Three New Estimates of India's All-Cause Excess Mortality during the COVID-19 Pandemic

Abhishek Anand, Justin Sandefur, and Arvind Subramanian

Abstract

India lacks an authoritative estimate of the death toll from the COVID-19 pandemic. We report excess mortality estimates from three different data sources from the pandemic's start through June 2021. First, extrapolation of state-level civil registration from seven states suggests 3.4 million excess deaths. Second, applying international estimates of age-specific infection fatality rates (IFR) to Indian seroprevalence data implies a higher toll of around 4 million. Third, our analysis of the Consumer Pyramid Household Survey, a longitudinal panel of over 800,000 individuals across all states, yields an estimate of 4.9 million excess deaths. Each of these estimates has shortcomings and they also diverge in the pattern of deaths between the two waves of the pandemic. Estimating COVID-deaths with statistical confidence may prove elusive. But all estimates suggest that the death toll from the pandemic is likely to be an order of magnitude greater than the official count of 400,000; they also suggest that the first wave was more lethal than is believed. Understanding and engaging with the data-based estimates is necessary because in this horrific tragedy the counting—and the attendant accountability—will count for now but also the future.

Center for Global Development
www.cgdev.org

Working Paper 589
July 2021

Three New Estimates of India's All-Cause Excess Mortality during the COVID-19 Pandemic

Abhishek Anand
Harvard University
abhishekanand@hks.harvard.edu

Justin Sandefur
Center for Global Development
jsandefur@cgdev.org

Arvind Subramanian
Center for Global Development and Brown University
asubramanian@cgdev.org and arvind_subramanian@brown.edu

We are grateful to Kaushik Krishnan and Mahesh Vyas for comments and help with understanding the CPHS Survey. We are grateful to Murad Banaji, our discussant at the CMIE-CPHS seminar, and Stephane Helleringer for insightful comments that helped improve our paper. Participants at the seminar, including Jean Drèze, Aashish Gupta, Anup Malani, and Sabareesh Ramachandran, provided very useful feedback. Discussions with and comments from Ruchir Agarwal, Prabhat Jha, Rukmini S, Saurav Das, and Chinmay Tumbe are also gratefully acknowledged. Errors remain our own.

Supported by the UK Foreign Commonwealth and Development Office (FCDO), the Covid Collective is based at the Institute of Development Studies (IDS). The Collective brings together the expertise of UK and Southern based research partner organisations and offers a rapid social science research response to inform decision-making on some of the most pressing Covid-19 related development challenges.

Anand, Abhishek; Justin Sandefur; and Arvind Subramanian, 2021. "Three New Estimates of India's All-Cause Excess Mortality during the COVID-19 Pandemic." CGD Working Paper 589. Washington, DC: Center for Global Development. https://cgdev.org/publication/three-new-estimates-indias-all-cause-excess-mortality-during-covid-19-pandemic

We provide code and workbooks to replicate all our calculations here: https://cgdev.org/sites/default/files/anand-et-al-2021-india-excess-mortality-replication-files.zip. The CPHS microdata are available by subscription from CMIE. More information on CGD's research data and code disclosure policy can be found here: www.cgdev.org/page/research-data-and-code-disclosure.

Center for Global Development
2055 L Street NW
Washington, DC 20036

202.416.4000
(f) 202.416.4050

www.cgdev.org

Contents

1. Introduction and summary of excess death estimates

India's official Covid death count as of end-June 2021 is 400,000.[1] The reality is, of course, catastrophically worse. A sense that the official estimates are under-counting deaths is suggested by simple cross-country comparisons. The end-June official death count implies deaths per capita of 0.3 for India, whereas the comparable numbers for large countries in Europe and the Americas are substantially greater (well in excess of 3 for Mexico and Peru and around 2 in Brazil, Italy, US, and UK) even though infection rates are lower.[2]

India is one of the few major economies that does not have estimates of excess deaths during the Covid pandemic, reflected for example, in its absence from global databases such as the Human Mortality Database or the World Mortality Database (Karlinsky and Kobak 2021) which underlie reporting by OurWorldinData.org and the Financial Times. The only all-India numbers we have are model-based estimates of all-cause excess mortality (for example, the Institute for Health Metrics and Evaluation (IHME) and The Economist) and some data-based estimates for the first wave (Banaji, 2021a).[3] A prominent set of estimates were published by the New York Times but they were based on hypothetical infection and infection fatality rates that were not linked to India data per se.[4][5]

But this picture is fast changing. And we are now, for the first time, getting data-based estimates of excess deaths at an all-India level. This owes, in part, to the heroic efforts of a number of journalists, newspapers, and researchers who have used domestic laws and unrelenting investigative sleuthing to obtain accurate and timely official data during the apocalyptic second wave.[6] These data are compiled by the Indian states as part of the civil registration of deaths (CRS).[7,8]

In part, new estimates are possible because of a number of sero-prevalence studies in India which, in conjunction with data from around the world on infection fatality rates (IFRs), allow estimation of excess deaths during Covid.

And finally, it owes to a new data source, the consumer pyramid household survey (CPHS) produced by the Center for the Monitoring of the Indian Economy (CMIE) which has been interviewing households all over India on a timely basis. Although the focus of the CPHS

---

[1] See https://www.mygov.in/covid-19, as compiled by the Johns Hopkins University Center for Systems Science and Engineering.

[2] https://www.ft.com/content/a2901ce8-5eb7-4633-b89c-cbdf5b386938

[3] http://www.healthdata.org/special-analysis/estimation-excess-mortality-due-covid-19-and-scalars-reported-covid-19-deaths and https://www.economist.com/graphic-detail/coronavirus-excess-deaths-tracker. Mukherjee et. al. (2020) also report model-based estimates of cases and deaths.

[4] https://www.nytimes.com/interactive/2021/05/25/world/asia/india-covid-death-estimates.html

[5] https://www.nytimes.com/2021/06/17/opinion/india-covid-ganges.html.

[6] A non-exhaustive list includes Rukmini S, Srinivasan Ramani, Anurabh Saikia, Dhanya Rajendran, Thejesh, Mariyam Alavi, Saurav Das, Chinmay Tumbe and scores of journalists working for India's English and vernacular newspapers.

[7] https://twitter.com/Rukmini/status/1407613966918901762

[8] https://www.researchgate.net/publication/352170008_Preliminary_Analysis_of_Excess_Mortality_in_India_during_the_Covid-19_Pandemic_Update_June_20

has been employment, income and consumption, a collateral benefit has been its data on mortality.

This paper has two objectives. First, to use these three sources to provide three new estimates of all-cause excess mortality for India, during the Covid pandemic for both the first and second waves. The second, which constitutes the bulk of this paper, is to elaborate on one of these estimates, namely those based on the CPHS.[9] We want to emphasize that we are not estimating Covid-caused deaths as CPHS has no information on cause of death. Rather, we focus on all-cause mortality, and estimate excess mortality from the onset of the pandemic relative to a pre-pandemic baseline, adjusting for seasonality.

The spirit of this paper is not to privilege any one estimate but simply to lay them out with transparency about data sources, assumptions, methodologies, and limitations. Given all the difficulties, getting at the true estimate will be difficult and only by piecing together data from different sources will we improve our understanding of the reality of the pandemic.

Table 1 summarizes the three data-based estimates of excess deaths.

Table 1. Comparing Alternative Estimates of All-Cause Excess Mortality (millions)

	Wave 1 (April 2020-March 2021)	*Wave 2 (April - June 2021)*	*Total*
1. States' Civil Registration Systems (CRS)	2 [0.1-2.3]	1.4 [1-2]	3.4 [1.1-4]
2. International age-specific infection fatality rates applied to Indian demography and seroprevalence	1.5	2.4	4.0
3. Consumer Pyramid Household Survey (CPHS)	3.4 [2-4.8]	1.5 [0.8-2.3]	4.9
Official	*0.16*	*0.24*	*0.4*

Notes: Strictly speaking, our second estimate is a Covid-caused one because it is based on Covid infections and Covid-related IFRs. The numbers in brackets for estimate 1 come from the alternative scenarios discussed in Section 2 below; and for estimate 3 come from the statistical confidence intervals of regression estimates discussed in Section 4.

We explain each of these three estimates below, including their shortcomings, with particular focus on the CPHS estimate. But three conclusions are evident. First, uncertainty afflicts all estimates, reflected in the wide range of each of them. Second, the three estimates are quite divergent in their assessment of the timing of Covid-related mortality (keeping in mind that the duration of the two waves is very different, 11 and 3 and a half months, respectively): the CRS-data suggest a roughly equal number of deaths in the two waves, the CPHS data implies that there were many more deaths in the first wave, while the sero-prevalence-cum-IFR procedure yields far greater deaths in the second wave.

[9] Independently of us, Malani and Ramachandran (2021) have also produced estimates based on the CPHS. As of writing, the paper is not available for circulation.

Third, and perhaps most important is that they all point to significantly greater deaths than the official estimates. Relatedly, it seems that the first wave was also more lethal than is widely believed. Estimating with confidence how many people died in Covid may prove elusive but at least now we have some data that will allow us to take a stab at that important question. What is tragically clear is that too many people, in the millions rather than hundreds of thousands, may have died.[10]

2. Estimates based on civil registration of deaths data

Official data on mortality are not collected and reported in a timely manner that would allow excess mortality to be estimated. India's Sample Registration System (SRS), managed by the central government, conducts annual mortality surveys but has only published numbers through 2019.

