The COVID-19 mortality effects of underlying health conditions in India: a modelling study Appendix

Paul Novosad*

Radhika Jain†

Alison Campion‡

Sam Asher§

July 2020

1 Appendix

1.1 Construction of Indian survey dataset

The 4th District Level Health Survey (DLHS-4) and Indian Annual Health Survey (AHS) were conducted between 2012 and 2014 with no overlap in geographic coverage and jointly cover all states and union territories in India except Jammu and Kashmir, Dadra and Nagar Haveli, and Lakshadweep (additionally, data for Gujarat were collected but not made publicly available). The DLHS-4 is a single cross-sectional survey, while the AHS is a three year panel survey. Both surveys are representative at the district level and use two-stage stratified cluster sampling. In rural areas the primary sampling unit (PSU) was a village and in urban areas the PSU was a census enumeration block in the AHS or an urban frame survey block in the DLHS-4. PSUs were selected randomly with probability proportional to population size using the 2001 Indian Population Census in the AHS and rural DLHS-4, and with equal probability in the urban DLHS-4. Households were the secondary sampling unit (SSU) and were selected through systematic random sampling [1].

The DLHS-4 and AHS surveys administered a household-level questionnaire that collected information on age, sex, and self-reported symptoms, diagnosis, and treatment of illness for each household member. Additionally, both surveys administered a Clinical, Anthropometric and Biochemical (CAB) module to collect data on height, weight, hemoglobin, blood pressure, and blood glucose for adults. The CAB module was completed for all individuals 18 years or older in all sampled households in the DLHS-4. In the AHS, the CAB module was conducted in the second round of the survey in 2014 for all individuals 18 years or older in a randomly selected subsample of twelve PSUs per district, on average [2]. The publicly available DLHS-4 data provides a merged dataset in which each individual in the CAB module is matched to their household survey responses. However, there is no merged AHS data available, and there is no unique identifier to merge individuals between the household and CAB survey modules. We merged individuals with a completed CAB module in the AHS survey to their records in the household module using state, district, stratum (urban and non-urban), household unit, household number, and individual serial number identifiers. Individuals missing one or more of these key identifying fields could not be uniquely identified or merged and were dropped from the analysis. Of 1,209,926 individuals covered in the CAB module of the AHS, 819,351 (67.7%) had all required identifying fields and were matched with their records in the household survey.

We combined the DLHS-4 and AHS datasets to create a nationally representative dataset for analysis. Individuals with missing age, sex, height, weight, glucose, or blood pressure measurement were dropped from the sample. Individuals with reported age greater than 99 were assumed to be outliers and also dropped from the dataset. Additionally, as the thresholds for defining risk factors like obesity during pregnancy are not well established, pregnant women were also dropped from the sample. The final analysis sample contained 1,375,548 individuals, of which 577,994 (42.0%) came from the AHS and the remaining 797,554 (58.0%) came from the DLHS.

1.2 Estimation of glucose, obesity, and blood pressure in India

The DLHS-4 and AHS surveys both used the same data collection methods for biomarkers. Details of biomarker measurement are described in the survey manuals and summarized briefly here [2]. Systolic and diastolic BP were measured twice on the upper left arm, while sitting, with an interval of at least 3 minutes for each individual. The mean of the two measures was used to generate a single continuous measure of BP that was then classified as hypertension if systolic BP was 140 mmHg or higher or diastolic BP was 90 mmHg or higher, or if the individual reported a diagnosis of hypertension. Height and weight were directly measured and BMI was calculated as weight in kilograms divided by the square of height in meters. Blood glucose was measured from a single capillary blood sample (finger prick) and automatically converted into plasma equivalents by the glucometer. Single

*Dartmouth College
†Stanford University
‡Development Data Lab
§Johns Hopkins University School of Advanced International Studies

