Social disadvantage, economic inequality, and life expectancy in nine Indian states

Sangita Vyas[*]   Payal Hathi[†]   Aashish Gupta[‡]

March 1, 2022

Abstract

An extensive literature documents the contributions of discrimination and social exclusion to health disparities. This study investigates life expectancy differentials along lines of caste, religion, and indigenous identity in India, home to some of the largest populations of marginalized social groups in the world. Using a large, high-quality survey that measured mortality, social group, and economic status, we estimate and decompose life expectancy differences between higher-caste Hindus, comprising Other Backward Classes and high-caste Hindus, and three of India's most disadvantaged social groups: Adivasis, Dalits, and Muslims. Relative to higher-caste Hindus, Adivasi life expectancy is more than four years lower, Dalit life expectancy is more than three years lower, and Muslim life expectancy is about one year lower. Economic status explains less than half of these gaps. The differences between the life expectancy of higher-caste Hindus and the life expectancies of Adivasis and Dalits are comparable to the Black-White gap in the US in absolute magnitude. The differences are larger in relative terms because overall life expectancy in India is lower. Our findings extend the literature on fundamental causes of global health disparities. Methodologically, we contribute to the literature on mortality estimation and demographic decomposition using survey data from low- and middle-income contexts.

Keywords: social inequality, life expectancy, India, caste, religion, indigenous identity

Significance Statement

India is one of the most hierarchical societies in the world. Because vital statistics are incomplete, mortality disparities are not quantified. Using novel survey data on more than 20 million individuals from nine Indian states representing about half of India's population, we estimate and decompose life expectancy differences between higher-caste Hindus, comprising Other Backward Classes and high-castes, and three marginalized social groups: Adivasis (indigenous peoples), Dalits (oppressed castes), and Muslims. The three marginalized groups experience large disadvantages in life expectancy at birth relative to higher-caste Hindus. Economic status explains less than half of these gaps. These large disparities underscore parallels between diverse systems of discrimination akin to racism. They highlight the global significance of addressing social inequality in India.

Introduction

Social disadvantage and health are closely linked. In the United States, for example, disparities in health and mortality between Black and White Americans have persisted over decades despite changes in technology, exposures, and diseases (Williams, Lawrence and Davis, 2019; Phelan and Link, 2015). However, the health impacts of social exclusion remain severely understudied in low- and middle-income countries (LMICs) (Deaton, 2002). Understanding health disparities in LMICs is important in part because social marginalization is no less present in poorer societies than in rich ones like the United States. Moreover, compared to high-income countries (HICs), patterns of disparities may be distinct in LMICs because population health is poorer, social safety nets are less robust, health care is less accessible, and mortality risk-factors differ (Tabutin and Masquelier, 2017).

This article describes and decomposes life expectancy disparities between socially marginalized and privileged groups in one of the most populated and stratified countries in the world, India. Marginalized social groups in India – Dalits, Adivasis, and Muslims – experience social exclusion based on caste, indigenous identity, and religion, respectively (Ambedkar, 1937; Xaxa, 1999; Sachar Committee, Government of India, 2006). Although each social group faces distinct forms of marginalization, together they comprise a population of over 450 million, greater than that of the US. Each group individually is also among the largest marginalized social groups in the world. We compare mortality for these three groups to the combined mortality of Other Backward Classes (OBC) and high-caste Hindus, who are relatively privileged in Indian society. We refer to this group as OBC/high-caste Hindus in the manuscript.

As in other contexts, social and economic disadvantages occur simultaneously in India. Dalits, Adivasis, and Muslims are poorer than privileged groups (Desai et al., 2010; Thorat and Neuman, 2012). Because health and economic status are also related (Case and Deaton, 2005), we quantify the extent to which differences in socioeconomic status (SES) can account for mortality differences between groups. To do this, we follow the literature on mortality disparities between Black and White Americans (Williams, Lawrence and Davis, 2019; Geruso, 2012), which uses standardization and decomposition techniques. These analyses are made possible in HICs by complete vital registration and multiple large surveys that reliably measure mortality, race, and SES (Elo, 2009). In many LMICs, however, decomposing life expectancy differences between groups is constrained by the sparseness of data on all-cause mortality linked with social conditions (Saikia and Kulkarni, 2016; Clark, 2019). For this reason, direct estimation of life tables disaggregated by social group and SES is not possible.

We overcome this limitation by using a unique and large-scale survey in nine Indian states from 2010 to 2011 that collected retrospective mortality information, social group, and SES from 4 million households. The sample is sufficient for directly estimating age-specific mortality rates. We find that the overall age-specific mortality rates estimated from this data correspond closely to official life tables. In particular, we compare our sex-specific, aggregated life tables to the life tables generated by the Government of India's Sample Registration System (SRS), a nationally representative system of mortality monitoring that does not disaggregate data by social group. After establishing the credibility of overall mortality estimates, we construct period sex-, group- and SES-specific life tables using standard demographic approaches. To examine the extent to which differences in SES between groups account for differences in life expectancy, we use a non-parametric standardization technique (Geruso, 2012). We use a cluster-bootstrap strategy to calculate standard errors for the life table and decomposition quantities we estimate (Cameron and Miller, 2015).

We document lower life expectancy at birth among Adivasis and Dalits compared to OBC/high-caste Hindus within each state. Relative to OBC/high-caste Hindus, Adivasi life expectancy at birth is about four years lower for females and five years lower for males. Life expectancy gaps between OBC/high-caste Hindus and Dalits is more than three years. We also provide estimates of Muslim life expectancy in India. In this region, Muslim life expectancy at birth is about one year lower than it is for OBC/high-caste Hindus. We find that lower life expectancy for all three groups relative to OBC/high-caste Hindus is not fully explained by differences in SES. Substantial disparities remain after accounting for differences in rural residence, wealth, and environmental factors.

Our findings have implications for health disparities worldwide, as well as for the global burden of mortality. The estimates of life expectancy at birth for Adivasis and Dalits are comparatively low globally. They are, for example, lower than contemporaneous population-level life expectancy at birth in many poorer contexts in sub-Saharan Africa. In terms of years, the disadvantages that we estimate for Adivasis and Dalits relative to OBC/high-caste Hindus are comparable to the Black-White gap in the US. In percentage terms, the disparities we observe are more substantial because life expectancy in India is less than four-fifths the level of life expectancy in the US. Compared to existing estimates on the extent to which SES accounts for the Black-White life expectancy gap (Geruso, 2012), the SES factors we use here account for a smaller fraction of the gaps between marginalized and privileged groups.

Our paper makes several contributions to the literature on social disadvantage and health in LMICs. First, we advance the scientific study of empirically estimating mortality in LMICs, which has been stymied by the lack of reliable and direct population-representative

estimates (Moultrie et al., 2013). We document an example in which a retrospective question on deaths in the household in the recent period produces estimates that follow typical patterns of mortality across the life course. These estimates also match expected levels of age-specific death rates in this context. Second, by examining the extent to which economic status can account for life expectancy disparities, our paper contributes to the existing literature using standardization and decomposition methods. Finally, we further the methodological literature on statistical inference of life table estimates constructed using survey data.

From a policy perspective, the Indian constitution provides protections for certain marginalized groups. However, discussions on health and well-being within and outside India often ignore inequalities based on indigenous identity, caste, and religion. This study highlights the importance of measuring and addressing social disparities within India and other LMIC contexts.

Data

This study uses the Government of India's Annual Health Survey (AHS) 2010-2011, a household survey that visited over 4 million households across approximately 20,000 primary sampling units (PSUs), and collected data on over 20 million individuals in nine relatively poor states in India: Assam, Bihar, Chhattisgarh, Jharkhand, Madhya Pradesh, Odisha, Rajasthan, Uttar Pradesh, and Uttarakhand. These states represent 48.5% of India's population (Office of the Registrar General, 2011), and are more rural than other Indian states. The total population in these states is twice that of the United States. We use data from the household roster, which recorded information on usual members living in surveyed households on January 1, 2010. We also use data from the mortality roster, which recorded the characteristics of surveyed households' usual members that died between January 1, 2007 and December 31, 2009. AHS instructions to enumerators are described in SI Table S1. SI Table S2 describes the sample.

The AHS recorded the caste group and religion of each household. We focus on estimating life expectancies for India's largest social groups: Adivasis (10% of our sample), Dalits (19%), Muslims (14%), and Other Backward Caste (OBC)/high-caste Hindus (56%). The data do not allow us to separately identify OBCs and high-caste Hindus. OBCs are relatively privileged compared to Dalits and Adivasis but marginalized compared to high-caste Hindus. Therefore, we expect the combined life expectancy of OBCs and high-caste Hindus to be lower than it would be for high-caste Hindus on their own. Since the OBC population is greater than that of high-caste Hindus (Desai et al., 2010), the combined life expectancy

figure is likely to be closer to that of OBCs than that of high-caste Hindus. We refer to the combined OBC and high-caste Hindu group as OBC/high-caste Hindus in this article. SI Table S3 shows the composition of the sample by responses to the AHS questions on caste group and religion. Further details on the social group categories we use in the analysis are in the SI section on 'Data preparation'.

Because the AHS asked members to list usual residents in the household and mortality rosters, these estimates are for usual residents of these states. In this respect, our estimates are similar to other demographic estimates. The AHS does not provide individual-level data on migrant status. According to the 2011 Census, out-of-state migration in the AHS states was small (Office of the Registrar General, 2011).

The AHS also recorded data on household SES, including rural residence, wealth, and environmental exposures including household solid fuel use and fraction of people defecating in the open in the PSU. To summarize a household's wealth, we construct a wealth index using a principal component analysis (PCA) of asset ownership and house infrastructure. The index is described in greater detail in the SI subsection on 'Socioeconomic status variables'. The relative disadvantage of Adivasis, Dalits, and Muslims, compared to OBC/high-caste Hindus, is evident in SI Figure S1 and Table S4, which display summary statistics by social group. Data are described further in the SI section on 'Data preparation'.

Methods

Estimating mortality rates and life expectancy

Using data from the AHS household and mortality rosters, we construct a dataset that records the number of person-years each individual contributes to each single-year age during the period January 2007 through December 2009, and whether the individual died at that age. We estimate social group-, age-, and sex-specific mortality rates (shown in SI Figure S2). Using standard procedures, we construct eight life tables to calculate life expectancy at birth separately for males and females of each social group. The number of person-years lived by those who died in each age interval, or $_na_x$, is calculated based on values from the Government of India's SRS 2007-2011 official life tables for states. SI Figure S3 shows that life expectancies and disparities calculated based on $_na_x$ values estimated directly from the AHS are similar to those calculated using SRS $_na_x$ values. In additional analyses, we estimate social group-, state-, and sex-specific life expectancies, as well as social group-, wealth-decile-, and sex-specific ones using the same methods. All estimates use the sample weights provided in the survey to make the data representative of the nine AHS states. SI

subsection 'Estimating mortality rates and life expectancy' further describes our methods for estimating life expectancy.

In order to rule out concerns regarding data quality from retrospective survey questions about mortality within the household (Timaeus, 1991), we compare age-specific mortality rates estimated from the AHS to those from the SRS and the National Family Health Survey (NFHS), India's Demographic and Health Survey (SI subsection 'Comparison of AHS mortality rates with SRS and NFHS' and SI Table S1 describe the SRS and NFHS in greater detail). Figure 1 displays this analysis. Age-specific rates estimated from the AHS match closely with rates from the SRS and NFHS. They are also smoother than those from the SRS and NFHS. Effects of age-misreporting, which is known to be high in India (Gerland, 2014), are apparent at age 75 in the AHS. This is also present to some extent in the SRS and NFHS surveys at older ages. Age-misreporting, which is more common among more disadvantaged groups, may lead to biases in estimates of mortality (Preston, Elo and Stewart, 1999). SI Figure S4 shows that for ages 40 to 85+, unadjusted mortality rates are similar to Gompertz rates, which adjust for age-misreporting. We show life expectancy differences at age 15 (e_{15}) in SI Figure S5 to clarify that our results are not entirely driven by differences in mortality at the youngest ages.