In the second wave, as the death toll rose, investigators forced state governments to release data on civil registration of deaths (CRS). As of June 29, 2021, excess deaths in 2020 and 2021 based on death registrations are shown in Table 2 for seven states.

The disaggregated data for the seven states (Andhra Pradesh, Bihar, Chhattisgar, Karnataka, Kerala, Madhya Pradesh, Tamil Nadu and Uttar Pradesh) underlying this table, which account for about half of India's total population, is presented in Appendix Table 1. There are a number of issues in going from the available data to an all-India estimate.

First, even for these seven states excess mortality may be mismeasured because CRS typically under-counts final estimated deaths in the SRS. Based on 2019 data, under-counting relative to SRS varies from zero for the Southern states to 37 percent for Uttar Pradesh and 48 percent for Bihar. NFHS-5 data for 2019-20 (but available for only 4 of the seven states) suggests that even the CRS is under-stating the magnitude of the problem: for example, NFHS indicates that the under-counting of deaths in Bihar is 67 not 48 percent and for Andhra Pradesh is 20 percent not zero.

Second, in terms of Covid mortality, it is possible that the rest of India is different from the seven states. Sero-prevalence could be different between more rural Bihar and dense, urban Maharashtra. Fatality rates could similarly be different between Kerala's well-run health system and Rajasthan's which is more challenged.

Third, the number for Uttar Pradesh in the second wave does not appear to pass the smell test. And finally, it must be emphasized that the CRS data for most states stops in May, so that they almost certainly fail to capture all of the second wave.

To cater to all these uncertainties, we present not one estimate but a range of estimates in which we make different assumptions about various parameters, always trying to err on the side of being conservative. In the baseline, we present numbers for all states, including UP, and we assume that the data for seven states, which account for half of India's population, are representative for the country. We also assume that the under-counting does not increase

[10] Recently, a potential fourth data source for estimating mortality has become available from the National Health Mission (Rukmini, 2021c). Its coverage is narrow (about 20 percent of total deaths in 2019) and the Covid mortality patterns are puzzling. More detailed analysis will be necessary before India-wide estimates can be derived from it.

because of Covid and the difficulty of recording deaths.[11] In other scenarios we exclude UP (Scenario 1), we assume heroically that under-counting magically disappears during the pandemic (Scenario 2), and we assume that NFHS-5 better measures under-counting than the CRS (Scenario 3).

In the baseline, we estimate excess deaths to be 2 million in the first wave and about 1.4 million in the part of the second wave that has been reflected in the data, yielding a total of about 3.5 million. Without UP, the total increases to about 4 million, mostly because UP's second wave appears implausibly low. And if NFHS-5 rather than CRS better captures under-counting excess deaths could be higher still than in the baseline. If we assume the CRS has suffered no undercounting during the pandemic, however unlikely that may be, our estimate would fall as low as 1 million total excess deaths across both waves.

Table 2. Excess deaths based on Civil Registration System (millions)

Description	Wave 1	Wave 2	Total
Baseline: 7 states are nationally representative in excess deaths and under-counting, and under-counting proportion remains constant pre-and post-Covid	1.98	1.39	3.37
Scenario 1. Baseline w/o Uttar Pradesh	2.11	1.97	4.08
Scenario 2. Baseline but assuming under-counting eliminated post-Covid	0.08	1.04	1.12
Scenario 3. Baseline but under-counting for four states based on NFHS-5 not CRS	2.25	1.58	3.83

Notes: The data underlying this table is in Appendix Table 1.

3. Estimates based on COVID-19 seroprevalence survey results and international estimates of the age-specific infection-fatality rate

Before we turn to our survey-based estimates of excess mortality, it is useful to step back and ask whether we can set some bounds for plausible excess deaths in India based on cross-country data.

Arguably the only reliable data on the true scale of the COVID-19 pandemic in India relate to infection rates that have come from sero-prevalence studies. Any number based on Indian deaths, such as case fatality rates or infection fatality rates, is questionable due to gaps in official mortality statistics. On the other hand, data for other countries, especially advanced ones, are more reliable both for infection rates (via sero-prevalence surveys) and for deaths.

[11] A simple example clarifies why under-counting is important for estimating the absolute number of excess deaths. Suppose pre-Covid deaths are 5 and post-Covid deaths rise to 7. Suppose that the under-counting proportion remains constant at say 50 percent. "True" absolute deaths are 10 and 14, respectively and "true" excess deaths are then 4 not 2.

So, the question is whether we can use cross-country data on age-specific infection fatality rates to estimate India's excess deaths.[12]

There have been many sero-prevalence studies done for different states and cities in India but two nationally representative ones are the third sero-survey done in December 2020-January 2021 and a recent WHO-AIIMS survey covering the period mid-March to early June. The former placed India's infection rate at about 22 percent. The latter placed India's infection rate at 57.7 percent, and 63.5 percent in the age group below 18 and above respectively.[13]

Since mortality estimates for India are less solid, so are IFR estimates. Here we have little choice but to rely on international estimates of IFRs. Recently, the United States' Center for Disease Control (CDC) disseminated its best estimates of age-specific IFRs which were based on the meta-analyses of Levin et. al. (2021). These age-specific IFRs can be combined with Indian demographic characteristics and the age-pattern of Indian infection rates to derive a plausible measure of IFR for India. The underlying assumption here is that the likelihood that any given infected person will die is the same across countries so that the international differences in aggregate IFRs are driven by the age structure of population and the age pattern of infections.[14]

Table 3 below illustrates the excess death calculations from this procedure. First, we have to address timing issues related to the sero-prevalance studies. Specifically, we need to get a measure of infections for the first wave (roughly mid-March) and for the second wave (end-June).

The mid-point of the third sero-prevalence study can be roughly dated at end-December 2020. So, we obtain infection for the first wave by extrapolating true infections by the evolution of actual infections. This yields an infection rate of about 25 percent in March (column 5 in Table 3).

To get an infection rate for the second wave, we need to date the mid-point of the WHO-AIIMS survey which was conducted over a wider time interval. To do so, we extrapolate the true infection rate based on the procedure described above and find the date when that yields the same number as the WHO-AIIMS survey. This occurs in early June. So, for end-June we once again extrapolate from this starting point again using the evolution of actual cases. This yields an infection rate of close to 66 percent at end-June (column 7 in Table 3).

Turning to infection fatality rates, we rely on the meta-analysis of Levin et. al. (2020), which is the basis for the latest best estimates recommended and used by the United States' Center for Disease Control (CDC). This yields a population IFR for the US of 0.56 percent. Applied

---

[12] Using aggregate IFRs from cross-country work would be less appropriate because of India's much younger demographic profile and because of the stark differences in fatality across ages.. Hence we build up from cross-country age-specific IFRs, the implicit assumption being that demography is the key driver of cross-country differences in aggregate IFRs.

[13] To our knowledge, both these surveys have only been reported on and have not been officially made public. Jha et. al. (2021) find much higher sero-prevalence in 12 large cities by end-2020.

[14] Age-specific IFRs can vary across countries. But this can go both ways: some genetic predisposition or exposure to BCG vaccinations could result in lower age-specific IFRs in India compared to other countries and indeed this was argued in the early stages of the pandemic. On the other hand, significantly more challenged health systems in India would increase India's age-specific IFRs (Sandefur et. al. 2020). A priori, it is not clear which dominates.

to India, it yields an IFR of 0.54 percent.[15] This difference stems from two factors: India's younger demography (implying a lower average IFR even when applying the same age-specific IFR), but also India's relatively flat relationship between age and seroprevalence compared to the US where infections are more concentrated in the young.[16]

Table 3. Deaths in India's first wave implied by combining international estimates of IFR with India's age structure and seroprevalence rates (millions)

Age	IFR	Popu-lation	Reported infection rate (28/12/20)	Implied infection rate (3/15/21)	Implied infection rate (4/6/21)	Implied infection rate (30/6/21)	Excess deaths (28/12/20)	Wave 1 excess deaths (15/3/21)	Wave 2 excess deaths (30/6/21)	Total excess deaths
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
10-59	0.08%	1,006	22.0%	24.6%	62.0%	65.7%	0.19	0.21	0.35	0.56
60-69	1.38%	87	23.4%	26.2%	63.5%	67.3%	0.28	0.31	0.49	0.81
70-79	4.62%	39	23.4%	26.2%	63.5%	67.3%	0.42	0.47	0.74	1.22
80+	15.46%	13	23.4%	26.2%	63.5%	67.3%	0.48	0.54	0.84	1.38
All	**0.54%**	**1,145**	**22.2%**	**24.8%**	**62.2%**	**65.9%**	**1.37**	**1.53**	**2.43**	**3.97**

Notes: This table is aggregated up from more disaggregated 5-year age categories. In the analysis, the under-10 age group is excluded because India's sero-prevalence surveys do not provide data on them. Age-specific IFRs in column 2 are from Levin et. al. (2020), Supplementary Appendix Q; Population and deaths are in millions; Infection rates at 28/12/2020 in column 4 are from the third national sero-prevalence survey; Infection rates in mid-March (end of the second wave) in column 5 are extrapolated by the ratio of reported cases between December and March; Infection rates in column 6 are from the WHO-AIIMS survey, whose midpoint was 4/6/21, assuming that actual infection rates between this survey and the third sero-survey grew in line with reported cases; Deaths in columns 8-10 are obtained by multiplying the relevant infection rate by the IFR (column 2) and population (column 3).