Novosad P, *et al. BMJ Open* 2020; 10:e043165. doi: 10.1136/bmjopen-2020-043165

capillary glucose measures are not ideal for clinical diagnosis of diabetes but have been recommended by the WHO for population surveillance in lower income countries [3]. Standard international thresholds were used to define diabetes as a plasma glucose reading ≥126mg/dL [7.0mmol/L] if fasting or ≥200 mg/dL [11.1 mmol/L] if reported not fasting. Individuals were asked to fast overnight before their glucose measurement. Self-reported fasting status was recorded in the DLHS-4 but not the AHS. We follow other studies that have used these data and use self-reported fasting status for all DLHS-4 participants and assume all AHS participants had fasted for the primary analysis [4, 5]. Assuming these individuals were not fasting changes estimated total diabetes prevalence in India from 9.8% to 8.5%.

All prevalence estimates from the DLHS-4 and AHS were weighted with a sample weight. Sample weights determined by the survey design of the DLHS-4 were provided in the publicly available data. These weights were multiplied by a district population weight, defined as the percentage of the national population in each district, to obtain the final sample weight. Due to a different survey design, the AHS does not have a sample weight in the data and so the sample weight was defined only by the district population weight.

1.3 Age-specific prevalence of risk factors in India and England

Table A1 shows age-specific prevalence of each risk factor considered in the study. Definitions of risk factors and data sources are given in the methods section and in Section 1.2.

Table A1
Health condition prevalences

	Prevalence (%)											
	India						**England**					
Age Group	18–39	40–49	50–59	60–69	70–79	80–99	18–39	40–49	50–59	60–69	70–79	80–99
Diabetes (Controlled)	0.3	1.7	3.1	3.9	4.3	3.3	0.7	5.0	8.8	11.3	13.9	13.9
Diabetes (Uncontrolled)	4.8	10.1	14.0	15.7	16.3	16.3	0.3	1.8	3.2	3.6	4.0	4.0
Hypertension	16.8	31.3	40.3	47.5	52.3	52.5	6.8	18.3	32.0	49.2	61.3	66.0
Obese (class I & II)	3.0	5.5	5.6	4.8	3.6	2.3	17.7	27.6	30.9	30.0	27.9	26.4
Obese (class III)	0.4	0.5	0.5	0.5	0.3	0.3	2.6	3.8	4.1	3.6	2.7	1.7
Chronic Heart Disease	1.4	4.4	8.1	15.0	24.5	31.4	1.3	5.0	10.5	21.5	34.7	42.6
Chronic Respiratory Disease	1.2	4.6	10.6	19.0	25.7	27.5	0.0	0.5	1.8	4.8	8.2	9.3
Asthma	1.3	2.7	4.2	6.2	8.1	8.6	9.3	8.5	8.4	8.5	8.4	7.8
Kidney Disease	6.6	13.1	17.3	24.4	37.0	51.7	3.0	5.5	7.9	14.2	27.2	45.8
Chronic Liver Disease	5.5	5.9	5.9	5.7	5.5	5.1	2.6	3.6	3.7	3.7	3.5	3.3
Haematological Cancer	0.0	0.0	0.0	0.1	0.1	0.1	0.1	0.2	0.4	0.7	1.0	1.0
Non-haematological Cancer	0.1	0.4	0.8	1.0	1.1	1.2	1.1	2.7	4.6	8.0	11.3	12.6
Stroke, Dementia	0.2	1.0	2.2	4.2	8.0	14.7	0.2	0.8	1.9	4.2	10.4	23.0
Other Neurological Condition	0.1	0.1	0.1	0.0	0.0	0.0	0.2	0.2	0.2	0.1	0.0	0.0
Psoriasis, Rheumatoid	0.8	1.4	1.8	2.3	2.5	2.3	2.0	3.4	4.5	5.2	4.8	3.7
Other Immunosuppressive Conditions	0.2	0.3	0.2	0.1	0.0	0.0	0.1	0.3	0.2	0.2	0.1	0.0

1.4 Demographics and prevalence of health conditions in England and the OpenSAFELY study population

We obtained estimated COVID-19 mortality hazard ratios for risk factors from the OpenSAFELY study [6]. The OpenSAFELY study sample includes adults 18 years or older enrolled with The Phoenix Partnership general practice system in England and covers 40% of the English population.