The disparities in life expectancy between marginalized groups and OBC/high-caste Hindus that we document are likely conservative for at least two reasons. First, because we cannot distinguish between high-caste Hindus and OBCs, we expect estimates of life expectancy for the combined OBC/high-caste Hindu group to represent a lower bound for high-caste Hindu life expectancy. Second, we are not able to capture mortality that occurs in households in which all members have died, which is more likely in Adivasi and Dalit households. We expect this particular bias to be relatively small. If mortality in single-person households is similar to mortality in two-person households, the closest counterfactual in our data, the number of deaths the AHS would have missed is less than one percent of total observed deaths.

Demographic Re-weighting

To understand the extent to which differences in SES can account for social-group differentials in life expectancy, we use a demographic re-weighting strategy (Geruso, 2012). Non-parametric re-weighting techniques allow for studying non-linear functions like life expectancy and, by matching on the full distribution of observed characteristics, are more flexible than regression techniques.

In practice, the demographic re-weighting technique estimates counterfactual life ex-

pectancies for marginalized social groups, re-weighting these groups so that they match the distribution of SES among OBC/high-caste Hindus. Because the marginalized social groups we study have lower SES than OBC/high-caste Hindus, the re-weighting strategy produces counterfactual life expectancies by up-weighting wealthier individuals and down-weighting less wealthy individuals in each marginalized group. This is implemented by estimating a re-weighting function as follows:

$$\psi^{MG}(c_i) = \frac{f(c_i|HC)}{f(c_i|MG)}, \tag{1}$$

where MG represents the marginalized group considered, and HC represents OBC/high-caste Hindus. c_i represents a vector of observable characteristics for individual i that are correlated with life expectancy, including sex, age group, social group, and SES. f represents the probability density function. Each individual in the sample is multiplied by her corresponding re-weighting function to produce counterfactual age-specific mortality rates and counterfactual life expectancies for each marginalized group. Re-weighting methods are described further in the SI subsection on 'Demographic re-weighting'.

The SES characteristics included in the re-weighting exercise are determined based on regression analysis of the characteristics associated with mortality (see SI Tables S5 and S6). They include rural residence, wealth (the intersection of wealth index quintile and land ownership), and environmental exposures (household solid fuel use for individuals age five and older, and household solid fuel use intersected with four categories of the fraction of people defecating in the open in the PSU for children younger than age five). These factors have been identified in the prior literature as important determinants of mortality (Cutler, Deaton and Lleras-Muney, 2006; Phelan and Link, 2015; Coffey and Spears, 2017). The regression equations are described further in the SI subsection 'Description of linear regression'.

Inference: Cluster-bootstrap Method

Standard errors are estimated using the cluster-bootstrap method described in Cameron and Miller (2015). We use this procedure because the AHS randomly sampled PSUs (villages or census enumeration blocks) rather than individuals, and both outcomes and explanatory variables are likely correlated within PSUs. This approach has been previously used to construct standard errors around life table quantities estimated from cluster sample surveys Cai et al. (2010); Gupta and Sudharsanan (2020).

Within districts, the AHS randomly sampled villages and urban areas stratified by population size. The number of PSUs randomly sampled from each stratum were determined

based on the district's population distribution across strata. All households in sampled PSUs were interviewed.

For the bootstrap, we resample with replacement $J_{strat,dist}$ PSUs within each district-stratum, with $J_{strat,dist}$ equal to the total number of PSUs in that district's stratum in the original AHS sample. Because our resampling procedure maintains the distribution of PSUs across strata within districts, we use the original AHS sample weights, which vary at the district-stratum level, to analyse each resample. Using the dataset generated by each resample, we estimate age-specific mortality rates and life tables. For the decomposition, we estimate a new re-weighting function and counterfactual life tables using each resample. We repeat this process 500 times, and the standard deviation of the 500 resulting estimates for each statistic are used for calculating 95% confidence intervals.

Results

Marginalized social groups have lower life expectancies

Figure 2 shows female and male life expectancies at birth for the four social groups we study. Confidence intervals are calculated using a cluster-bootstrap approach. Compared to OBC/high-caste Hindus, we observe lower life expectancies at birth among marginalized social groups. Adivasis have the lowest life expectancy among the four groups. Differentials between Adivasis and OBC/high-caste Hindus are almost four years for women and almost five years for men. The gap between Dalits and OBC/high-caste Hindus is of similar magnitude: more than three years for both women and men. Muslim life expectancy is about one year less than that of OBC/high-caste Hindus. Both overall levels of mortality among marginalized social groups and the absolute mortality differentials between groups are comparatively large. Life expectancies for Dalits and Adivasis are similar to those of the poorest countries in the world. The overall gaps are similar in absolute terms to the contemporaneous Black-White gap in the United States (Arias and Xu, 2018) and the Arab-Jewish gap in Israel (Saabneh, 2016).

Muslims have overall lower life expectancy at birth compared to OBC/high-caste Hindus. The gap between Muslims and OBC/high-caste Hindus is smaller relative to other marginalized groups. This is consistent with the prior literature (Sachar Committee, Government of India, 2006; Bhalotra, Valente and Van Soest, 2010). As reasons for lower mortality among Muslims, research has identified lower exposure to open defecation among Muslim children (Coffey and Spears, 2017), lower rates of cervical cancers among Muslim women (Gomes et al., 2017), lower consumption of alcohol (International Institute for Population Sciences -

IIPS/India and ICF., 2017), and lower incidence of suicide (Thimmaiah et al., 2016).

Research on caste and social identity has emphasized that although there are features of social stratification that are common across India, marginalization manifests differently from region to region (Srinivas et al., 1962). Mortality risks also vary across states. Figure 3 shows life expectancy by social group, sex, and state. We do not estimate mortality rates for social groups that constitute less than 5% of a state's population. We find that across states, Dalits and Adivasis have lower life expectancy at birth compared to OBC/high-caste Hindus. Except for one state, Muslims have similar or lower life expectancy than OBC/high-caste Hindus.

Among the nine states, Adivasi life expectancy is highest in Assam, a society in which they face less discrimination compared to other AHS sample states (Gogoi and Saikia, 2020). The life expectancy of Dalits is lowest in Uttar Pradesh, and that of Adivasis is lowest in Madhya Pradesh. OBC/high-caste life expectancy and absolute disparities are low in Uttar Pradesh. These facts contribute to Uttar Pradesh having the lowest life expectancy among all Indian states (Registrar General and Census Commissioner of India, 2013). Lower life expectancy in Uttar Pradesh across social groups is likely a result of poor environmental health (Coffey and Spears, 2017) and healthcare provision (Drèze and Sen, 2013). States with more than five years of absolute disparities in life expectancy between OBC/high-caste Hindus and a marginalized social group include Uttarakhand, Jharkhand, Odisha, Chattisgarh, and Madhya Pradesh.

Figure 4 shows female-male differences in life expectancy at birth by social group. Gaps between female and male life expectancy are greatest for Adivasis, and smallest for OBC/high-caste Hindus. Prior research has emphasized relatively less gender inequality among Adivasis compared to other social groups in India (Maharatna, 2000; Xaxa, 2004). The extent to which the patterns documented here are driven by variation in gender inequality across India's social groups deserves further scientific scrutiny.

Life expectancy disparities remain after accounting for SES

Given that marginalized social groups are also poorer, to what extent are social group differences in life expectancy driven by economic disadvantage? Figure 5 shows life expectancy by social group and household wealth decile. These estimates are not adjusted by rural and environmental factors. SI Figure S6 shows that using state-specific PCAs to construct wealth deciles, instead of a combined nine-state PCA, does not meaningfully change these results.

We find that Adivasis and Dalits have lower life expectancies than OBC/high-caste Hin-

dus across wealth categories. Comparing Muslims and OBC/high-caste Hindus, we find similar life expectancies at poorer deciles, but lower life expectancies among Muslims at richer deciles. Although further scientific investigation is needed to understand these patterns, a part of the explanation for why OBC/high-caste Hindu and Muslim life expectancies diverge at richer deciles may be due to differences across deciles in the share of OBCs and high-caste Hindus within the OBC/high-caste Hindu group (see SI Figure S7).

Figure 6 explores the extent to which dimensions additional to household wealth can explain differences in life expectancy between social groups. It uses the re-weighting approach outlined in the Methods Section. We find that differences in rural residence, wealth, and environmental exposures do not fully account for the life expectancy gaps between marginalized social groups and OBC/high-caste Hindus.

The figure shows gaps in life expectancy at birth between marginalized groups and OBC/high-caste Hindus, separately for females and males. The vertical lines in the figure reflect 95% confidence intervals. The left-most estimates show the raw gaps. The differences are about one year for Muslim men and women; more than three years for Dalit women, Adivasi women, and Dalit men; and about five years for Adivasi men. The second set of estimates show gaps that remain after re-weighting the marginalized groups to reflect the distribution across rural and urban residence among OBC/high-caste Hindus. Accounting for rural residence reduces the gap for Adivasis and Dalits, but not substantially. For Muslims, it increases the gap slightly, because Muslims are more likely to live in urban areas than OBC/high-caste Hindus. The third set of estimates adds wealth quintile and land ownership to the re-weighting characteristics. Except for Muslim women, this reduces gaps. The right-most set of estimates adds environmental exposures. For children under age five, these include household solid fuel use and the fraction of individuals defecating in the open in the PSU. For individuals age five and older, we only include household solid fuel use. Environmental exposures do not explain gaps substantially, and for Muslims, they actually increase gaps, given lower exposure to open defecation among Muslims relative to OBC/high-caste Hindus (Coffey and Spears, 2017).

In the United States, observable SES characteristics explain about three-quarters of the life expectancy gap between Black and White Americans (Geruso, 2012). In the AHS states, they explain less than half. We find that the unexplained gaps are about half a year for Muslim men, one year for Muslim women, more than two years for Adivasi women, Dalit women, and Dalit men, and more than three years for Adivasi men.

Discussion

This study examines relationships between social disadvantage, economic status, and life expectancy in India. Using survey data we document large and important disadvantages in life expectancy at birth for Adivasis, Dalits, and Muslims compared to OBC/high-caste Hindus. These disparities cannot be explained by differences in rural residence, wealth, or environmental exposures.

In addition to having comparatively lower life expectancies, marginalized social groups in India are also larger in terms of population than most countries in the world. Our study therefore underscores the global significance of challenging social inequality in health in India. Indeed, extreme social stratification and exploitation in India may be contributing to global population health deficits and slower improvements in health worldwide.

From a comparative perspective, life expectancy gaps in India are similar in magnitude to ethnic and identity-based disparities such as by race in the US, Brazil, and South Africa, indigenous identity in New Zealand and Australia, and religion in Israel (Arias and Xu, 2018; Chiavegatto Filho, Beltrán-Sánchez and Kawachi, 2014; Gupta and Sudharsanan, 2020; Phillips et al., 2017; Saabneh, 2016). Even within India, the disparities in life expectancy between Adivasis and Dalits on the one hand, and OBC/high-caste Hindus on the other, are large. Life expectancy for OBC/high-caste Hindus in the nine AHS states is higher than the contemporaneous all-India life expectancy observed around 2010 (Registrar General and Census Commissioner of India, 2013). Life expectancies for Adivasis and Dalits, however, are lower than those observed for all of India in 1996-2000, more than ten years before the AHS survey (Registrar General and Census Commissioner of India, 2013). The gaps documented in this paper are also larger than the within-sample gap in life expectancy at birth between individuals at the 25th and 75th percentiles of the wealth distribution. Using survey data from the NFHS and a similar estimation approach, Gupta and Sudharsanan Gupta and Sudharsanan (2020) provide evidence that even when overall life expectancy is higher in other regions of India, such as in South India, disparities are still large.

Our findings highlight several first-order concerns for future research. Given ongoing epidemiological transitions in India and other LMICs, continuous monitoring of mortality within countries and globally is important. This study documents the value of several approaches for studying mortality and its determinants in LMICs. These include large-scale data collection exercises that contain retrospective questions on household deaths, empirical estimation of age-specific mortality rates, non-parametric re-weighting techniques, and cluster-bootstrap variance estimation approaches suitable for multi-stage sample survey data. These approaches are particularly valuable in the context of the large mortality

and economic impacts of the COVID-19 pandemic, which has likely affected marginalized social groups more severely. In addition, research on age contributions, causes of death, segregation, behaviors, and risk-factors, such as occupational exposures, may help further understand disparities in mortality and life expectancy. Following the qualitative and auto-biographical literature (Kamble, 1990; Valmiki and Mukherjee, 2008; George, 2019), studies that are able to document causal pathways from violence, exploitation, and discrimination to mortality are also valuable.