Combining the Indian sero-prevalence data and applying the best estimate of international age-specific IFRs to Indian demography and sero-prevalence patterns, yields excess deaths of 1.5 million and 2.4 million in the two waves, respectively. Together, the toll of Covid until June 2021 is estimated at 4 million. Even though the second wave encompasses three months compared to eleven months for the first, the estimated death toll is significantly greater. This follows simply from the fact that the second wave in our calculations witnessed more infections (about 41 percent) compared to the first wave (25 percent).[17]

---

[15] For comparison, Banaji (2021b), based on early data from Mumbai and a global meta-analysis, suggests an IFR for India between 0.25 and 0.5 percent. See *https://www.theindiaforum.in/article/estimating-covid-19-fatalities-india* and *https://www.medrxiv.org/content/10.1101/2021.04.08.21255101v1*. We are grateful to Murad Banaji who suggested the use of the more age-disaggregated data in Levin et. al (2021) over the more aggregated suggestions of the CDC. Because of the sharp increase in mortality for the over-80 age group and because India's share of this group is much lower, aggregate IFRs are sensitive to aggregation.

[16] The difference in the age-seroprevalence gradient between India and the U.S. can be seen by comparing the results of India's 3rd National Seroprevalence Survey (in column 3 of Table 3) to the U.S. figures from Bajema et al (2020).

[17] It is possible that the age-specific IFRs are lower for Indian than assumed here, leading to an overestimation of deaths in both waves. But it is almost certain that the second wave estimates have been under-estimated relative to the first wave because of the assumption of constant age-specific IFRs across the two waves. The sudden

4. Estimates based on the Consumer Pyramids Household Survey Data

4.1 Data description

To our knowledge, the only survey data collecting death information for all of India for the entire pandemic period is the Consumer Pyramids Household Survey (CPHS) conducted by the private firm Centre for Monitoring Indian Economy (CMIE). The CPHS constitutes a longitudinal panel of individuals, with information spanning roughly 868,000 individuals across roughly 177,000 households interviewed once every four months, with about one-fourth of the sample interviewed each calendar month.

Since September 2014, the CPHS has recorded whether any member of the family has died in the four months covered by the survey. No information on the cause of death is recorded, nor on the precise timing of deaths between rounds. The spacing of the survey rounds and the lack of a birth roster make the data poorly suited for estimating infant mortality, but better suited for adult mortality which is of particular relevance given the known age-profile of COVID-19 fatality rates.

So, in principle, the CPHS provides a source -- and perhaps the only survey data source -- for estimating excess death estimates for India as a whole. To date there have been 22 rounds of the CPHS's People Survey, with the most recent round covering January to April 2021. The 23rd round of this survey is underway and the data will be released later this year, covering the four months of May to August. Although data from this survey is released every four months, the CPHS's Income Survey releases monthly data based on interviews being conducted in this 23rd Round. Data for surveys done in May and June 2021 have been released in the income survey.

Thus, the CPHS can be used to capture mortality in the first and some of the second wave;[18] not all of the second wave will be captured because of the lag between cases and deaths and because households are surveyed over the course of the month and are asked about deaths in the previous four months. Surveys done in July and August 2021 are therefore likely to capture additional mortality from the second wave.

There is reason for caution when relying on the CPHS for mortality estimates though. While CPHS has become a critical source of timely information on labour market and consumption trends, especially in the absence of timely and reliable official data, its

increase in infections and the overwhelming of domestic health systems, including the shortages of oxygen, hospital beds and ventilators, must have raised age-specific IFRs substantially. New York, for example, experienced an increase in the overall IFR, from 0.5 percent (trough) to about 1.75 percent (peak) when cases surged in March-April 2020 and overwhelmed the health system (Yang, et. al. 2020). Banaji (2021c) calculates that for Andhra Pradesh, IFRs may have been higher by about 0.2 percent in the second wave compared to the first. If this is true more broadly for India, our second wave deaths could be under-estimated by about 35 percent (https://maths.mdx.ac.uk/research/modelling-the-covid-19-pandemic/andhrapradeshifr/). In the second wave, some of the IFR surge might have been moderated by vaccinations, especially of the older age groups: for example, by July 2, about 90 million doses were administered to above-60 year-olds, comprising 140 million.

[18] Malani and Ramachandran (2021) alerted us to the fact that the CPHS income survey could be merged to the earlier employment rounds to extract mortality for May and June 2021.

representativeness has recently been questioned.[19] Dreze and Somanchi (2021) report that the CPHS sample appears more affluent than the government's nationally representative sample for the National Family Health Survey. All our estimates use sampling weights provided with the CPHS.[20] Using these weights, the CPHS matches the rural-urban breakdown of India from the census quite closely: 65% rural in 2020, compared to 69% in the 2011 census. To gauge representativeness on a socio-economic dimension, we compare the share of adults aged 18 or above who report no schooling in CPHS: as of 2015 this figure stood at 29%, compared to 28% in the NFHS 2015-16.

Table 4. Summary statistics for the Consumer Pyramids Household Survey

	2015-2018	2019	2020-21	Benchmark
Observations	9,159,043	2,547,524	3,871,040	n.a.
Households	206,638	174,405	177,541	n.a.
Individuals	1,036,623	860,117	884,605	n.a.
By age (% , weighted)				UN (2020)
[0-60)	91.7	91.3	91.3	91.1
[60-70)	5.8	5.8	6.0	5.5
[70-80)	1.9	2.2	2.1	2.6
80 and above	0.5	0.7	0.6	0.8
By rural/urban (%, weighted)				UN (2020)
Rural	68.5	68.3	68.4	65.1
Urban	31.5	31.7	31.6	34.9
By male/female (%, weighted)				UN (2020)
Male	52.9	53.0	53.5	52.0
Female	47.1	47.0	46.5	48.0
By education level (%, weighted				NFHS (2015-16)
Above median	66.4	63.9	63.7	51.7
Below median	33.6	36.1	36.3	48.3
Raw no. of deaths in sample				

[19] 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

[20] Note deaths in the CPHS are recorded, logically, in the first survey wave after a death occurs, and deceased individuals receive no sampling weight. We rely on lagged weights from the previous survey wave for the deceased.

By age				
[0-60)	8,838	2,476	3,064	n.a.
[60-70)	4,031	1,448	2,202	n.a.
[70-80)	3,215	1,434	2,132	n.a.
80 and above	1,943	959	1,296	n.a.
By rural/urban				
Rural	6,004	2,354	4,679	n.a.
Urban	13,575	4,534	9,372	n.a.
By male/female				
Male	7,619	2,944	3,172	n.a.
Female	5,627	2,433	2,698	n.a.
Average Response rate	84.13	82.53	59.53	n.a.

Nevertheless, the sampling process for the CPHS panel is not elaborately documented, and further examination of its representativeness is surely justified in future work. One possibility is that the stark discrepancies between the CPHS and NFHS that Dreze and Somanch (2021) document relate to how questions are asked, rather than who is sampled; though our results using education levels are far from establishing this definitively. Vyas (2021) provides further explanation of these discrepancies.

Apart from sampling concerns, it is important to note that the CPHS was not originally intended to produce vital statistics. The survey question relating to deaths was included for the purpose of maintaining the household roster up to date in the longitudinal panel, and prior to the pandemic had not received close scrutiny. The deficits of the CPHS questionnaire design must be weighed against one of its key advantages, which is its large sample size and time-series coverage. Prior to the COVID pandemic, the survey records roughly 4,000 to 5,000 deaths per year, roughly a third of which occur among individuals 65 years or older.

A third concern with the CPHS, particularly in the COVID pandemic period, is the response rate. Prior to the pandemic, CPHS maintained a response rate above 80%, but this fell to just 60% in 2020-21. Looking wave by wave, this decline was disproportionately driven by a sharp drop in responses in the months during and immediately after the 2020 lockdown, reaching a low of 43.8%.

Non-response becomes problematic if it is non-random: specifically, if households with higher (or lower) probability of mortality among members are less likely to respond. This seems quite likely given the nature of disruptions, migration, etc. during mid 2020. There is some consolation here though due to the panel nature of the survey and the phenomenon of interest here: death. Individuals or households who fail to respond in one wave may return to the sample in subsequent waves. Any deaths which were missed will be recorded, albeit late. In total, we find that 88% of respondents answered the CPHS questionnaire at some point after the 2020 lockdown. Thus to some extent, the drop in response rate should impact the timing of deaths we observe as much or more than the total rate.

4.2 Validating (or not) CPHS death estimates?

What numbers does the CPHS data yield? Before we can answer that, we need to check for the plausibility of the CPHS estimates for the pre-Covid years. We can compare the death estimates from CPHS with two other sources: India's own SRS and the UN's estimates in their World Population projections.[21] In turn we can compare the estimates at the aggregate level, by age, by gender, and at state level. This is what we do in the next few charts.

What can we infer from all these charts?