In order to calculate population risk for England for this study, we obtained national age, sex, and risk factor prevalence for the entire English population from a combination of census data, population health surveys and the GBD, as described in the methods. In Table A2 we present characteristics of the OpenSAFELY study sample against those of the entire English population as represented in our population statistics. Age, sex, and prevalence of most risk factors are very similar in the two. Hypertension prevalence in the OpenSAFELY sample is higher than in the English population (34.2% vs 28.1%) and asthma and class I obesity are lower (1.7% vs 9.2% and 19.1% vs 24.8%).

Table A2

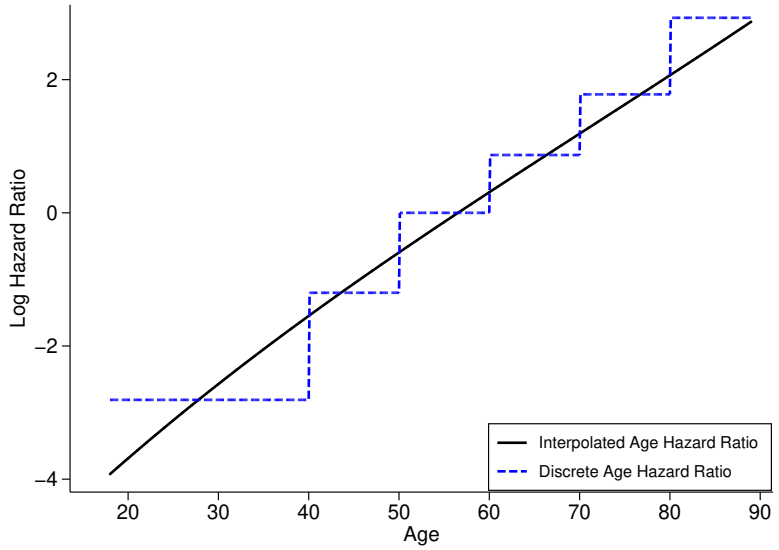
Demographics and prevalence of conditions in England and in OpenSAFELY

	England Prevalence (%)	
	OpenSafely Sample	This Study
Age 18-39	34.4	36.6
Age 40-49	16.5	16.3
Age 50-59	17.6	17.0
Age 60-69	13.8	13.3
Age 70-79	11.2	10.4
Age 80-99	6.5	6.3
Male	49.9	48.9
Diabetes (Controlled)	6.0	6.4
Diabetes (Uncontrolled)	2.8	2.1
Hypertension	34.2	28.0
Obese (class I & II)	19.1	24.8
Obese (class III)	2.7	3.1
Chronic Heart Disease	6.7	5.9
Chronic Respiratory Disease	4.1	2.5
Asthma	1.7	9.2
Kidney Disease	6.3	5.6
Chronic Liver Disease	0.7	2.6
Haematological Cancer	0.1	0.2
Non-haematological Cancer	0.5	2.6
Stroke, Dementia	2.1	1.5
Other Neurological Condition	1.0	0.1
Psoriasis, Rheumatoid	5.1	2.4
Other Immunosuppressive Conditions	1.6	0.1

1.5 Interpolation of age bin relative risks

Hazard ratios for age in OpenSAFELY are reported in the discrete bins 18–39, 40–49, 50–59, 60–69, 70–79, 80+. To obtain hazard ratios at continuous ages, we first converted hazard ratios to natural logs and then fitted a cubic polynomial to the midpoints of each bin. The hazard ratio is almost linear in age and the polynomial provides a very good fit (Figure A1). These continuous log hazard ratios were converted into relative risks assuming a mortality rate of 1% (as described in the methods section), and then collapsed into integer age bins for the analysis.