From a policy perspective, these findings suggest that population health interventions that explicitly challenge social disadvantage are essential because addressing economic inequality may not be sufficient (Ambedkar, 1943). Unfortunately, health policy in India and globally largely ignores exploitation, violence, and discrimination rooted in social inequality. This study justifies further action on social disparities in health within India, and advances the global conversation addressing inequalities based on race, ethnicity, indigenous identity, caste, and religion.

Data availability

AHS data are available at Zenodo (https://zenodo.org/ record/6062984.YgmCoUntyUk). Replication files are available at Zenodo (https://zenodo.org/record/6067096).

References

Ambedkar, Bhimrao Ramji. 1937. *Annihilation of caste: With a reply to Mahatama Gandhi.* University of Columbia.

Ambedkar, Bhimrao Ramji. 1943. *Ranade, Gandhi & Jinnah.* Columbia University: Dr. B.R. Ambedkar's life and beyond.

Arias, Elizabeth, and Jiaquan Xu. 2018. "United States life tables, 2015."

Bhalotra, Sonia, Christine Valente, and Arthur Van Soest. 2010. "The puzzle of Muslim advantage in child survival in India." *Journal of Health Economics*, 29(2): 191–204.

Cai, Liming, Mark D Hayward, Yasuhiko Saito, James Lubitz, Aaron Hagedorn, and Eileen Crimmins. 2010. "Estimation of multi-state life table functions and their variability from complex survey data using the SPACE Program." *Demographic research*, 22(6): 129.

Cameron, A Colin, and Douglas L Miller. 2015. "A practitioner's guide to cluster-robust inference." *Journal of Human Resources*, 50(2): 317–372.

Case, Anne, and Angus Deaton. 2005. "Health and wealth among the poor: India and South Africa compared." *American Economic Review*, 95(2): 229–233.

Chiavegatto Filho, Alexandre Dias Porto, Hiram Beltrán-Sánchez, and Ichiro Kawachi. 2014. "Racial disparities in life expectancy in Brazil: challenges from a multiracial society." *American journal of public health*, 104(11): 2156–2162.

Clark, Samuel J. 2019. "A General Age-Specific Mortality Model With an Example Indexed by Child Mortality or Both Child and Adult Mortality." *Demography*, 56(3): 1131–1159.

Coffey, Diane, and Dean Spears. 2017. *Where India goes: abandoned toilets, stunted development and the costs of caste.* Harper Collins.

Cutler, David, Angus Deaton, and Adriana Lleras-Muney. 2006. "The determinants of mortality." *Journal of economic perspectives*, 20(3): 97–120.

Deaton, Angus. 2002. "Commentary: The convoluted story of international studies of inequality and health." *International Journal of Epidemiology*, 31(3): 546–549.

Desai, Sonalde B, Amaresh Dubey, Brij Lal Joshi, Mitali Sen, Abusaleh Shariff, and Reeve Vanneman. 2010. "Human development in India." *New York: Oxford University.*

Drèze, Jean, and Amartya Sen. 2013. *An uncertain glory: India and its contradictions.* Princeton University Press.

Elo, Irma T. 2009. "Social class differentials in health and mortality: Patterns and explanations in comparative perspective." *Annual review of sociology*, 35: 553–572.

George, Sobin. 2019. "Reconciliations of Caste and Medical Power in Rural Public Health Services." *Economic and Political Weekly*, 54(40).

Gerland, Patrick. 2014. "UN Population Division's methodology in preparing base population for projections: case study for India." *Asian Population Studies*, 10(3): 274–303.

Geruso, Michael. 2012. "Black-white disparities in life expectancy: how much can the standard SES variables explain?" *Demography*, 49(2): 553–574.

Gogoi, Suraj, and Parag Jyoti Saikia. 2020. "Sharing and exchanging: understanding common grounds in northeast India." *Asian Ethnicity*, 21(1): 22–38.

Gomes, Mireille, Rehana Begum, Prabha Sati, Rajesh Dikshit, Prakash C Gupta, Rajesh Kumar, Jay Sheth, Asad Habib, and Prabhat Jha. 2017. "Nationwide mortality studies to quantify causes of death: relevant lessons from India's Million Death Study." *Health Affairs*, 36(11): 1887–1895.

Gupta, Aashish, and Nikkil Sudharsanan. 2020. "Large and persistent life expectancy disparities among India's social groups."

International Institute for Population Sciences - IIPS/India and ICF. 2017. *India National Family Health Survey NFHS-4 2015-16.* Mumbai, India: IIPS and ICF.

Kamble, Baby. 1990. "Jina Amucha." *Pune: Mansanman Prakashan. 2007 The Prisons We Broke.*

Maharatna, Arup. 2000. "Fertility, mortality and gender bias among tribal population: an Indian perspective." *Social Science & Medicine*, 50(10): 1333–1351.

Moultrie, Tom A, RE Dorrington, Allan G Hill, Kenneth Hill, IM Timæus, and Basia Zaba. 2013. *Tools for demographic estimation.* International Union for the Scientific Study of Population.

Office of the Registrar General. 2011. "Census of India 2011: Household Listing Operations." *Office of the Registrar and General Census Commissioner, India. Indian Census Bureau.*

Phelan, Jo C, and Bruce G Link. 2015. "Is racism a fundamental cause of inequalities in health?" *Annual Review of Sociology*, 41: 311–330.

Phillips, Bronwen, John Daniels, Alistair Woodward, Tony Blakely, Richard Taylor, and Stephen Morrell. 2017. "Mortality trends in Australian Aboriginal peoples and New Zealand Māori." *Population health metrics*, 15(1): 25.

Preston, Samuel H, Irma T Elo, and Quincy Stewart. 1999. "Effects of age misreporting on mortality estimates at older ages." *Population studies*, 53(2): 165–177.

Registrar General and Census Commissioner of India. 2013. *Compendium of India's Fertility and Mortality Indicators ,1971 - 2013.* Government of India.

Saabneh, Ameed M. 2016. "Arab–Jewish gap in life expectancy in Israel." *The European Journal of Public Health*, 26(3): 433–438.

Sachar Committee, Government of India. 2006. "Social, Economic and Educational Status of the Muslim Community of India." *Prime Minister's High Level Committee, Cabinet Secretariat, Government of India.*

Saikia, Nandita, and PM Kulkarni. 2016. "Data for Research into Health Inequality in India." *The Economic and Political Weekly*, 51(26): 111–116.

Srinivas, Mysore Narasimhachar, et al. 1962. "Caste in modern India and other essays." *Caste in modern India and other essays.*

Tabutin, Dominique, and Bruno Masquelier. 2017. "Mortality inequalities and trends in low-and middle-income countries, 1990-2015." *Population*, 72(2): 221–296.

Thimmaiah, Rohini, Vijayalakshmi Poreddi, Rajalakshmi Ramu, Sugavana Selvi, and Suresh Bada Math. 2016. "Influence of religion on attitude towards suicide: An Indian perspective." *Journal of religion and health*, 55(6): 2039–2052.

Thorat, Sukhadeo, and Katherine S Neuman. 2012. *Blocked by caste: economic discrimination in modern India.* Oxford University Press.

Timaeus, Ian M. 1991. "Measurement of adult mortality in less developed countries: a comparative review." *Population index*, 552–568.

Valmiki, Omprakash, and Arun Prabha Mukherjee. 2008. *Joothan: An Untouchable's Life.* Columbia University Press.

Williams, David R, Jourdyn A Lawrence, and Brigette A Davis. 2019. "Racism and health: evidence and needed research." *Annual review of public health*, 40: 105–125.

Xaxa, Virginius. 1999. "Tribes as indigenous people of India." *Economic and political weekly*, 3589–3595.