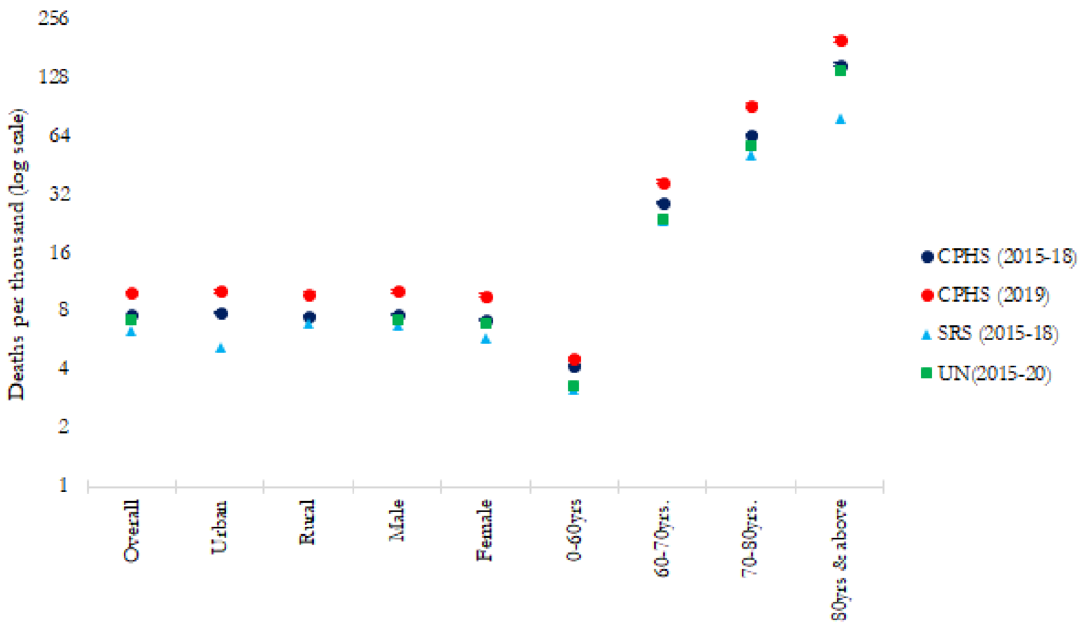
First, that CPHS overall death estimates are consistently greater than those of the UN, SRS, and also with the Census of 2011. The discrepancy is relatively small until 2018 but widens dramatically in 2019. Estimates for 2019 really stand out as an outlier. Surprisingly, the discrepancy is much greater for the urban estimates than rural ones.

Second, since Covid's mortality is skewed toward older age groups and non-linearly, validating the age pattern of mortality against other data sources is critical. Figure 2 shows that CPHS mortality estimates for the period 2015-18 are substantially greater than estimates by the SRS and UN for younger age groups (below 40); but for the older age groups, CPHS tracks SRS and UN very closely. But this changes in 2019 when CPHS diverges very sharply especially for groups above 60 (the log scale in Figures 1 and 2 masks the magnitude of discrepancy).

Third, at state level, the CPHS estimates are much more variable than those of the SRS: in Figure 2, for example, the spread is greater along the y-axis than the x-axis. Consistent with Figure 1, urban estimates are greater in the CPHS (blue symbols) than SRS while the rural estimates (green symbols) are less biased.

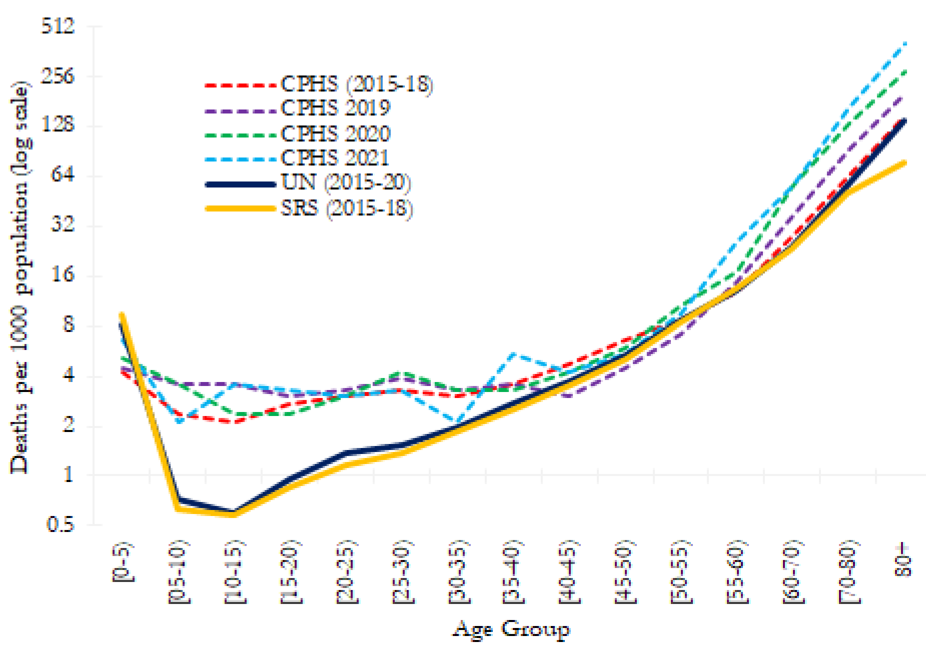
[21] The 2011 Census even though it is relatively outdated is a useful benchmark because of all the estimates it is the most reliable—it is a census not a survey; moreover, since mortality rates don't bounce around much and exhibit relatively stable trends (if at all), they serve as a useful check.

Figure 1. Comparing CPHS Death Estimates with Other Sources by Geography, Gender, and Age, pre-Covid



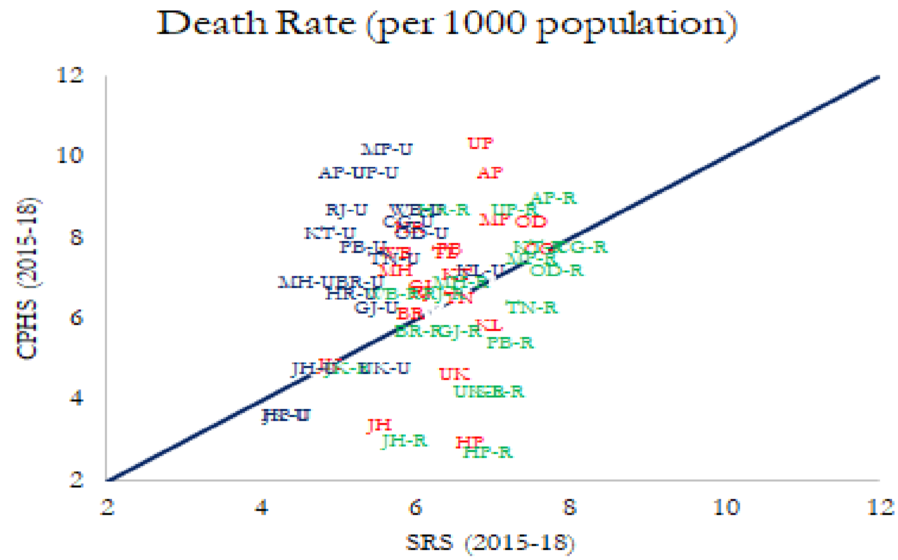
Note: CPHS estimates based on authors' calculations. SRS estimates from https://censusindia.gov.in/2011-Common/Sample_Registration_System.html. UN estimates from https://population.un.org/wpp/Download/Standard/Mortality/.

Figure 2. Comparing CPHS Death Estimates with Other Sources By Disaggregated Age Categories, pre-Covid



Note: CPHS estimates based on authors' calculations. UN estimates from https://population.un.org/wpp/Download/Standard/Mortality/.

Figure 3. Comparing CPHS Death Estimates with Other Sources by State, pre-Covid



Note: Each point represents average mortality per 1,000 inhabitants, averaged over three years. State averages are shown in red, averages for urban areas in blue with suffix -U, and for rural areas in green with suffix -R.

The 2019 CPHS numbers pose a real dilemma. If they are indeed unreliable, did the inherent errors continue in the post-Covid period and if so in which direction and to what extent?

To understand the data--and indeed the predicament--better, we plot the monthly death estimates from the CPHS by age group, focussing on the older cohorts affected disproportionately by Covid (Figures 4 and 5, the latter showing a log scale)).