Figure A1
Age Interpolation: Fully-Adjusted Model



1.6 Sensitivity of results to sampling error

The two sources of sampling error in the data underlying the analysis are the hazard ratios in OpenSAFELY (which are reported with 95% confidence intervals) and the health condition prevalence estimates. We examined sensitivity to sampling error by resampling from simulated datasets with distributions indicated by the standard errors in the underlying data.

To examine sensitivity to sampling error in hazard ratios, we ran the analysis on 1000 samples where each hazard ratio was replaced with a number drawn from a distribution with mean and standard deviation equal to the reported hazard ratio and calculated standard error from OpenSAFELY. Draws were conducted in logs and then converted back to levels and relative risks for the analysis. We examined the distribution of estimates for: (A) the combined population relative risk from health conditions in India relative to England; (B) the modeled share of deaths under age 60 in England; and (C) the modeled share of deaths under age 60 in India (Figure A2). The red line in the figure indicates the statistic reported in the results.

In each case, the entire distribution of results is highly consistent with what is reported in the results. Population relative risk due to underlying health conditions is consistently lower in India than in England, with a 95% confidence interval of [0.86,0.97]. Nearly all draws for the share of deaths under 60 in England and in India are within half a percent of the primary estimate.

We performed a similar exercise to examine sensitivity to sampling error in prevalence estimates (Figure A3). For biomarker data from India, standard errors were calculated directly from the data. For prevalences obtained from GBD and for the England biomarkers, we used the standard errors or 95% confidence intervals reported with the underlying data.

The sensitivity of the results to sampling error in prevalences is even smaller than the sensitivity with respect to hazard ratios, likely because prevalence estimates are based on very large samples, while hazard ratios are calculated from only a small number of deaths. The 95% confidence interval for population relative risk in this simulation was [0.88,0.93], and the percentage of deaths under 60 in both countries also showed little variation.