Xaxa, Virginius. 2004. "Women and Gender in the Study of Tribes in India." *Indian Journal of Gender Studies*, 11(3): 345–367.
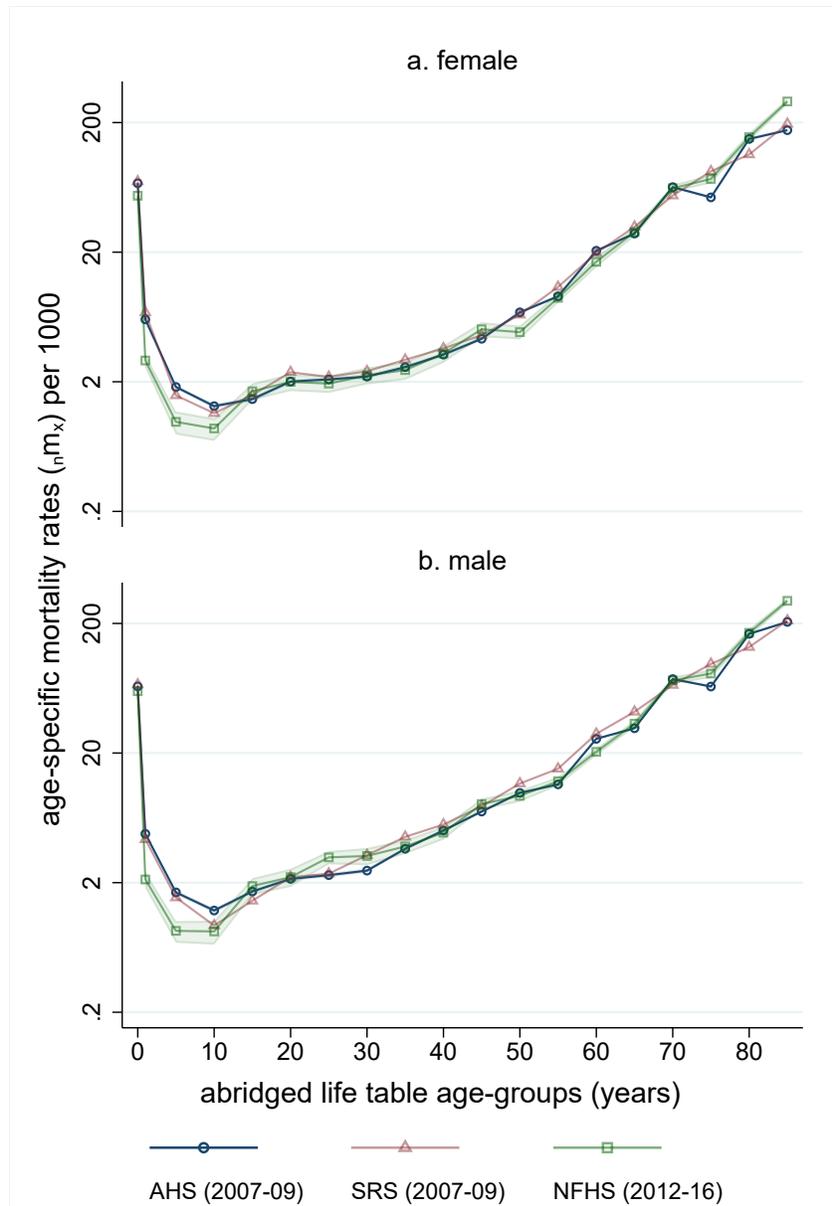
Figure 1: Similar age-specific mortality rates from the Annual Health Survey, Sample Registration System, and National Family Health Survey. Age-specific mortality rates are shown for females (Panel a) and males (Panel b) separately, from three different surveys. All AHS states are included, except for Uttarakhand, for which SRS data are missing. For the AHS, mortality rates are estimated based on the procedure described in the Methods Section. NFHS mortality rates are estimated according to the procedure described in Gupta and Sudharsanan (2020). NFHS rates are lower because they are for a later period. For the SRS, we average published age-specific mortality rates for 2007, 2008, and 2009, and weight states by their 2011 Census populations to produce an average across the eight states. Estimates from the AHS and NFHS use sample weights. 95% confidence intervals calculated using a cluster-bootstrap procedure are shown as shaded areas around the AHS (not visible because they are small) and NFHS lines. SRS age-specific mortality rates do not have confidence intervals because underlying SRS micro-data is not publicly available, and the SRS does not estimate clustered standard errors in its reports. Source: Annual Health Survey 2010-2011, Sample Registration System 2007-2009, and National Family Health Survey 2015-2016.
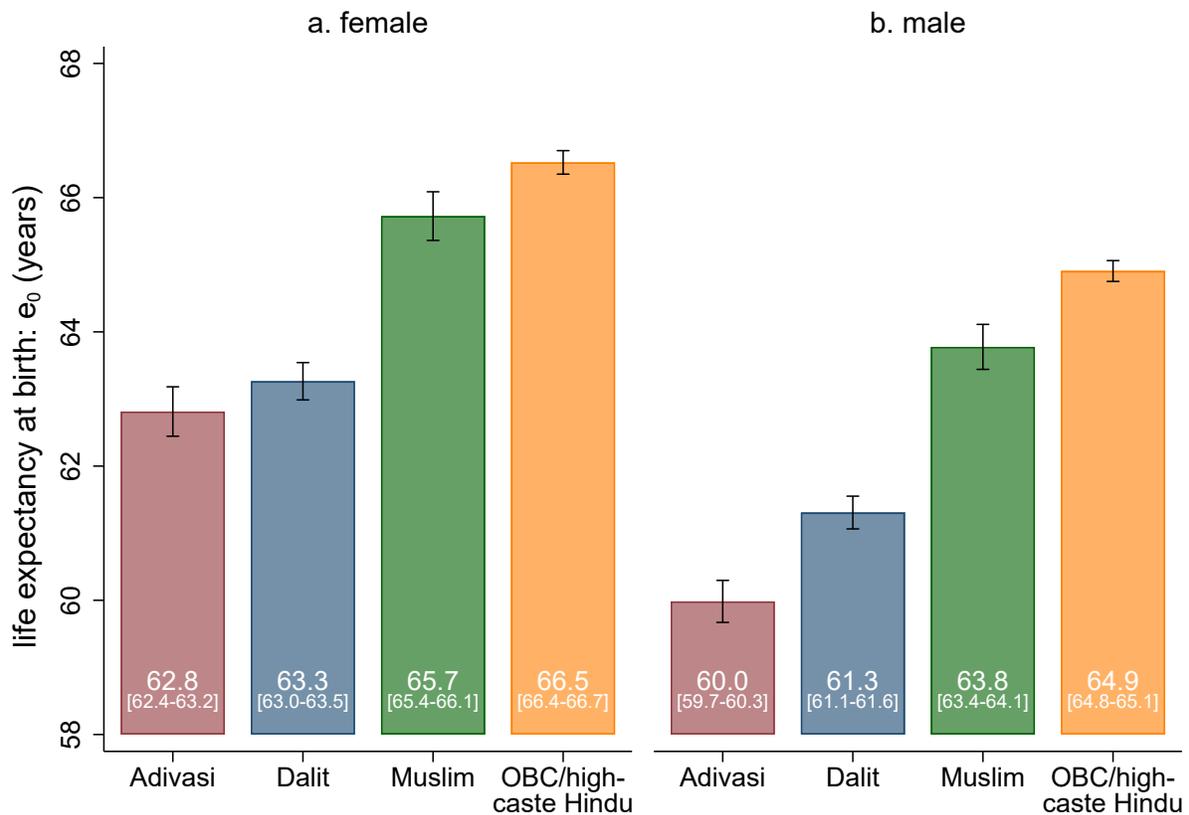
Figure 2: Lower life expectancy at birth among marginalized groups compared to OBC/high-caste Hindus. Life expectancy estimates for each sex and social group are calculated using standard life table procedures. Panel (a) shows female life expectancies, Panel (b) shows male life expectancies. Estimates use sample weights. The vertical lines around each estimate represent 95% confidence intervals calculated using a cluster-bootstrap procedure. Source: Annual Health Survey 2010-2011.
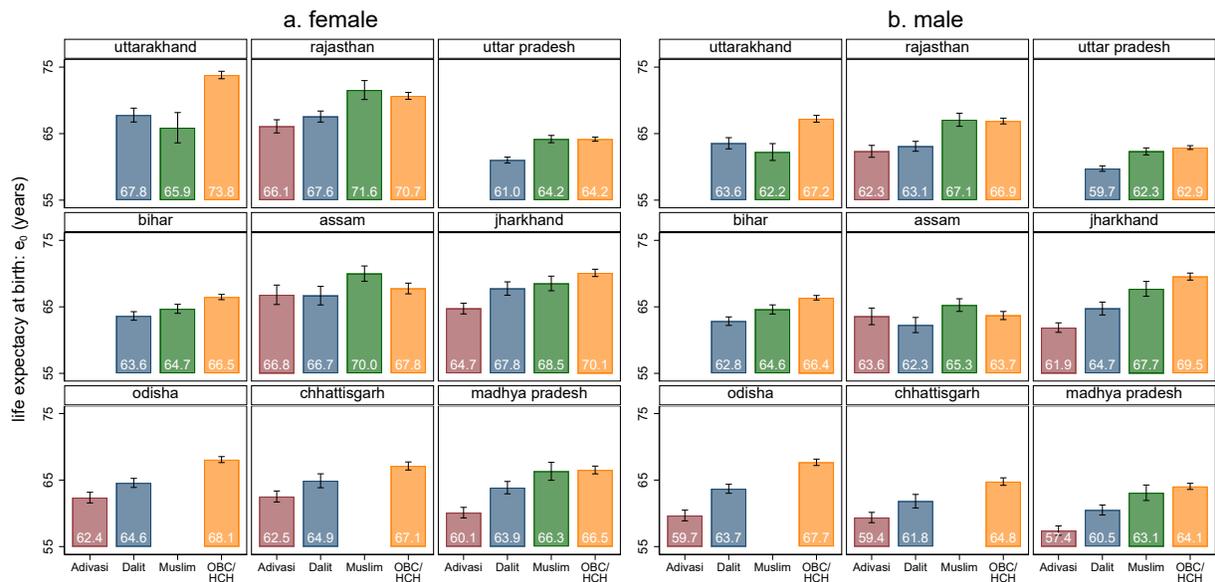
Figure 3: Life expectancy by state: Lower life expectancy at birth among Dalits and Adivasis within states. Life expectancy estimates for each state, sex, and social group are calculated using standard life table procedures. Panel (a) shows estimates for females, Panel (b) shows estimates for males. We generate state- and sex-specific life expectancy estimates for a group if the group's population is greater than 5% of a state's population. Estimates use sample weights. The vertical lines around each estimate represent 95% confidence intervals calculated using a cluster-bootstrap procedure. OBC/HCH: OBC/high-caste Hindu. Source: Annual Health Survey 2010 to 2011.
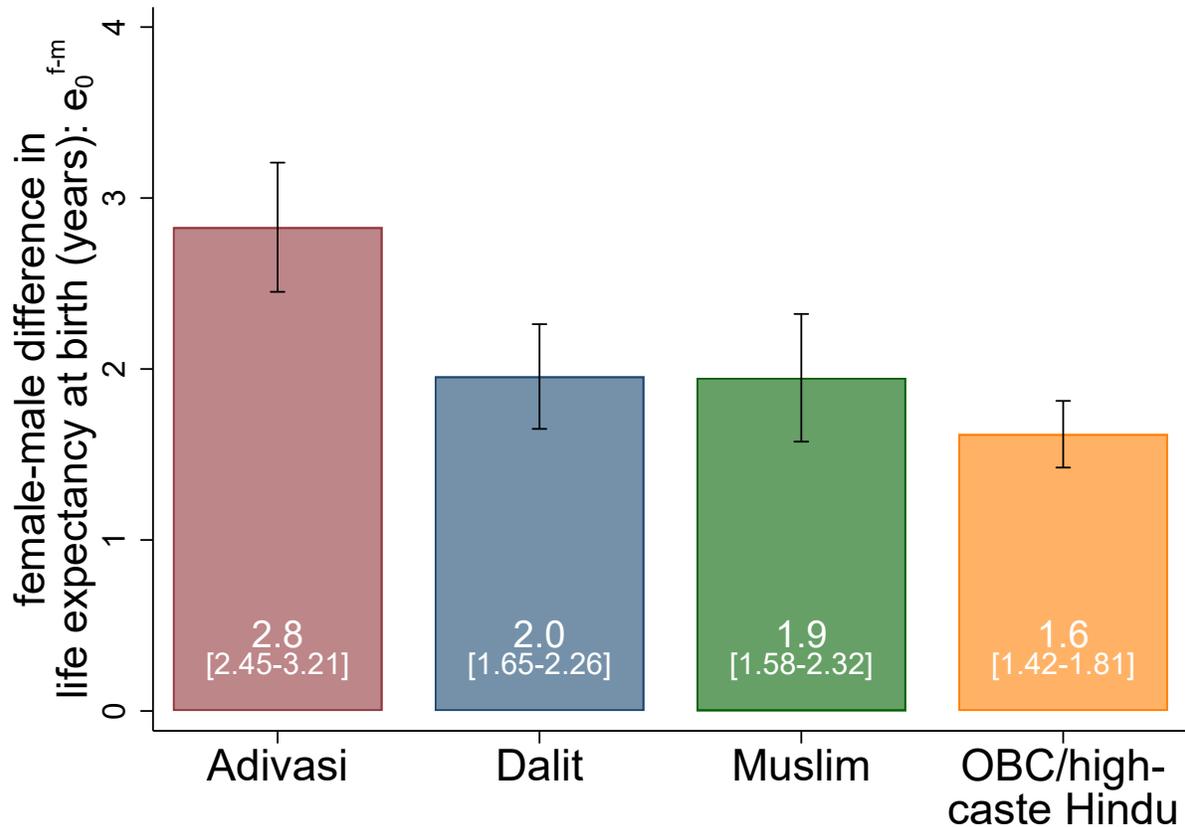
Figure 4: Female-male difference in life expectancy at birth by social group. Life expectancy estimates for each sex and social group are calculated using standard life table procedures. Estimates use sample weights. The vertical lines around each estimate represent 95% confidence intervals calculated using a cluster-bootstrap procedure. OBC/HCH: Other Backward Classes / high caste Hindu. Source: Annual Health Survey 2010-2011.
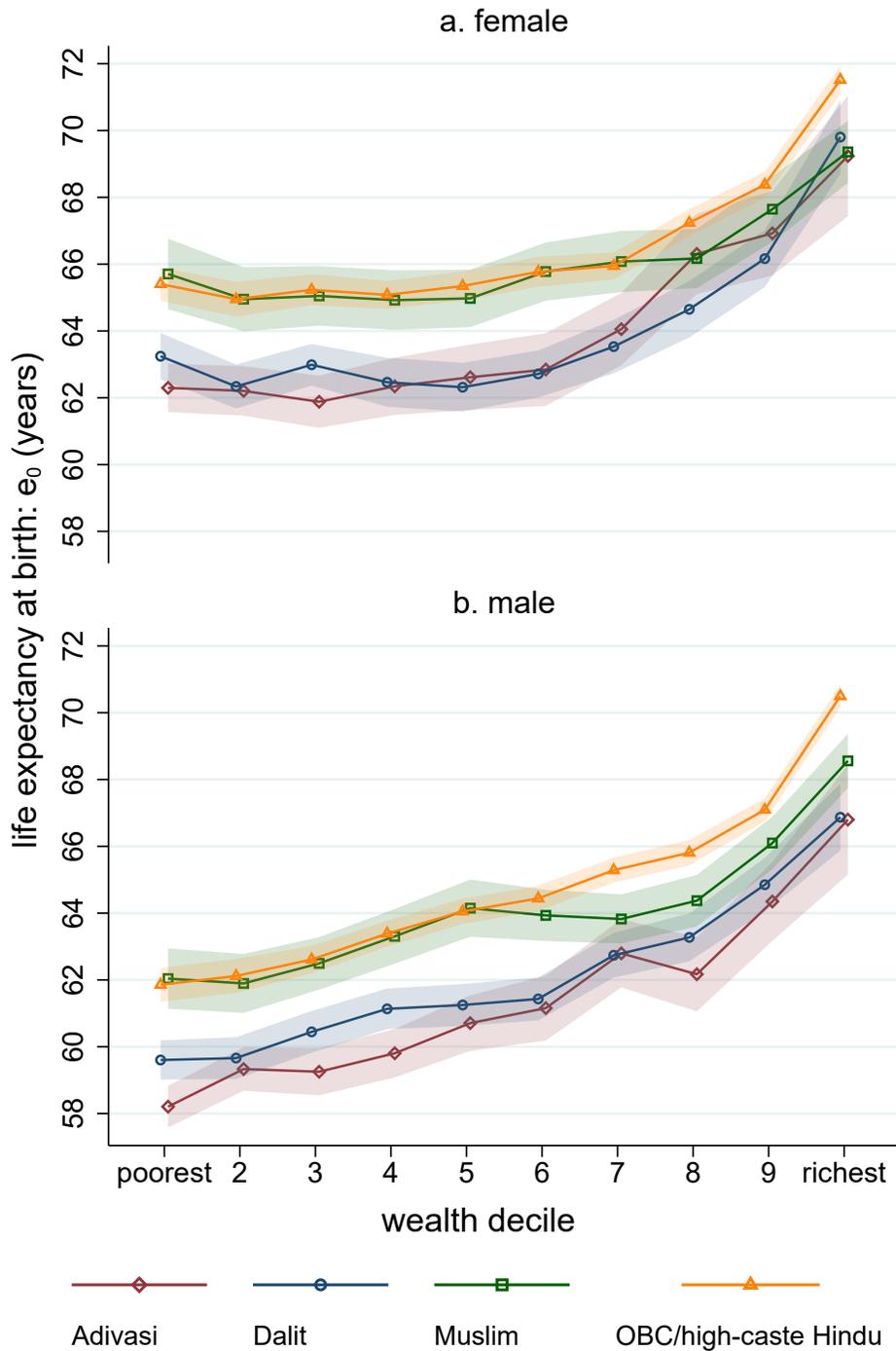
Figure 5: Life expectancy by household wealth: Lower life expectancy at birth among Dalits and Adivasis at all levels of wealth. Wealth decile are deciles of a wealth index constructed using a principal component analysis of household assets and house infrastructure. Life expectancy estimates for each wealth decile, sex, and social group are calculated using standard life table procedures. Panel (a) shows estimates for female life expectancies, Panel (b) shows male life expectancies. Estimates use sample weights. The shaded areas around the lines represent 95% confidence intervals calculated using a cluster-bootstrap procedure. Source: Annual Health Survey 2010-2011.
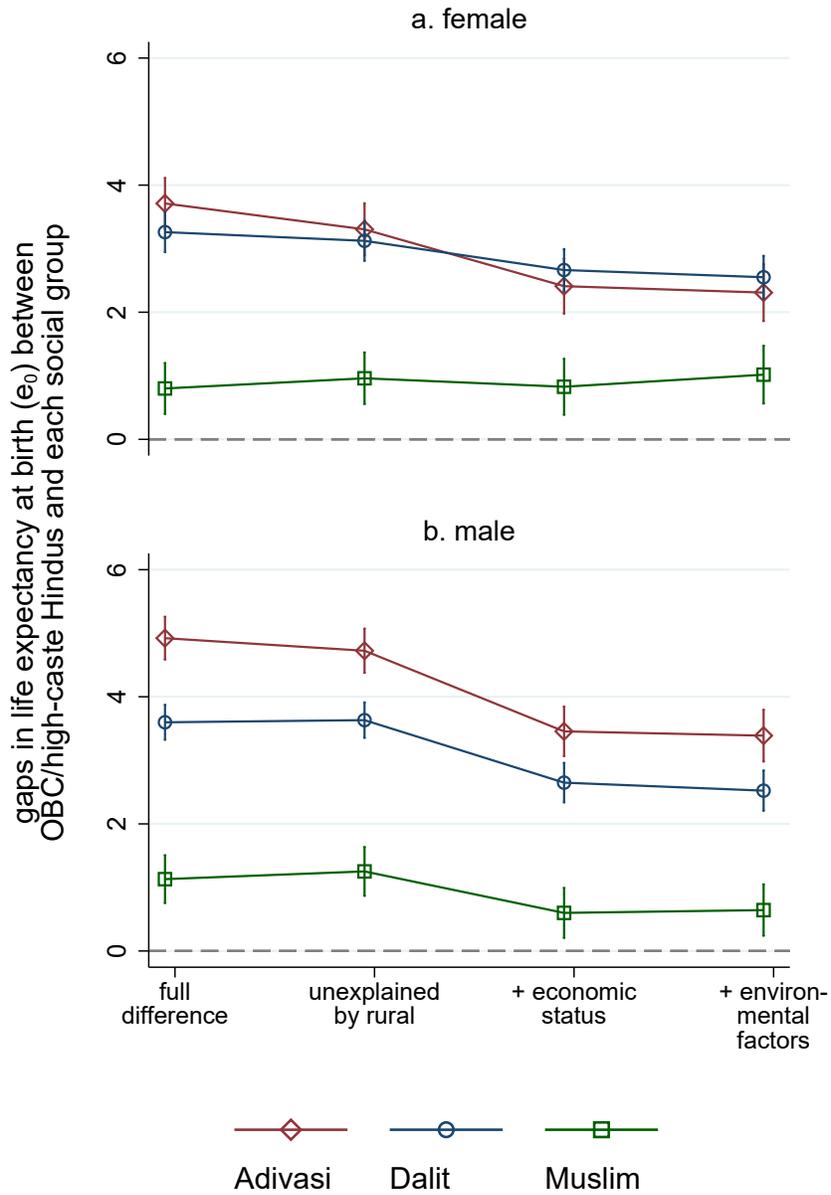
Figure 6: Demographic re-weighting: Unaccounted-for gaps in life expectancy (years) between OBC/high-caste Hindus and marginalized groups. Marginalized groups are re-weighted to reflect the same distribution of characteristics as among OBC/high-caste Hindus in three ways. Panel (a) shows estimates for females, Panel (b) shows estimates for males. 'Full difference' shows the full life expectancy gap between OBC/high-caste Hindus and marginalized social groups (HC e_0 - Adivasi e_0, for instance). 'Unexplained by rural' shows the gap that remains after accounting for differences in rural residence. '+ economic status' shows the gap that remains after additionally accounting for differences in wealth index quintile intersected with land ownership. '+ environmental factors' shows the gap that remains after additionally accounting for differences in environmental exposures. For children under age five, this includes household solid fuel use intersected with four categories of PSU open defecation. For individuals age five and older, environmental factors only include household solid fuel use. The vertical lines around each estimate represent 95% confidence intervals calculated using a cluster-bootstrap procedure. Source: Annual Health Survey 2010-2011.