Figure 4. Monthly Mortality CPHS, May 2015-June 2021

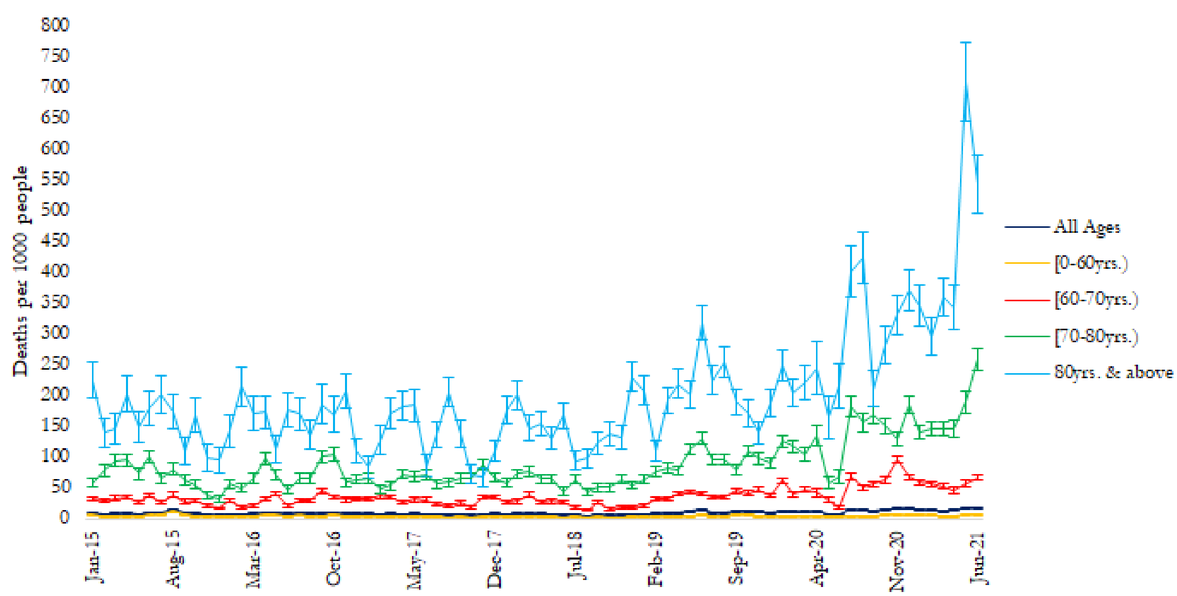
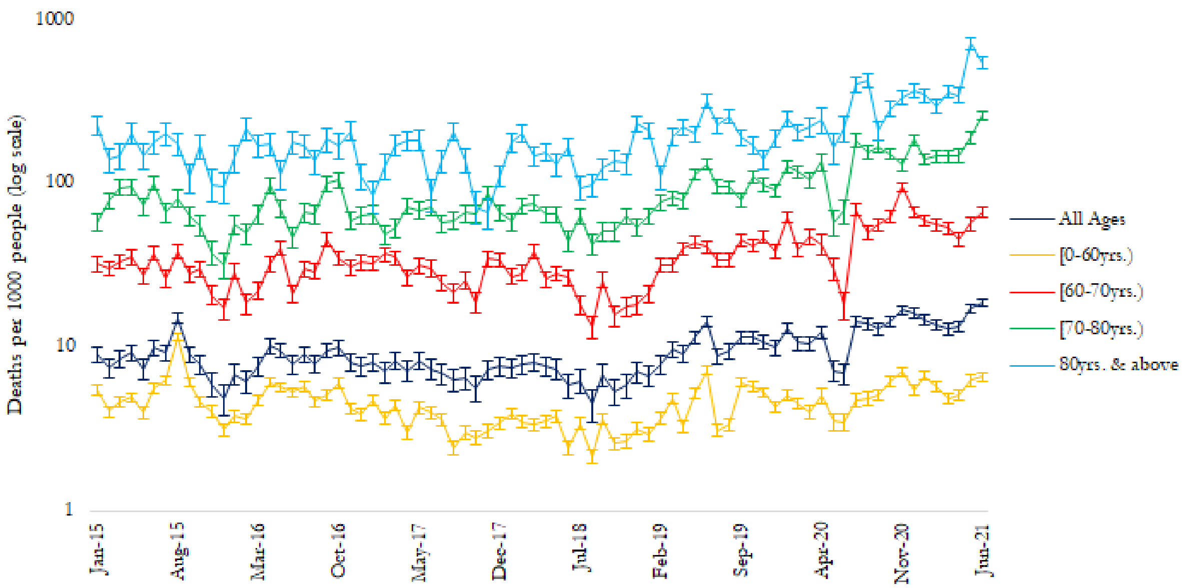


Figure 5. Monthly Mortality CPHS, May 2015-June 2021 (log scale)



CPHS monthly estimates for mortality, especially for older cohorts, exhibit a fair amount of temporal volatility. The short vertical lines for each of the four mortality trends represent 95 percent confidence intervals. These are quite tight for all groups except for the oldest where these intervals are quite wide, suggesting greater variability in the mortality estimates. That Covid has impacted older cohorts differentially is also evident in the figures. The normal scaling in Figure 4 obscures some important trends in the data. Figure 5 (shown on a log scale) suggests a slow, trend decline in mortality in the below 60 age group since 2016 and then a sharp increase in aggregate mortality until mid-2019.

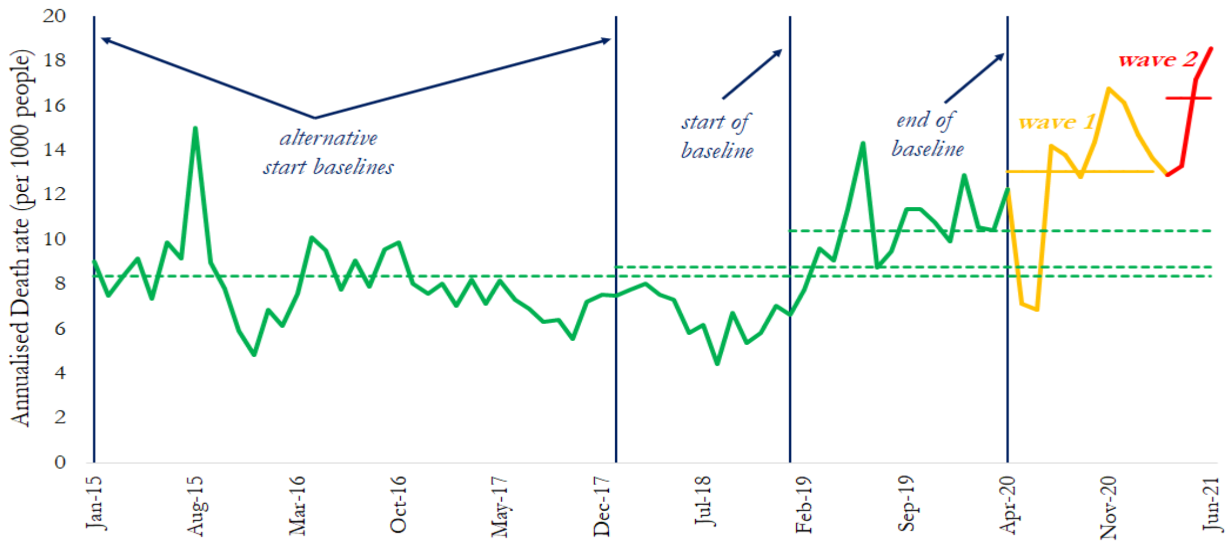
There is also a sharp drop in mortality the month of the lock-down and rising mortality thereafter. It seems from the figures that peak mortality in the first wave was reached in the last quarter of 2020 which appeared to subside through March-April of 2021 and then spiked sharply in the second wave in May and June 1991.

This raw data raises a number of questions. Why did mortality increase in the middle of 2019 well before the pandemic across all age groups? Is this partly just some correction for the decline in mortality before mid-2019? Are these ups and downs plausible? Is the sharp decline in the lockdown phase of the first wave (March-April 2020) real or could it be due to the decline in the response rate?

4.3 Estimating excess mortality

Given these deficiencies, can we nevertheless estimate or guesstimate the Covid-period excess mortality by comparing the death estimates in 2020 and 2021 with a reference period? But what should that reference period be? Had the 2019 estimates been similar to those between 2015 and 2018, that choice would not have been tricky. But the 2019 estimates which suddenly jump makes the comparison exercise fraught. If the 2019 estimates are off, could the 2020 and 2021 estimates be even more off than 2019, and if so in which direction (i.e. could the bias be increasing or decreasing in unpredictable ways?).

Fig 6. Alternative Baselines for Excess Mortality Estimates



Based on Figure 6, we instead chose three alternative reference periods: January 2019-April 2020, January 2018-April 2020 and January 2015-April 2020. Essentially, we treat the post-April 2020 period as the one impacted by Covid, in turn divided into Wave 1 (May 2020-March 2021) and Wave 2 (April - June 2021). The first reference period of 2019 would be most appropriate if the methodology (and the errors) of 2019 carried over to the post-Covid period. The other two reference periods would be more appropriate if 2019 was the aberration in the CPHS.

We measure all-cause excess mortality as the difference between the death rate from May 2020 to June 2021 compared to each of these three reference periods. So, for example, excess mortality in wave 1 would be the difference between the horizontal yellow line and each of the dashed green lines. Since some of the reference periods include partial years, we control for seasonality with month effects. All calculations use survey weights. We implement this calculation in a simple regression framework, using the specification detailed in the appendix.[22]

[22] Note that the exact same total excess mortality figures can be achieved by calculating total deaths in each calendar month of the pandemic, subtracting average deaths in the same calendar month in the baseline period, and summing those differences. The regression approach simply provides a parsimonious way of doing this calculation and estimating its standard errors.

