2019. *The Lancet 396*(10258), 1204–1222.

Vyas, M. (2021). View: There are practical limitations in cmie's cphs sampling, but no bias. *The Economic Times June 23*(June 23).

Woolf, S. H., D. A. Chapman, R. T. Sabo, and E. B. Zimmerman (2021). Excess deaths from covid-19 and other causes in the us, march 1, 2020, to january 2, 2021. *JAMA 325*(17), 1786–1789.

www.covid19india.org (2021). Coronavirus outbreak in india. Report.

Appendix

A 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 districts 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.

Table A1: Death rate by consecutive survey response status

	Annualised death rates(%)
[1em] Pandemic	-0.0265
	(0.0547)
Respondent	0.267***
	(0.0463)
Pandemic × Respondent	0.182**
	(0.0638)
2019 Non-respondent mean	0.780***
	(0.0426)
N	1907785

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 A2: Excess death rates for different samples of responders

	Annualised death rates(%)			
	(1)	(2)	(3)	(4)
Pandemic	0.250***	0.251***	0.301***	0.323***
	(0.0330)	(0.0311)	(0.0309)	(0.0308)
2019 mean	1.038***	1.075***	1.083***	1.097***
	(0.0189)	(0.0185)	(0.0185)	(0.0187)
Sample composition	Responses 4 mo apart	Responses 4/8 mo apart	Responses 4/8/12 mo apart	All
% of full sample	83 %	95 %	99 %	100 %
N	2816368	3214238	3334706	3382762

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 Non-random response

Table A1 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. Table A2 present estimates of baseline death rate in 2019 and excess deaths during the pandemic 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 A2 have the same sample as that in Table A1.

C Comparing different pandemic definitions

Table A3 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 A3: Robustness of excess death rates to pandemic definition

	Nov 2018	Dec 2018	Jan 2019	Feb 2019	Mar 2019
Dec 2019	0.25	0.22	0.20	0.19	0.16
	0.99	1.02	1.05	1.06	1.08
Jan 2020	0.28	0.26	0.23	0.22	0.20
	0.99	1.02	1.04	1.05	1.07
Feb 2020	0.30	0.27	0.25	0.24	0.22
	0.99	1.02	1.04	1.05	1.06
Mar 2020	0.29	0.27	0.25	0.24	0.22
	1.00	1.02	1.04	1.05	1.07
Apr 2020	0.33	0.30	0.28	0.28	0.26
	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 (1).

D Second solution to the timing-of-death problem

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 Figure A1 Panel A, which plots monthly excess death rates (relative to 2019 rates) based on (a) estimating (2), 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 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 Figure A1.

Including observations from households that may not answer consecutive surveys mitigates the saw-tooth pattern (Figure A1 Panel B). 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 A4 reports our estimate of excess deaths 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 A4: Excess death rates under the second timing solution for different pandemic definitions

Panel A: Consecutive observations

Pandemic start month	Excess death rates (Annualised rate (%))	Standard error
Dec 2019	1.271	(0.611)
Jan 2020	1.150	(0.642)
Feb 2020	1.763	(0.625)
Mar 2020	1.897	(0.665)
Apr 2020	1.953	(0.711)
May 2020	2.275	(0.764)

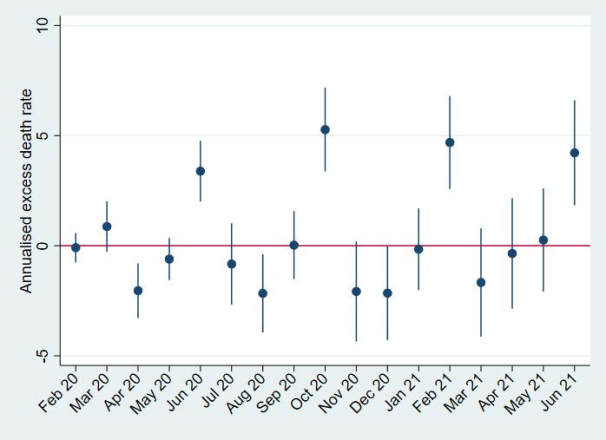
Panel B: All observations

Pandemic start month	Excess death rates (Annualised rate (%))	Standard error
Dec 2019	0.453	(0.525)
Jan 2020	0.255	(0.552)
Feb 2020	0.497	(0.510)
Mar 2020	0.778	(0.542)
Apr 2020	0.928	(0.579)
May 2020	1.358	(0.622)

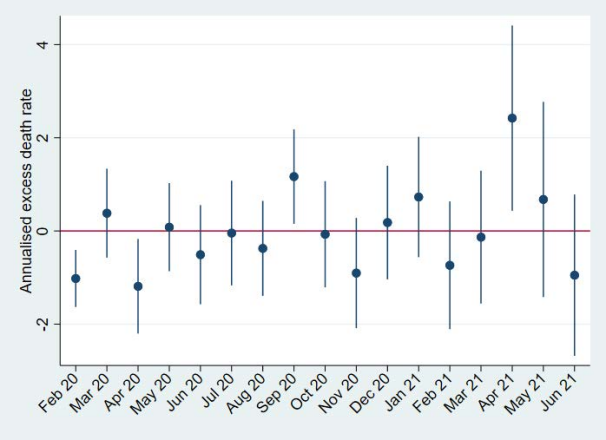
Notes. This table reports the excess death rates computed using the specification in equation 2. The different rows represent different start dates for the pandemic period. The excess deaths rates are the mean of the monthly death rates multiplied by 12. The table on the left only uses observations from households that answer consecutive surveys. The table on the right uses all observations.

Figure A1: Monthly death rates from second solution

Panel A: Consecutive observations



Panel B: All observations



Notes. This figure plots monthly death rates for each month. The coefficients from regression 2 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.

E Age-wise deaths

Table A5 reports estimates of a regression based on (1), 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 A5: Excess death rate during the pandemic and its waves in different age groups

Panel A: Pandemic overall

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	0-17	18-29	30-39	40-49	50-59	60-69	70+
During Pandemic	-0.0475	-0.0287	-0.0330	0.0223	0.195	1.542***	5.540***
	(0.0590)	(0.0421)	(0.0534)	(0.0508)	(0.117)	(0.344)	(0.938)
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	585655	571141	350753	455934	347423	164896	78370

Panel B: By waves

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	0-17	18-29	30-39	40-49	50-59	60-69	70+
Wave 1	-0.0176	0.00249	-0.0467	0.0302	0.207	1.277**	2.162**
	(0.0710)	(0.0488)	(0.0566)	(0.0585)	(0.140)	(0.423)	(0.824)
Wave 2	-0.103	-0.0863	-0.00670	0.00806	0.172	2.007***	11.79***
	(0.0669)	(0.0487)	(0.0818)	(0.0716)	(0.157)	(0.471)	(1.077)
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	585655	571141	350753	455934	347423	164896	78370

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 (1); the model in Panel B replaces the pandemic indicator in (1) 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 $p < 0.05/0.01/0.001$.