Supplementary Information for

## Social disadvantage, economic inequality, and life expectancy in nine Indian states

Sangita Vyas, Payal Hathi, Aashish Gupta

E-mail: sangita.vyas@utexas.edu, phathi@berkeley.edu, and aashishg@sas.upenn.edu

This PDF file includes:

Supporting Information Text

Supplementary Information: Methods

Estimating mortality rates and life expectancy. The AHS was designed to generate estimates of death rates and other demographic indicators at the district level, thus it has a large sample that is well-suited to estimating age-specific mortality rates. The household roster recorded the date of birth and sex of all members living in surveyed households on January 1, 2010. The survey also included a mortality roster, which recorded sex, age at death, and date of death, among other characteristics, of any usual members of surveyed households that died between January 1, 2007 and December 31, 2009. For individuals who were alive in January 2010, we estimate exact age using the date of birth provided in the household roster. For household members that died during the period, we use the exact age at death and date of death provided on the mortality roster. Using this information, we construct a dataset that records the number of person-years each individual contributed to each single-year age during the period January 2007 through December 2009, and also includes an indicator for whether or not the individual died at that age. The procedure we use to construct this dataset is discussed in greater detail in the section "Estimating person-years contributed to single-year ages."

With this dataset, we estimate age-specific mortality rates separately for each sex-by-social group using the following equation:

$$_nm_x^{s,g} = \frac{\sum_{i=1}^{I}(died\ between\ age\ x\ to\ x+n_i) \times (weight_i)}{\sum_{i=1}^{I}(person\ years\ lived\ between\ age\ x\ to\ x+n_i) \times (weight_i)}. \tag{1}$$

In this equation, m refers to mortality rate, x to the interval's beginning age, and n to the length of the age interval. We use the abridged life table age ranges used by India's Sample Registration System (SRS).[*] i refers to individuals of sex s and social group g, and I to the total number of individuals of sex s and social group g covered in the survey. Observations are weighted by the sample weight provided in the survey to make the data representative of the nine AHS states. The reference period for deaths and person-years lived is January 2007 to December 2009.

We use standard life table procedures to estimate life expectancies for each social group, as well as each state-by-social group, and wealth decile-by-social group. To estimate life expectancies by additional characteristics, we modify Equation 1 to generate mortality rates for individuals defined by age group, sex, social group, and additional characteristics z, $_nm_x^{s,g,z}$. In our analyses, z represents states or wealth decile.

The number of person-years lived by those who died in age intervals x to $x+n$, or $_na_x$, is calculated based on values from SRS 2007-2011 life tables for states. The SRS does not produce life tables for different social groups. We therefore estimate group and sex-specific $_na_x$ values for the overall sample of nine states by taking a weighted average of $_na_x$ across states, where each state's weight is the fraction of deaths of individuals of the social group that occurred in the state.[†] Within states, for each of the social groups, we use state and sex-specific $_na_x$ from the SRS. We estimate wealth decile, group, and sex-specific $_na_x$ by taking a weighted average of $_na_x$ across states, where each state's weight is the fraction of deaths of individuals of a particular social group and wealth decile that occurred in the state.

In practice, the use of different $_na_x$, for instance, $_na_x$ directly estimated from the AHS, do not change our substantive results. This can be seen in Figure S3. Given that changes in $_na_x$ values do not change life expectancy estimates in general (1, pg. 47), this is expected. We prefer the life expectancy estimates based on SRS $_na_x$, as AHS $_na_x$ values are likely more affected by age-misreporting and the age-composition of the AHS sample.

Demographic re-weighting. The re-weighting function, defined in Equation 1 of the main paper, is directly applied in the estimation of counterfactual age-specific mortality rates for marginalized groups. Counterfactual age-specific mortality rates are:

$$_nm_x^{s,MG,CF} = \frac{\sum_{i=1}^{I}(died\ between\ age\ x\ to\ x+n_i) \times (weight_i) \times \psi^{MG}(c_i)}{\sum_{i=1}^{I}(person\ years\ lived\ between\ age\ x\ to\ x+n_i) \times (weight_i) \times \psi^{MG}(c_i)}, \tag{2}$$

where MG represents the marginalized group considered and HC represents OBC/high-caste Hindus. c_i represents a vector of observable characteristics for individual i that are correlated with life expectancy, including sex, age group, social group, and SES. ψ^{MG} is the re-weighting function. We use counterfactual age-specific mortality rates to estimate counterfactual life expectancies for each marginalized group.

Within each bin defined by sex, age group, and social group, we split the sample into additional bins based on exhibiting various combinations of SES. In the first scenario, the sample is split into two bins based on whether the individual lives in a rural area or not. The second scenario creates 20 bins by adding wealth status in five quintiles and a binary indicator for land ownership. The third scenario additionally includes environmental exposures. For children under age five, this includes household solid fuel use and four levels of PSU open defecation, generating 160 bins. For individuals age five and

---

[*] The age intervals in years are as follows: zero to one, one to five, five to ten, ..., 80 to 85, and 85+.

[†] The SRS did not produce life tables for the AHS states of Chhattisgarh, Jharkhand, or Uttarakhand. For Chhattisgarh, we use $_na_x$ values from Madhya Pradesh life tables, since Chhattisgarh was earlier a part of Madhya Pradesh. Similarly, we use $_na_x$ values from Bihar life tables for Jharkhand. Although Uttarakhand was formed from the state of Uttar Pradesh, the demography of Uttarakhand is very different, and so we use $_na_x$ values from India as a whole for this state.

over, environmental exposures include only household solid fuel use, generating 40 bins.[‡] The characteristics included in the re-weighting exercise are determined based on regression analysis, discussed in Appendix .