Figure A2
Sensitivity Test 1: Hazard Ratio Uncertainty

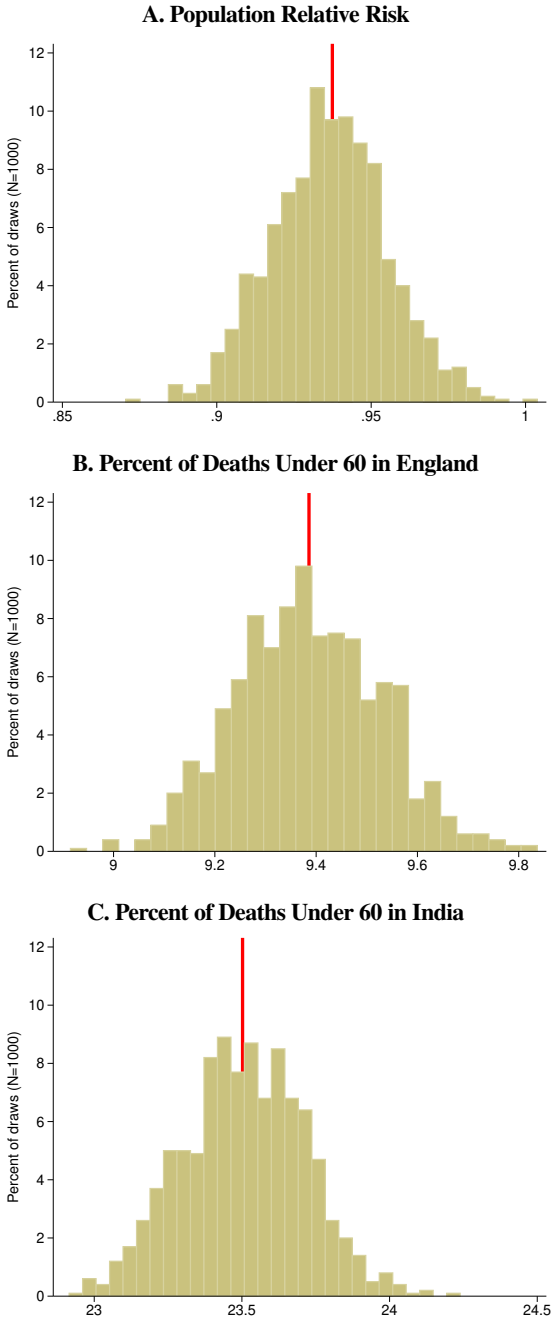
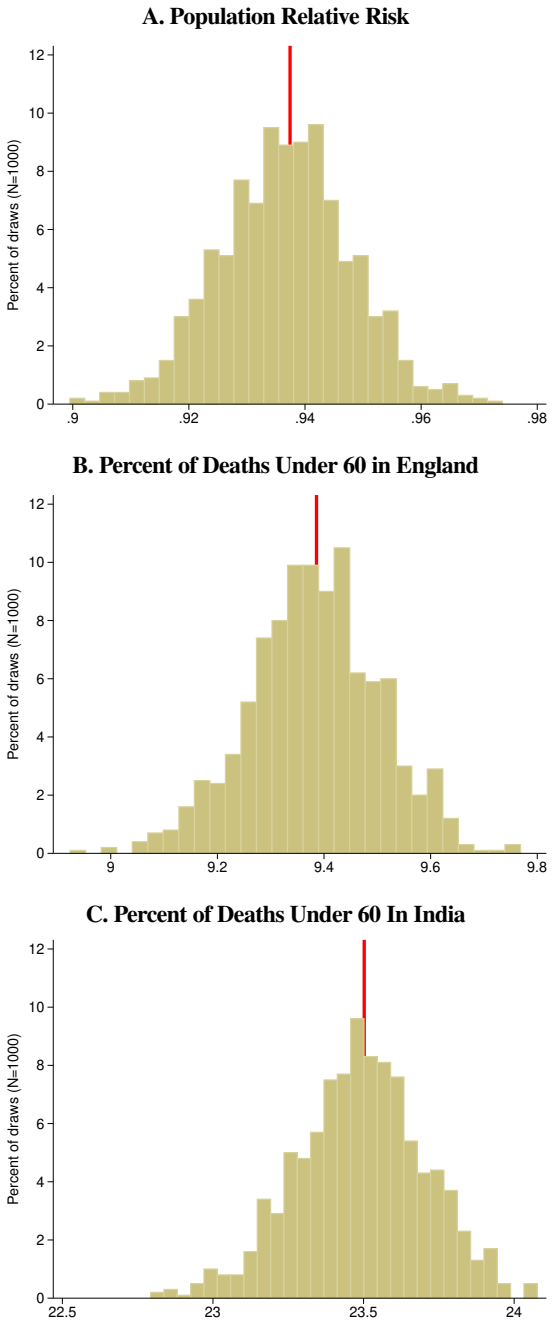


Figure A3
Sensitivity Test 2: Prevalence Uncertainty



1.7 Sensitivity to correlated health conditions

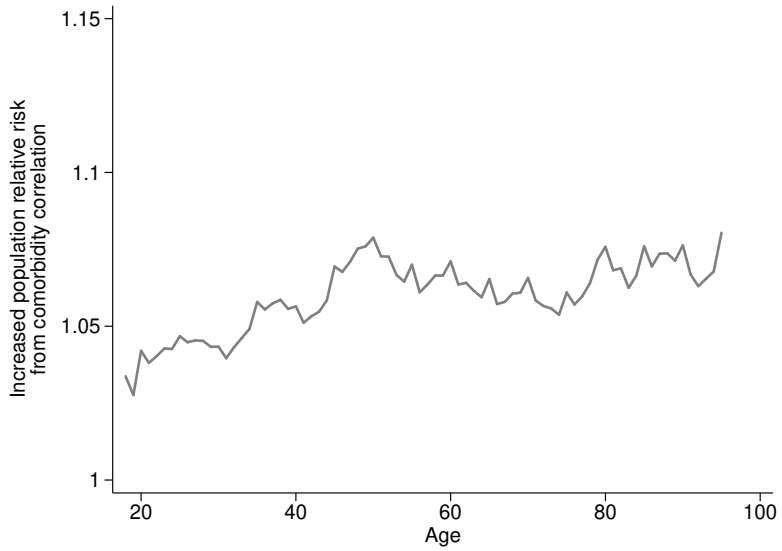
Our primary analysis implicitly assumes that COVID-19 mortality risk factors are uncorrelated with each other. This assumption is unavoidable given that many of the risk factors in India are drawn from the Global Burden of Disease studies which report age-specific prevalence but not the correlation across conditions.