Table 5. India, nationwide excess mortality from CPHS (annualised, deaths per thousand)

	Excess mortality relative to benchmark of...			
		Jan15- Apr20	Jan18-Apr20	Jan19-Apr20
Overall (1430 mn)	Baseline mortality	8.33	8.67	10.52
		[0.12]	[0.18]	[0.29]
	Excess mortality (wave 1)	4.76	4.4	2.46
		[0.44]	[0.45]	[0.51]
	Excess mortality (wave 2)	7.48	6.93	4.47
		[0.91]	[0.94]	[1.12]
Urban (454 mn)	Baseline mortality	8.59	8.94	10.4
		[0.21]	[0.34]	[0.53]
	Excess mortality (wave 1)	4.66	4.39	3.00
		[0.88]	[0.88]	[0.95]
	Excess mortality (wave 2)	9.96	9.29	7.26
		[1.72]	[1.74]	[1.99]
60yrs.-70yrs (79 mn)	Baseline mortality	30.92	32.25	39.92
		[0.68]	[1.07]	[1.61]
	Excess mortality (wave 1)	25.16	23.88	16.08
		[2.87]	[2.91]	[3.12]
	Excess mortality (wave 2)	23.72	21.67	14.76
		[5.01]	[5.39]	[6.07]
70yrs.-80yrs (37 mn)	Baseline mortality	73.76	80.14	98.84
		[1.77]	[2.91]	[4.43]
	Excess mortality (wave 1)	66.04	58.85	39.54
		[7.34]	[7.44]	[8.14]
	Excess mortality (wave 2)	117.67	112.2	85.83
		[15.33]	[15.77]	[17.12]

Above 80yrs. (12 mn)	Baseline mortality	165.82	180.85	208.29
		[4.95]	[7.43]	[9.99]
	Excess mortality (wave 1)	145.4	129.38	102.3
		[21.56]	[22.76]	[24.43]
	Excess mortality (wave 2)	349.21	329.3	283.66
		[49.96]	[50.53]	[53.97]
Above Median Educated HHs (911 mn)	Baseline mortality	8.5	8.94	10.87
		[0.19]	[0.30]	[0.45]
	Excess mortality (wave 1)	4.96	4.53	2.61
		[0.67]	[0.69]	[0.78]
	Excess mortality (wave 2)	5.00	4.10	0.81
		[1.19]	[1.27]	[1.57]

Source: Authors' calculations based on CPHS micro data. Estimates are based on a unit-record level regression of an indicator for death on calendar month indicators and their interaction with indicators for 2020 and 2021, respectively. Numbers in brackets are standard errors, clustered at the level of the primary sampling unit. For presentational simplicity, we do not present all sub-categories: for example, we present results for the urban sector, omitting the rural sector; for older age cohorts, excluding the 0-60 year cohorts; and above median-educated households, excluding below-median educated households. Coefficients for these excluded categories can be inferred by reference to the overall coefficients reported in the first row.

The results are presented in Table 5, with standard errors in brackets. The most cautious/conservative excess mortality estimates are those in column 3 where the reference period is January 2019 to April 2020. The annualized excess mortality rate of 2.46 and 4.47 in the two waves, (which is statistically significant at the 1% level) implies that mortality jumped by about 23 percent (2.46/10.52) and 42 percent (4.47/10.52) in the two waves, respectively. This excess mortality is also close to the average for developing countries as a whole.

Row 2-4 present results for the above 60-year age groups. Mortality is significantly greater during the Covid pandemic for these age groups. For example, for ages above 80, excess mortality in the second wave was 284 out of 1000 or 136 percent greater than in normal times. Given possible concerns that the age structure of the CPHS sample could be unrepresentative especially at older age cohorts, we estimated the coefficients weighting the sample according to the UN's age structure of India's population. The results were close to those shown in the first row of Table 5.[23] Row 5 indicates that Covid did not differentially impact educated compared to uneducated households: in the first wave, they were impacted more but less in the second wave.

[23] Results available from the authors on request.

The numbers on the left hand side of Table 5 are the populations of the relevant samples which allow the per capita estimates to be translated into total deaths shown in Table 6. For example, in the first wave, an excess mortality coefficient of 2.46 (row 1, column 3 of Table 5) translates into total mean deaths of 3.4 million (95% confidence interval of 2.0 to 4.8 million). Dissimilar coefficients in Table 5 (for example, urban versus rural) may nevertheless translate into more equal impacts in terms of absolute deaths because of different shares of the population.[24]

Table 6. Total Excess Deaths in Millions

(Derived from Estimates in Column 3 of Table 5)

		Mean	Confidence Interval	
Total	Wave 1	3.4	2.0	4.8
	Wave 2	1.5	0.8	2.3
Urban	Wave 1	1.3	0.5	2.1
	Wave 2	0.8	0.4	1.2
60-69 yrs	Wave 1	1.3	0.8	1.8
	Wave 2	0.3	0.1	0.5
70-79 yrs	Wave 1	1.1	0.6	1.5
	Wave 2	0.6	0.4	0.8
80 plus	Wave 1	0.7	0.4	1.0
	Wave 2	0.5	0.3	0.7
Above-Median Educated HHs	Wave 1	2.3	0.9	3.7
	Wave 2	0.2	-0.5	0.9

Figures represent the estimated number of excess deaths for all of India in millions, from May 2020 through June 2021 (divided into two waves as described above) relative to January 2019 through April 2020. Estimates are based on the regressions using CPHS data reported in Column 3 of Table 2, and in greater detail in Appendix Table 4.

[24] Our estimates for excess deaths are somewhat greater than those estimated by Malani and Ramachandran (2021); their reference periods and dating of the pandemic are a little different from ours.

5. Conclusions

Of another existential threat, Bob Dylan accusingly asked, "How many deaths will it take till we know that too many people have died?" This Covid pandemic in India has seen a curious but no less tragic inversion of that sentiment: only a sense that too many died in the second wave has really galvanized the effort to find out the true number of deaths.

In this paper, we have presented three different data-based estimates for all-cause excess mortality for India. Each of these estimates has shortcomings. The CRS-based numbers are still not available for all the states; they are known to under-count deaths and whether that under-counting remained the same or changed during the pandemic is still not clear; and the CRS numbers also stop for most states in May 2021.

The second estimate is only partly based on Indian data (on infection rates) and assumes IFRs based on other countries whose validity for India is open to question; and we cannot adjust the IFRs for wave 2 when mortality is likely to have increased as domestic health systems were overwhelmed.

The important caveat on the death estimates from the CPHS is that its pre-Covid mortality does not track closely estimates from other official sources. Perhaps even more important is that the CPHS shows a big and inexplicable spike in mortality in 2019 before Covid. If some of the measurement errors from the CPHS pre-Covid carry over to the Covid period, the reliability of the excess deaths estimates is not assured.

The time pattern of the excess death estimates is also different across these three estimates.

Even with all these caveats, there are three important take-aways from our findings.

First, unsurprisingly, there is considerable uncertainty within and across estimates. They range from about 1 million to 6 million overall, with central estimates varying between 3.4 to 4.9 million. And the timing on the estimates between the two waves also varies across estimates. It is imperative, therefore, that research continue to estimate more precisely Covid-related deaths. It is equally imperative that government aid this effort by making available all the data on sero surveys and deaths it has generated.

Second, the first wave seems to have been more lethal than is popularly believed. Because it was spread out in time and space, unlike the sudden and concentrated surge of the second wave, mortality in the first wave appeared moderate. But even the CRS data suggest that up to 2 million might have died in that period. In fact, not grasping the scale of the tragedy in real time in the first wave may have bred the collective complacency that led to the horrors of the second wave.

Finally, and perhaps the most critical take-away is that regardless of source and estimate, actual deaths during the Covid pandemic are likely to have been an order of magnitude greater than the official count. True deaths are likely to be in the several millions not hundreds of thousands, making this arguably India's worst human tragedy since partition and independence.

Our purpose in this paper has been to report--without in any way endorsing--different data-based estimates. A collective understanding of and engagement with them and indeed with estimates from other data sources, warts and all, is now necessary.

References

Arora, R., A. Joseph, J. Van Wyk, S.Rocco, A. Atmaja, E.May, T. Yan, N. Bobrovitz, J. Chevrier, M. Cheng, T. Williamson, and D. Buckeridgeh, 2021, "SeroTracker: a global SARS-CoV-2 seroprevalence dashboard," https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7402646/

Bajema, K, R. Wiegand, K. Cuffe, et al., 2020, "Estimated SARS-CoV-2 Seroprevalence in the US as of September 2020", *JAMA Internal Medicine.* doi:10.1001/jamainternmed.2020.7976

Banaji, Murad, 2021a, "Estimating Covid-19 Fatalities in India", *The India Forum*, https://www.theindiaforum.in/article/estimating-covid-19-fatalities-india

Banaji, Murad, 2021b, "Estimating COVID-19 infection fatality rate in Mumbai during 2020," *https://www.medrxiv.org/content/10.1101/2021.04.08.21255101v1*

Banaji, Murad, 2021c, "Point estimates of COVID-19 infection fatality rate in Andhra Pradesh," https://maths.mdx.ac.uk/research/modelling-the-covid-19-pandemic/andhrapradeshifr/

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

Gupta, A, and N. Sudharsanan, 2020, "Large and persistent life expectancydisparitiesamongIndia's social groups," https://doi.org/10.31235/osf.io/hu8t9

Jha, P, Arokiaswamy Velumani, Chaitali Nikam, Wilson Suraweera, Sze Hang Fu, Hellen Gelband, Patrick Brown, Isaac Bogoch, and Nico Nagelkerke, 2021, "SARS-CoV-2 Seroprevalence in 12 Cities of India from July-December 2020," https://www.medrxiv.org/content/10.1101/2021.03.19.21253429v1.full.pdf

Karlinsky, A and D. Kobak, 2021, "The World Mortality Dataset: Tracking excess mortality across countries during the COVID-19 pandemic," https://pubmed.ncbi.nlm.nih.gov/33532789/

Leffler, C., 2021, "Preliminary Analysis of Excess Mortality in India during the Covid-19 Pandemic," https://www.researchgate.net/publication/352170008_Preliminary_Analysis_of_Excess_Mortality_in_India_during_the_Covid-19_Pandemic_Update_June_20