It is important to note that the unexplained gap in life expectancy should not be interpreted as the gap that can be attributed to discrimination or oppression for at least two reasons. First, differences in observed SES may arise from discriminatory processes. If this is the case, then the unexplained gap would underestimate the portion that can be attributed to discrimination. Second, the socioeconomic characteristics that we use to re-weight the sample of marginalized groups are necessarily limited to the characteristics measured in the survey. There may be other relevant characteristics that are not accounted for. If this is the case, then the unexplained gap in life expectancy may overestimate the portion that can be attributed to discrimination.

Description of linear regression. To motivate the re-weighting exercise, we present a regression analysis of the characteristics associated with mortality. We estimate these regressions separately by age, rather than in a pooled sample, because we hypothesize that the characteristics that affect child mortality may be different from the characteristics that affect adult mortality. Moreover, because the life table estimates mortality separately for different age groups, it is possible to use different characteristics for children and adults in the re-weighting exercise.

We estimate linear probability models predicting death separately for children under age five and individuals age five and older. We cluster standard errors at the level of the primary sampling unit and use survey weights. We estimate regressions of the following form:

$$died_{ihv} = group_{hv}\beta +$$
$$\alpha rural_v + economic_{hv}\delta + environmental_{hv}\gamma +$$
$$\theta_{ihv} + \epsilon_{ihv}. \qquad [3]$$

where $died$ is a dummy variable taking on values 0 or 1,000, indicating that the individual died between January 2007 and December 2009; $group$ is a vector of dummy variables indicating social group with OBC/high-caste Hindus as the omitted category; $rural$ is a dummy variable indicating rural residence; $economic$ is a vector including wealth decile and a dummy variable indicating land ownership; and $environmental$ is a vector including dummy variables for household open defecation and household solid fuel use, and continuous variables indicating the fraction of individuals in the PSU defecating in the open and the fraction eating food cooked on solid fuels. θ represents a vector of life table age group fixed effects. For regressions predicting child mortality, θ also includes fixed effects for potential exposure in months, which is the total number of months the child could have lived from January 2007 through December 2009. For children who were born before January 2007, potential exposure is 36 months. For children born after January 2007, potential exposure is the number of months the child could have lived before December 2009.

Supplementary Information: Data preparation

Annual Health Survey (AHS) data have been made available to the public by the Government of India.[§] Data from the household roster are available in "COMB" datasets, and data from the mortality roster are available in "MORT" datasets. We use data on usual residents of households surveyed in the first round of the AHS survey, which collected data in 2010-2011.

Caste and religion categories. The AHS records $social\_group$ and $religion$ variables separately for each household. Options for the $social\_group$ variable are 'SC', 'ST', and 'Other' Social Group. Options for the $religion$ variable are 'Hindu', 'Muslim', 'Christian', 'Sikh', 'Buddhist', 'Jain', 'Others', and 'No Religion'. We apply these household responses to each individual in the household. Table S3 shows the breakdown of the total sample of individuals by answers to these two questions. For simplicity, for the $religion$ variable, we collapse 'Christian', 'Sikh', 'Buddhist', 'Jain', 'Others', and 'No Religion' into the category of Other Answers in this table. Using the responses to these two questions, we classify individuals into five caste and religion categories used in the analysis: Adivasi, Dalit, OBC/high-caste Hindu, Muslim, or Other.

Our assignment strategy uses the following approach. Irrespective of the response to the $religion$ question, individuals in 'ST' households are assigned to the Adivasi category, and individuals in 'SC' households are assigned to the Dalit category. Individuals are assigned to the Muslim category if they are in households for which the response to the $religion$ question is 'Muslim', and the response to the $social\_group$ question is 'Other' Social Group. A small proportion of the total sample identifies as 'Muslim' and 'ST', or 'Muslim' and 'SC'. Following Desai et. al. (2), we classify these households as Adivasis and Dalits, respectively. Similarly, individuals are assigned to the OBC/high-caste Hindu category if they are in households for which the response to the $religion$ question is 'Hindu', and the response to the $social\_group$ question is 'Other' Social Group. Finally, the Other caste and religion category (used in the analysis) consists of individuals in households that are recorded to have Other Answers (i.e. not 'Hindu', 'Muslim', or missing) to the $religion$ question *and* 'Other' Social Group (i.e. not 'SC', 'ST', or missing) to the $social\_group$ question. We do not estimate life expectancy for the Other caste and religion category because of its small sample.

We are unable to assign 0.02% of observations to caste and religion categories because data are missing for $social\_group$ or $religion$. 93% of observations for which $social\_group$ is unknown are from the mortality roster, and leaving them out

---

[‡]We exclude sanitation as an explanatory factor for adult mortality because sanitation is unlikely to have direct adult mortality consequences.

[§]Data are available from this website: https://nrhm-mis.nic.in/hmisreports/AHSReports.aspx.

biases mortality estimates downwards. For our main analysis, we randomly assign these observations to caste and religion categories based on the existing distribution of the sample population across the five caste and religion categories. Dropping these observations changes our estimates of disparities very little.

Although OBCs are not as privileged in Indian society as high-caste Hindus, they were not considered untouchable, as Dalits were. As we note in the main text, we are unable to separate OBCs and high-caste Hindus. This makes our estimates for disparities conservative.

Socioeconomic status variables. Following Filmer and Pritchett (3), we construct a wealth index through principal component analysis of household ownership of assets (including radio, television, landline, mobile phone, washing machine, refrigerator, sewing machine, bicycle, motorcycle, car, tractor, water pump) and house infrastructure (number of rooms in the house, cooking location, and materials used to construct house). For some analyses we divide the wealth index into deciles, and for others, we divide the index into quintiles. We separately control for land ownership as another measure of wealth. For environmental exposures, we use indicators for household open defecation and use of solid fuels for cooking, as well as PSU-level fraction defecating in the open and cooking on solid fuels.

Comparison of AHS mortality rates with SRS and NFHS. In addition to the AHS, we also use reported summary mortality data from India's Sample Registration System (SRS) for the years 2007 through 2009, and estimated age-specific mortality rates from India's Demographic and Health Survey 2015-2016, also called the National Family Health Survey (NFHS).[¶] Across all nine states, the AHS visited over 20,000 primary sampling units (PSU)[‖] and over four million households. The survey collected data on over 20 million individuals. In comparison, the SRS produced estimates of age-specific mortality rates and life expectancies from a nationally representative sample of eight million individuals across most Indian states (4). Similarly, the National Family Health Survey (NFHS) collected data from approximately 600,000 households in 2015-2016 (5) across all Indian states. The AHS is larger than these other national surveys, and was in fact one of the largest data collection exercises in the world.

The main purpose of referencing these alternative datasets is to show that AHS mortality rates match closely to those from the SRS and NFHS. The AHS is the primary data source for this paper because it contains data on social group, which the SRS does not, and is adequately large to study differences in mortality based on social group and SES, which the NFHS is not.

Estimating person-years contributed to single-year ages. For individuals who were alive in January 2010, we estimate exact age using the date of birth provided in the household roster. For household members that died between January 2007 and December 2009, we estimate exact age at death and date of death using the data provided for each of these fields on the mortality roster.

Because the day of the month for both date of birth and date of death is missing for the majority of observations, we randomly assign a day of the month for all records. Some observations are also missing month of the year. Month of birth is missing for 45% of individuals alive on January 2010, and month of death is missing for 1% of individuals who died between January 2007 and December 2009. For observations for which month is missing, we randomly assign a month of the year.

Using data on day, month, and year of birth, we estimate the exact age of individuals who were alive on January 2010, and the number of person-years each individual contributed to single-year ages during the period January 2007 to December 2009. For individuals who died between January 2007 and December 2009, we use data from the fields on age at death, and data on day, month, and year of death, to estimate the number of person-years that these individuals contributed to single-year ages during the study period before they died.

Identifying Primary Sampling Units (PSUs). Primary sample units (PSUs) in the AHS are Census Enumeration Blocks (CEBs) in urban areas and villages in rural areas. Based on conversations with a statistician in the Government of India's Ministry of Statistics and Programme Implementation (MOSPI), we learned that for most records, PSU identifiers can be extracted from *fid* or *fidx* variables. Records either have a value under *fid* or a value under *fidx*. If the record has a value for *fid*, PSU is the three digit number starting at the fifth digit of the 12 or 13 digit number. If the record has a value for *fidx*, PSU is the last three digits of the 11 or 12 digit number.

For records that do not have a value for *fid* or *fidx*, PSU can be extracted using a third method. Because data were sorted by *state* and PSU before assigning the variable *psu id*, sorting the data by *state* and *psu id* can help in the process of backing out PSU for the remaining records. We sort the data by these two variables and assign records to PSUs based on the PSU of the surrounding records. If records that are above have a different PSU than records that are below an observation for which PSU is unknown, we randomly assign to the observation the PSU of one of the surrounding PSUs. Random PSU assignment occurs for less than 0.05% of observations.

Description of appendix figures and tables

Appendix Figure S1 shows the distribution of social groups by wealth deciles. We note that all deciles have representation from each of the four social groups. As expected, marginalized social groups are more likely to be in poorer deciles.

[¶] Additional details on these three datasets, the deaths questions asked, and the estimation processes are available in Appendix Table S1.

[‖] Sample units in the AHS are Census Enumeration Blocks (CEBs) in urban areas, and villages in rural areas.

Sangita Vyas, Payal Hathi, Aashish Gupta

164 **Appendix Figure S2** shows age-specific mortality rates for the four social groups. Muslims have lower or similar mortality
165 rates compared to OBC/high-caste Hindus until about age 50. After age 50, OBC/high-caste Hindus have lower mortality
166 rates. Dalits and Adivasis have higher mortality rates across the life course.

168 **Appendix Figure S3** shows life expectancy estimates based on a) sex-specific $_na_x$ values borrowed from the SRS; and b) sex-
169 and group-specific $_na_x$ values calculated from the AHS data. Life expectancy estimates from empirical $_na_x$ values are lower
170 than life expectancy estimates from borrowed $_na_x$ values. This is because empirical $_na_x$ are influenced by both age-composition
171 of the dead individuals, as well as age-misreporting of age at death (1). They are biased downwards. Life expectancy disparities
172 are similar in both sets of estimates.

174 **Appendix Figure S4** shows unadjusted mortality rates and predicted Gompertz mortality rates. Predicted mortality is
175 similar to unadjusted mortality. Gaps between unadjusted and predicted mortality rates are slightly larger for marginalized
176 social groups than for OBC/high-caste Hindus. This is likely because of greater age-misreporting among marginalized social
177 groups and suggests that the unadjusted life expectancy gaps reported in this study are conservative.

179 **Appendix Figure S5** shows female and male life expectancies at age 15 for the four social groups we study. Confidence
180 intervals are calculated using a cluster-bootstrap approach. The overall life expectancy patterns are similar to female and male
181 life expectancies at birth for the four social groups. Compared to OBC/high-caste Hindus, we observe lower life expectancies at
182 age 15 among marginalized social groups.

184 **Appendix Figure S6** shows life expectancy by social group and household wealth decile, using deciles created at the state level.
185 Similar to using a nine-state wealth index, we find that Adivasis and Dalits have lower life expectancies than OBC/high-caste
186 Hindus across wealth categories. Muslims and OBC/high-caste Hindus have similar life expectancies at lower wealth deciles,
187 but Muslims have lower life expectancies at higher wealth deciles.

189 **Appendix Figure S7** shows the fraction of the within-wealth-decile sample comprised by each social group. Compared to
190 poorer deciles, richer deciles have greater representation of high-caste Hindus.

192 **Appendix Table S1** describes the AHS, SRS, and NFHS surveys, and how each survey measured mortality within households.

194 **Appendix Table S2** provides a description of the survey sample. Across all nine states, the AHS visited over 20,000 primary
195 sampling units (PSU) and over four million households. The survey collected data on over 20 million individuals. Some states
196 have small populations of certain marginalized groups. For instance, Adivasis represent a small fraction of the population
197 in Uttarakhand, Uttar Pradesh, and Bihar, and Muslims represent a small fraction of the population in Odisha and Chhattisgarh.

199 **Appendix Table S3** shows how individuals were classified to one of five caste and religion categories based on household
200 responses to questions on *social_group* and *religion* in the survey.