If risk factors are correlated, then actual mortality risk will be higher than what is estimated because the relative risks of joint comorbidities will be compounded. Even if mortality risk is biased downward, our results may not be substantially affected because they rely on comparisons between England and India rather than focusing on the relative risk directly. If England and India have similar covariances of health conditions, then the biases will balance out exactly. Note that our analysis accounts for the most significant correlation of risk factors, which is the positive correlation between health conditions and age. This is accounted for because all health conditions are aggregated at each age.

To study the impact of correlated health conditions on population relative risk, we focused on the subset of health conditions available in the Indian microdata: obesity, diabetes, and hypertension. We calculate individual relative risk from the combined set of these risk factors by compounding the relative risks for these conditions at the individual level. We then aggregated the data across individuals and compared the combined relative risk to the same measure generated from the aggregate data, i.e. without accounting for correlation between comorbidities (Figure A4). The covariance of conditions increases the mortality rate by 5–10% across the age distribution. The effect does not vary much by age and is small relative to the aggregate risk factor shown in Figure 2.

The close match of our model to the age-specific death rates in OpenSAFELY is a second indication that condition covariance is unlikely to substantially bias the findings. As noted in the body of the paper, our model predicts 8.8% of deaths below the age of 60 in England; OpenSAFELY reports a figure of 8.6%. Because our hazard ratios come from an individual model that takes correlation between conditions into account, the close match of the age distribution of deaths suggests that our inability to observe condition covariance has not biased the results substantially.

Figure A4
Sensitivity Test 3: Covariance of Health Conditions



References

[1] Office of the Registrar General & Census Commissioner GoI India Ministry of Home Affairs. Annual Health Survey Report: A Report on Core and Vital Health Indicators Part I; 2014.

[2] Office of the Registrar General & Census Commissioner GoI India Ministry of Home Affairs. Annual Health Survey Report: A Report on Clinical, Anthropometric and Bio-Chemical Survey Part II. Office of the Registrar General & Census Commissioner, India Ministry of Home Affairs, Government of India; 2014.

[3] Organization WH, Fund ID. Definition and diagnosis of diabetes mellitus and intermediate hyperglycemia. World Health Organizatoin; 2006.

[4] Geldsetzer P, Manne-Goehler J, Theilmann M, Davies JI, Awasthi A, Vollmer S, et al. Diabetes and Hypertension in India: A Nationally Representative Study of 1.3 Million Adults. JAMA Internal Medicine. 2018;178(3):363–372.

[5] Bischops AC, Manne-Goehler J, Jaacks LM, Awasthi A, Theilmann M, Davies JI, et al. The prevalence of concurrently raised blood glucose and blood pressure in India: a cross-sectional study of 2,035,662 adults. Journal of Hypertension. 2019;37(9):1822–1831.

[6] Williamson E, Walker AJ, Bhaskaran KJ, Bacon S, Bates C, Morton CE, et al. OpenSAFELY: factors associated with COVID-19-related hospital death in the linked electronic health records of 17 million adult NHS patients. medRxiv. 2020:2020.05.06.20092999. Available from: `http://medrxiv.org/content/early/2020/05/07/2020.05.06.20092999.abstract`.