Levin, A.T., W. P. Hanage, N. Owusu-Boaitey, K. B. Cochran, S. P. Walsh and G. Meyerowitz-Katz, 2020, "Assessing the age specificity of infection fatality rates for COVID-19: systematic review, meta-analysis, and public policy implications," *European Journal of Epidemiology,* volume 35, pp. 1123–1138, https://link.springer.com/article/10.1007/s10654-020-00698-1

Meyerowitz-Katz, G., and L. Merone, 2020, "A systematic review and meta-analysis of published research data on COVID-19 infection fatality rates," *International Journal of Infectious Diseases,* Volume 101, December 2020, Pages 138-148, https://reader.elsevier.com/reader/sd/pii/S1201971220321809?token=ADA43F95496 23DC544EEEB513667BCA7A6B018A656FC7CA19ABD63BB5F09F90A42124A905F

BD56CC23DB592E6821CB0E&originRegion=us-east-1&originCreation=20210705000910
Malani, A, and S. Ramachandran, 2021," All-cause mortality during SARS-CoV-2 Pandemic in India," (draft). Presented at CMIE Seminar, July 8, 2021.
Malani, A, S. Ramachandran, V. Tandel, R. Parasa, S. Sudharshini, V. Prakash, Y. Yogananth, S. Raju, T.S. Selvavinayagam, 2021, "SARS-CoV-2 Seroprevalence in Tamil Nadu in October-November 2020," https://www.medrxiv.org/content/10.1101/2021.02.03.21250949v1
Mukherjee, B., R. Bhattacharyya, R. Bhaduri, R. Kundu, and M. Salvatore, 2020, "Reconciling epidemiological models with misclassified case-counts for SARS-CoV-2 with seroprevalence surveys: A case study in Delhi, India," https://bhramarm.medium.com/what-can-we-learn-from-the-delhi-seroprevalence-study-and-a-single-number-that-was-reported-45e4cf185357
O'Driscoll, M., G. Ribeiro Dos Santos, L. Wang, D. A. T. Cummings, A. S. Azman, J. Paireau, A. Fontane, S. Cauchemez, and H. Salje, 2021, "Age-specific mortality and immunity patterns of SARS-CoV-2," *Nature*, Vol. 590, https://www.nature.com/articles/s41586-020-2918-0.pdf
Rukmini S, 2021a, "The Challenge Of Saying How Many Excess Deaths Could Be Due To Covid-19, *Indiaspend*, " June 25, 2021 https://www.indiaspend.com/covid-19/deaths-first-wave-second-wave-pandemic-757701
Rukmini S, 2021b, " Gauging pandemic mortality with civil registration data," https://www.thehindu.com/opinion/op-ed/gauging-pandemic-mortality-with-civil-registration-data/article35157185.ece
Rukmini S, 2021c, " Deaths By 'Unknown Causes' On National Health Mission Portal 2X Official Covid Toll" https://www.indiaspend.com/covid-19/deaths-unknown-causes-national-health-mission-portal-covid-toll-760219
Tumbe, C, 2021, "Why 'excess mortality' figures for Covid must be calculated," *Indian Express*, May 26, 2021, https://indianexpress.com/article/opinion/columns/why-excess-mortality-figures-for-covid-must-be-calculated-7330348/
Vyas, M, 2021, "There are practical limitations in CMIE's CPHS sampling, but no bias," *Economic Times,* June 23, 2021 https://economictimes.indiatimes.com/opinion/et-commentary/view-there-are-practical-limitations-in-cmies-cphs-sampling-but-no-bias/articleshow/83788605.cms
Yang, W., S. Kandula, M. Huynh, S. K. Greene, G. Van Wye, W. Li, H. T. Chan, E. McGibbon,
A. Yeung, D. Olson, A. Fine, and J. Shaman, 2021, "Estimating the infection-fatality risk of SARS-CoV-2 in New York City during the spring 2020 pandemic wave: a model-based analysis," *Lancet Infectious Diseases,* 21: 203–12 https://www.thelancet.com/action/showPdf?pii=S1473-3099%2820%2930769-6

Appendix

CRS estimates

Appendix Table 1 provides the state-wise disaggregated data that is used to estimate the excess deaths based on the CRS shown in Tables 1 and 2 of the text.

Appendix Table 1. Deaths and Excess Deaths in the CRS by States (in millions)

State	*Population*	*Pre-Covid deaths (Wave 1)*		*Pre-Covid deaths (Wave 2)*		*Coverage rate pre-COVID (2019)*		*All-cause deaths*		*Excess deaths (CRS under-reporting unchanged)*		*Excess deaths (NFHS-5 under-reporting unchanged)*		*Excess deaths (under-reporting stops)*	
		SRS	*CRS*	*SRS*	*CRS*	*CRS vs SRS*	*NFHS*	*Wave 1*	*Wave 2*	*Wave 1*	*Wave 2*	*Wave 1*	*Wave 2*	*Wave 1*	*Wave 2*
Andhra Pradesh	52.3	0.32	0.32	0.08	0.08	100%	80%	0.43	0.22	0.11	0.14	0.13	0.17	0.11	0.14
Bihar	120.1	0.51	0.27	0.07	0.04	52%	37%	0.39	0.10	0.23	0.13	0.32	0.18	-0.13	0.03
Chhattisgarh	28.9	0.02	0.02	0.00	0.00	82%	na	0.09	0.07	0.08	0.08	0.08	0.08	0.07	0.07
Karnataka	66.0	0.43	0.43	0.07	0.07	100%	87%	0.55	0.12	0.13	0.06	0.15	0.06	0.13	0.06
Kerala	35.2	0.23	0.23	0.06	0.06	100%	97%	0.24	0.05	0.01	0.00	0.01	0.00	0.01	0.00
Madhya Pradesh	82.6	0.45	0.40	0.07	0.06	89%	na	0.45	0.23	0.05	0.20	0.05	0.20	-0.01	0.17
Tamil Nadu	75.8	0.52	0.52	0.09	0.09	100%	na	0.62	0.15	0.10	0.06	0.10	0.06	0.10	0.06
Uttar Pradesh (UP)	226.0	1.10	0.70	0.06	0.04	63%	na	0.87	0.06	0.28	0.04	0.28	0.04	-0.23	0.00
Total	**686.8**	**3.58**	**2.88**	**0.50**	**0.43**	**80%**	**na**	**3.63**	**1.01**	**0.99**	**0.69**	**1.12**	**0.79**	**0.04**	**0.52**
Total (w/o UP)	**460.9**	**2.48**	**1.95**	**0.44**	**0.34**	**79%**	**na**	**2.51**	**0.90**	**0.71**	**0.66**	**0.84**	**0.75**	**0.27**	**0.52**

Notes: CRS data are from the Development Data Lab (*http://www.devdatalab.org/covid*). The column "Excess deaths (CRS under-reporting unchanged) corresponds to our baseline estimate reported in Table 2 in the text. This assumes that the all-cause deaths calculated based on states' data is representative for all-India. The estimate in this column excluding UP corresponds to scenario 1 reported in table 2 in the text. The column "Excess deaths (NFHS under-reporting unchanged)" corresponds to our Scenario 3 reported in Table 2 in the text. The column "Excess deaths (NFHS under-reporting stops)" corresponds to our Scenario 2 reported in Table 2 in the text.

CPHS estimates

Using the full panel of individuals in the CPHS data, we regress an indicator variable for whether individual *i* was reported dead for the first time during this survey round, $died_{it}$, on indicators for the calendar month (with coefficients α_m), and the interaction of these month dummies with an indicator that takes the value of one from May 2020 onward (yielding twelve additional coefficients β_m).

$$died_{it} = \sum_{m=1}^{12} I[month = m](\alpha_m + \beta_m I[t \geq May2020]) + \varepsilon_{it}$$

All regressions use survey weights. We repeat the regression restricting the sample in two different ways: (a) restricting the range of pre-pandemic dates that we use as a baseline for excess mortality, and (b) restricting the sample to only urban, rural, young, or old individuals, yielding 15 separate estimates. These regressions are presented in Appendix Tables 2-4 below. For simplicity, for each of these estimates, Table 1 in the text reports total excess mortality expressed in deaths per 1,000 population, calculated as follows:

$$Total\ excess\ mortality, May\ 2020\ to\ Apr\ 2021, per\ thousand\ population = \sum_{m=1}^{12} \beta_m * 1000/4$$

Division by 4 is necessary as households are interviewed only once every four months, but we include monthly dummies. (Interviews are staggered so some portion of households are interviewed each month.). All calculations are done in Stata, version 15.1.

The CPHS data is available with subscription from the CMIE website. All code necessary to replicate our results here is available at https://cgdev.org/sites/default/files/anand-et-al-2021-india-excess-mortality-replication-files.zip.