202 **Appendix Table S4** shows summary statistics by social group. Panel A shows entries and exits in the AHS sample between
203 January 2007 and December 2009. More than 19 million individuals were alive in January 2007. Over the three-year period,
204 there were almost 1.4 million births and 446,085 deaths in sample households. More than 20 million individuals were alive at
205 any point in this study period, contributing over 59 million person-years.
206   Panel B of Table S4 presents characteristics of sample individuals alive at any point between January 2007 and December
207 2009. Of this sample, 2.2% of individuals died in the period. Relative to the fraction of OBC/high-caste Hindu individuals
208 in the sample that died, a higher fraction of Adivasis and Dalits in the sample died, and a lower fraction of Muslims in the
209 sample died. 48.5% of the sample were female, and the average age was 23.9. The next set of characteristics in Panel B reflect
210 SES. With the exceptions of Muslim advantage in urban residence, owning sewing machines, and sanitation, and Adivasi
211 advantage in land ownership, this set of characteristics reflects that Adivasis, Dalits, and Muslims are disadvantaged across
212 various socioeconomic indicators. Adivasis, Dalits, and Muslims own fewer assets, live in poorer quality homes, and are less
213 likely to cook using clean fuels.

215 **Appendix Table S5** shows results from linear regressions predicting the probability of death of children under age five. The
216 regression results motivate the variables included in the decomposition. Column 4 includes all of the SES variables. Rural
217 residence, wealth decile, land ownership, PSU and household open defecation, and household solid fuel use are predictive of
218 child mortality. In order to maintain a manageable number of bins in the re-weighting exercise, we use PSU open defecation
219 only, rather than both household and PSU open defecation variables, since we hypothesize that there are important externalities
220 associated with open defecation.

222 **Appendix Table S6** shows results from linear regressions predicting the probability of death of individuals age five and older.
223 The regression results motivate the variables included in the decomposition. Column 4 includes all of the SES variables. Rural
224 residence, wealth decile, land ownership, and household solid fuel use are predictive of mortality.
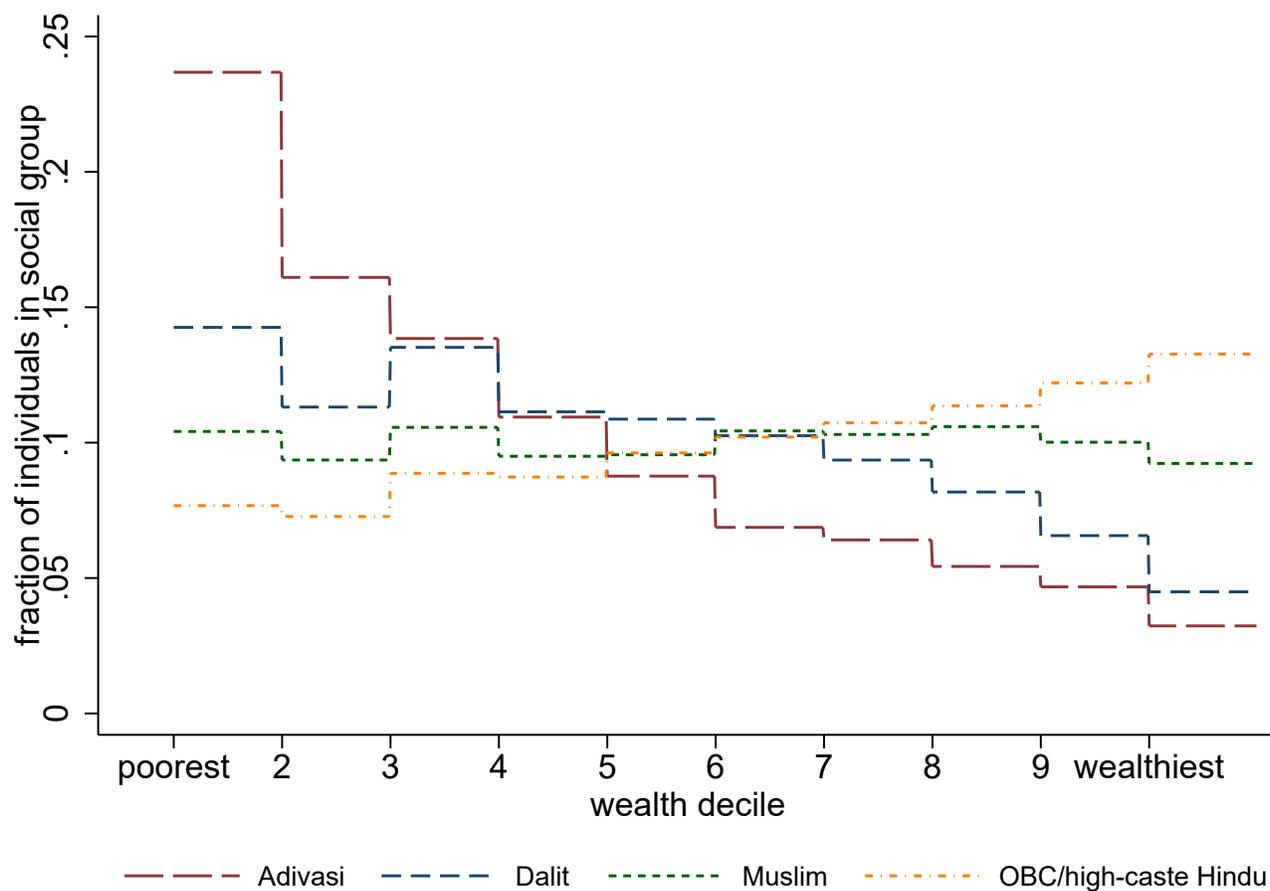
Fig. S1. Distribution of social groups across wealth deciles. Wealth decile are deciles of a wealth index constructed using a principal component analysis of household assets and house infrastructure, including radio, television, landline, mobile phone, washing machine, refrigerator, sewing machine, bicycle, motorcycle, car, tractor, water pump, number of rooms in house, presence of a kitchen, and materials of house. Estimates use sample weights. Source: Annual Health Survey 2010-2011.
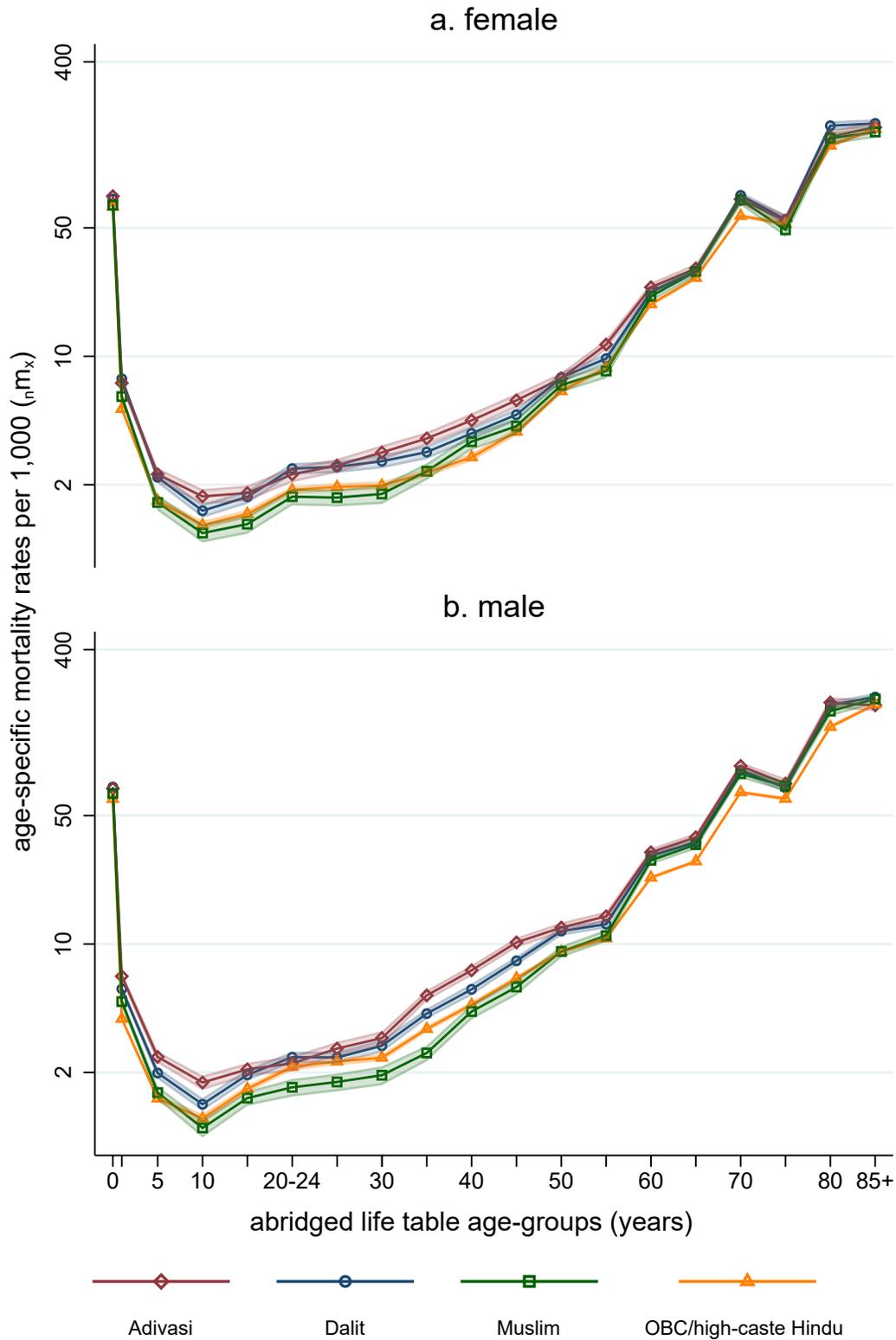
Fig. S2. Age-specific mortality rates by sex and social group from AHS data. Age-specific mortality rates are based on deaths and person-years lived during the period Jan 2007 to Dec 2009, estimated from information provided in the mortality and household rosters. Panel (a) shows estimates for females, Panel (b) shows estimates for males. Estimates use sample weights. The shaded areas around the lines represent 95% confidence intervals calculated using a cluster-bootstrap procedure. Source: Annual Health Survey 2010-2011.

a) $_na_x$ values borrowed from SRS

female

male

life expectancy at birth: e_0 (years)

	Adivasi	Dalit	Muslim	OBC/high-caste Hindu
female	62.8 [62.4-63.2]	63.3 [63.0-63.5]	65.7 [65.4-66.1]	66.5 [66.4-66.7]
male	60.0 [59.7-60.3]	61.3 [61.1-61.6]	63.8 [63.4-64.1]	64.9 [64.8-65.1]

b) $_na_x$ values calculated from AHS data

female

male

life expectancy at birth: e_0 (years)