Appendix Table 2. Regression Results (Reference Baseline is Jan.2015-April 2020)

	Overall	Urban	60-70yrs.	70-80yrs.	Above 80yrs.	Above median educated HHs
May-20	-0.0005	-0.0008*	-0.0013	-0.007	0.0019	-0.0010*
	(0.00031)	(0.00033)	(0.00225)	(0.00670)	(0.01581)	(0.00046)
Jun-20	-0.0007*	-0.0003	-0.0043*	-0.0035	0.0058	-0.0006
	(0.00030)	(0.00044)	(0.00170)	(0.00556)	(0.01910)	(0.00046)
Jul-20	0.0021***	0.0008	0.0140***	0.0374***	0.0795**	0.0021**
	(0.00051)	(0.00072)	(0.00365)	(0.01133)	(0.02713)	(0.00069)
Aug-20	0.0017*	0.0035*	0.0076**	0.0291**	0.0821	0.0018
	(0.00071)	(0.00160)	(0.00263)	(0.00915)	(0.04718)	(0.00134)
Sep-20	0.0014***	0.0022**	0.0073*	0.0322***	0.0202	0.0012*
	(0.00037)	(0.00071)	(0.00289)	(0.00734)	(0.01522)	(0.00060)
Oct-20	0.0021***	0.0023**	0.0113***	0.0245***	0.0469**	0.0019***
	(0.00038)	(0.00074)	(0.00257)	(0.00625)	(0.01760)	(0.00050)
Nov-20	0.0031***	0.0025*	0.0217***	0.0198**	0.0674***	0.0034***
	(0.00055)	(0.00113)	(0.00440)	(0.00632)	(0.01907)	(0.00073)
Dec-20	0.0029***	0.0022**	0.0130***	0.0409***	0.0721***	0.0036***
	(0.00045)	(0.00067)	(0.00295)	(0.00753)	(0.01878)	(0.00070)
Jan-21	0.0021***	0.0018**	0.0083***	0.0226***	0.0527**	0.0021***
	(0.00037)	(0.00061)	(0.00243)	(0.00614)	(0.01764)	(0.00052)
Feb-21	0.0019***	0.0013*	0.0083**	0.0234***	0.0434*	0.0019***
	(0.00039)	(0.00052)	(0.00285)	(0.00644)	(0.01695)	(0.00043)
Mar-21	0.0014***	0.0016*	0.0064**	0.0227**	0.0610**	0.0018***
	(0.00037)	(0.00075)	(0.00229)	(0.00711)	(0.02131)	(0.00054)
Apr-21	0.0015**	0.0021*	0.0044	0.0208**	0.0510*	0.0023***

	(0.00045)	(0.00085)	(0.00278)	(0.00695)	(0.02157)	(0.00065)
May-21	0.0028***	0.0040***	0.0075*	0.0368***	0.1835***	0.00110
	(0.00053)	(0.00090)	(0.00297)	(0.00881)	(0.03786)	(0.00064)
Jun-21	0.0032***	0.0038***	0.0118***	0.0600***	0.1147***	0.0016*
	(0.00052)	(0.00101)	(0.00289)	(0.01063)	(0.02367)	(0.00063)
R-sqr.	0.0030	0.0030	0.0130	0.0330	0.0760	0.0030
Obs.	10340076	6803642	662199	241518	65502	4738852

* p<0.05, ** p<0.01, *** p<0.001; columns refer to different samples; std. errors clustered at primary sampling unit level

Appendix Table 3. Regression Results (Reference Baseline is Jan.2018-April 2020(

	Overall	Urban	60-70yrs.	70-80yrs.	Above 80yrs.	Above median educated HHs
May-20	-0.0007*	-0.0008	-0.0019	-0.0105	-0.0006	-0.0013*
	(0.00035)	(0.00044)	(0.00245)	(0.00720)	(0.01657)	(0.00052)
Jun-20	-0.0010**	-0.0005	-0.0052**	-0.0063	-0.0102	-0.001
	(0.00038)	(0.00055)	(0.00193)	(0.00617)	(0.02062)	(0.00058)
Jul-20	0.0022***	0.0009	0.0142***	0.0343**	0.0799**	0.0020**
	(0.00052)	(0.00072)	(0.00371)	(0.01147)	(0.02793)	(0.00070)
Aug-20	0.0023**	0.0034*	0.0087**	0.0288**	0.0816	0.0026
	(0.00071)	(0.00163)	(0.00269)	(0.00939)	(0.04862)	(0.00135)
Sep-20	0.0012**	0.0023**	0.0068*	0.0339***	0.0173	0.0011
	(0.00041)	(0.00073)	(0.00292)	(0.00744)	(0.01626)	(0.00068)
Oct-20	0.0020***	0.0023**	0.0112***	0.0233***	0.0429*	0.0017***
	(0.00039)	(0.00071)	(0.00271)	(0.00651)	(0.01789)	(0.00052)
Nov-20	0.0028***	0.0024*	0.0210***	0.0167*	0.0650**	0.0034***

	(0.00057)	(0.00112)	(0.00441)	(0.00670)	(0.01973)	(0.00075)
Dec-20	0.0026***	0.0018**	0.0129***	0.0376***	0.0539**	0.0031***
	(0.00045)	(0.00065)	(0.00301)	(0.00756)	(0.01991)	(0.00070)
Jan-21	0.0019***	0.0016**	0.0074**	0.0191**	0.0445*	0.0020***
	(0.00039)	(0.00061)	(0.00257)	(0.00622)	(0.01812)	(0.00053)
Feb-21	0.0016***	0.0013*	0.0075*	0.0189**	0.0418*	0.0016***
	(0.00040)	(0.00053)	(0.00296)	(0.00655)	(0.01743)	(0.00045)
Mar-21	0.0012**	0.0014	0.0050*	0.0200**	0.0583**	0.0015**
	(0.00038)	(0.00074)	(0.00240)	(0.00733)	(0.02180)	(0.00055)
Apr-21	0.0015**	0.0017	0.0038	0.0217**	0.0496*	0.0021**
	(0.00046)	(0.00090)	(0.00306)	(0.00731)	(0.02149)	(0.00069)
May-21	0.0026***	0.0041***	0.0070*	0.0333***	0.1810***	0.00080
	(0.00055)	(0.00092)	(0.00313)	(0.00916)	(0.03743)	(0.00068)
Jun-21	0.0028***	0.0035***	0.0109***	0.0572***	0.0987***	0.00110
	(0.00057)	(0.00104)	(0.00305)	(0.01081)	(0.02478)	(0.00071)
R-sqr.	0.004	0.003	0.004	0.001	0.009	0.004
Obs.	5203573	1867345	3336228	3564741	1638832	1700036

* $p<0.05$, ** $p<0.01$, *** $p<0.001$; columns refer to different samples; std. errors clustered at primary sampling unit level

Appendix Table 4. Regression Results (Reference Baseline is Jan.2019-April 2020)

	Overall	Urban	60-70yrs.	70-80yrs.	Above 80yrs.	Above median educated HHs
May-20	-0.0014**	-0.0009	-0.0044	-0.0184*	-0.012	-0.0020**
	(0.00048)	(0.00055)	(0.00284)	(0.00814)	(0.01927)	(0.00066)
Jun-20	-0.0025***	-0.0017	-0.0075**	-0.0208**	-0.034	-0.0029**
	(0.00058)	(0.00089)	(0.00246)	(0.00782)	(0.02450)	(0.00090)
Jul-20	0.0018**	0.0001	0.0115**	0.0290*	0.0588*	0.0016*
	(0.00056)	(0.00081)	(0.00388)	(0.01182)	(0.02957)	(0.00077)
Aug-20	0.0014	0.0025	0.0054	0.0202*	0.056	0.0019
	(0.00075)	(0.00172)	(0.00295)	(0.01029)	(0.05145)	(0.00143)
Sep-20	0.0005	0.00130	0.0036	0.0294***	0.007	0.0001
	(0.00051)	(0.00086)	(0.00312)	(0.00814)	(0.01759)	(0.00089)
Oct-20	0.0010*	0.0018*	0.0070*	0.0135	0.0371	0.0006
	(0.00046)	(0.00079)	(0.00311)	(0.00790)	(0.01959)	(0.00060)
Nov-20	0.0020***	0.0021	0.0162***	0.0109	0.0638**	0.0028***
	(0.00060)	(0.00114)	(0.00460)	(0.00755)	(0.02045)	(0.00078)
Dec-20	0.0021***	0.0019**	0.0096**	0.0315***	0.0611**	0.0029***
	(0.00048)	(0.00072)	(0.00325)	(0.00811)	(0.02095)	(0.00074)
Jan-21	0.0016***	0.0014*	0.0059*	0.0152*	0.0389*	0.0018**
	(0.00040)	(0.00062)	(0.00269)	(0.00644)	(0.01901)	(0.00055)
Feb-21	0.0015***	0.0012*	0.0068*	0.0162*	0.0465**	0.0015**
	(0.00041)	(0.00057)	(0.00302)	(0.00668)	(0.01725)	(0.00048)
Mar-21	0.0010*	0.0013	0.0050*	0.0181*	0.0520*	0.0012*
	(0.00041)	(0.00082)	(0.00254)	(0.00783)	(0.02226)	(0.00061)
Apr-21	0.0011*	0.001	0.0017	0.0177*	0.0392	0.0015
	(0.00051)	(0.00108)	(0.00352)	(0.00805)	(0.02156)	(0.00084)
May-21	0.0019**	0.0039***	0.00450	0.0255*	0.1696***	0.00010

	(0.00063)	(0.00090)	(0.00349)	(0.00996)	(0.03881)	(0.00076)
Jun-21	0.0014*	0.00240	0.0086*	0.0427***	0.0749**	(0.00070)
	(0.00072)	(0.00124)	(0.00346)	(0.01151)	(0.02788)	(0.00099)
R-sqr.	0.004	0.004	0.017	0.044	0.101	0.004
Obs.	3627972	2347562	239053	90973	26824	1819670

* p<0.05, ** p<0.01, *** p<0.001; columns refer to different samples; std. errors clustered at primary sampling unit level