	Adivasi	Dalit	Muslim	OBC/high-caste Hindu
female	62.3 [62.0-62.7]	62.8 [62.5-63.1]	65.3 [64.9-65.6]	66.1 [65.9-66.2]
male	59.5 [59.2-59.8]	60.8 [60.6-61.1]	63.3 [63.0-63.6]	64.4 [64.3-64.6]
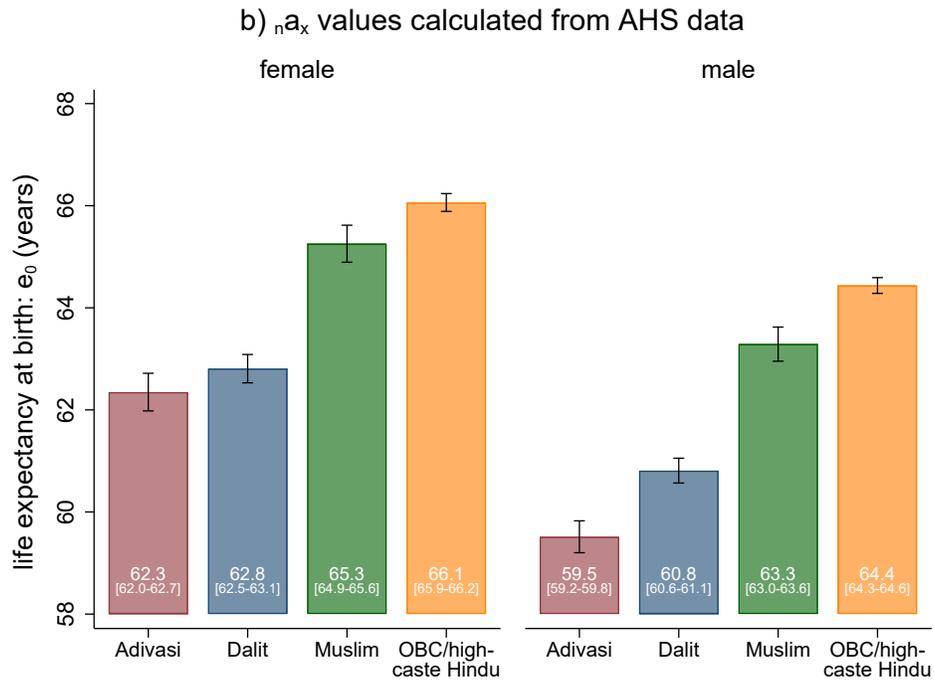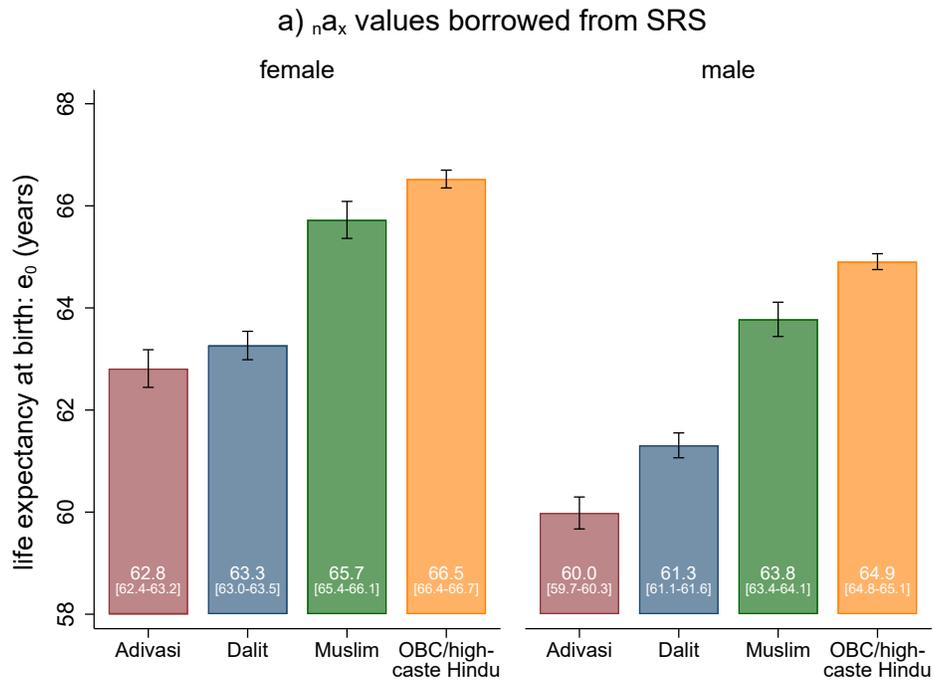
Fig. S3. Comparison of disparities in life expectancy when using $_na_x$ values borrowed from the SRS and $_na_x$ values estimated from AHS data. Life expectancy estimates for each sex and social group are calculated using standard life table procedures. Estimates for females are on the left, estimates for males are on the right. Estimates use sample weights. The vertical lines around each estimate represent 95% confidence intervals calculated using a cluster-bootstrap procedure. Estimates in Panel (a) are based on $_na_x$ values estimated from the Sample Registration System. We calculate sex- and group-specific $_na_x$ values from state-specific SRS $_na_x$ values. $_na_x$ values for each social group weight the state-specific $_na_x$ values from the SRS according to the distribution of social groups across states. Estimates in Panel (b) use sex- and group-specific $_na_x$ values calculated from the AHS. Source: Sample Registration System 2007-11 and Annual Health Survey 2010-2011.

Sangita Vyas, Payal Hathi, Aashish Gupta
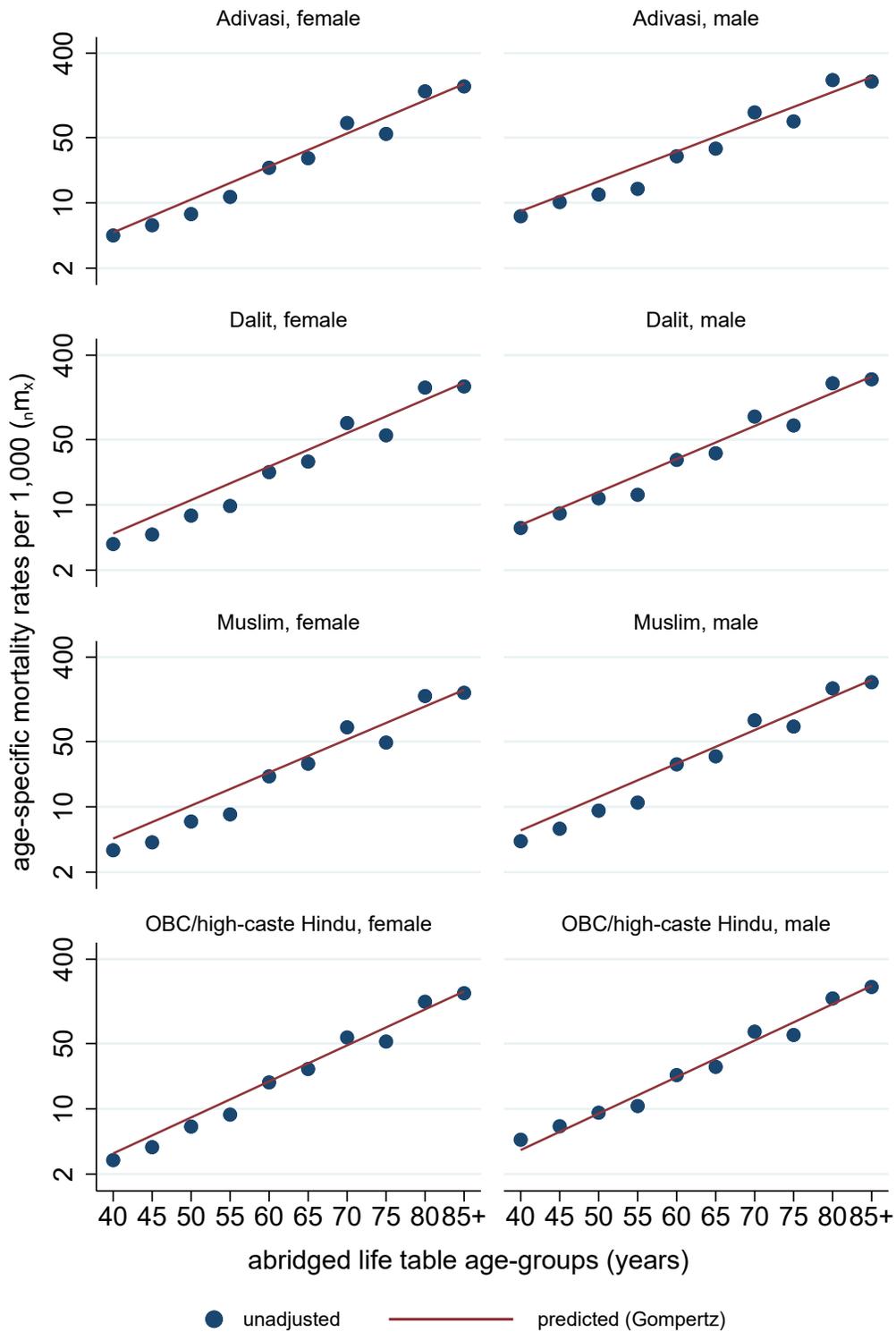
Fig. S4. Actual and Gompertz mortality rates estimated from AHS data. Estimates for females are on the left, estimates for males are on the right. Estimates use sample weights. Unadjusted age-specific mortality rates are based on deaths and person-years lived during the period Jan 2007 to Dec 2009, estimated from information provided in the mortality and household rosters. Gompertz mortality rates are estimated using a non-linear least squares model of the form: $m_x = \alpha \times e^{\beta x}$, where x represents age in five-year groups starting at age 40, with the final age interval as 85+ (6). Source: Annual Health Survey 2010-2011.

Sangita Vyas, Payal Hathi, Aashish Gupta
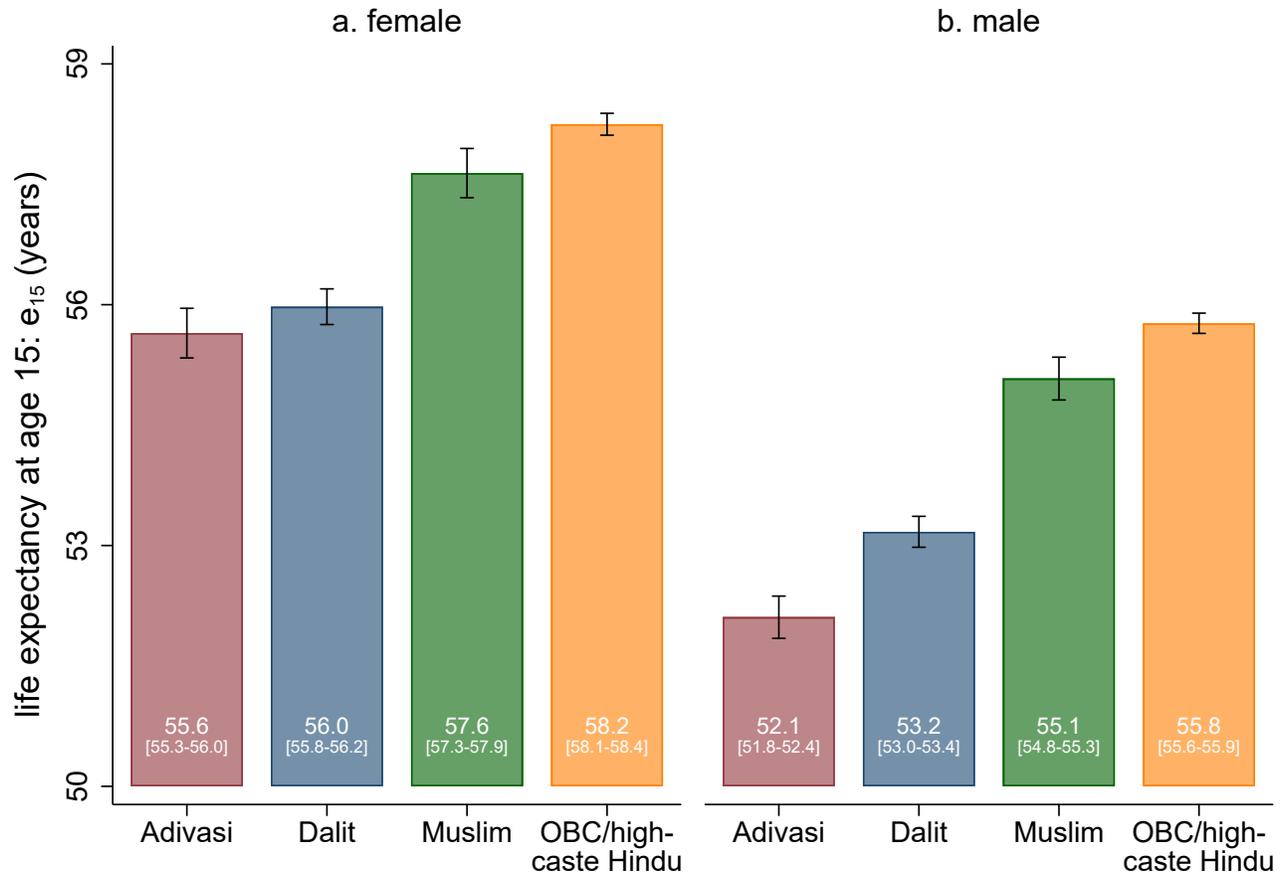
Fig. S5. Lower life expectancy at age 15 among marginalized groups compared to OBC/high-caste Hindus. Life expectancy estimates for each sex and social group are calculated using standard life table procedures. Panel (a) shows estimates for females, Panel (b) shows estimates for males. Estimates use sample weights. The vertical lines around each estimate represent 95% confidence intervals calculated using a cluster-bootstrap procedure. Source: Annual Health Survey 2010-2011.

Sangita Vyas, Payal Hathi, Aashish Gupta
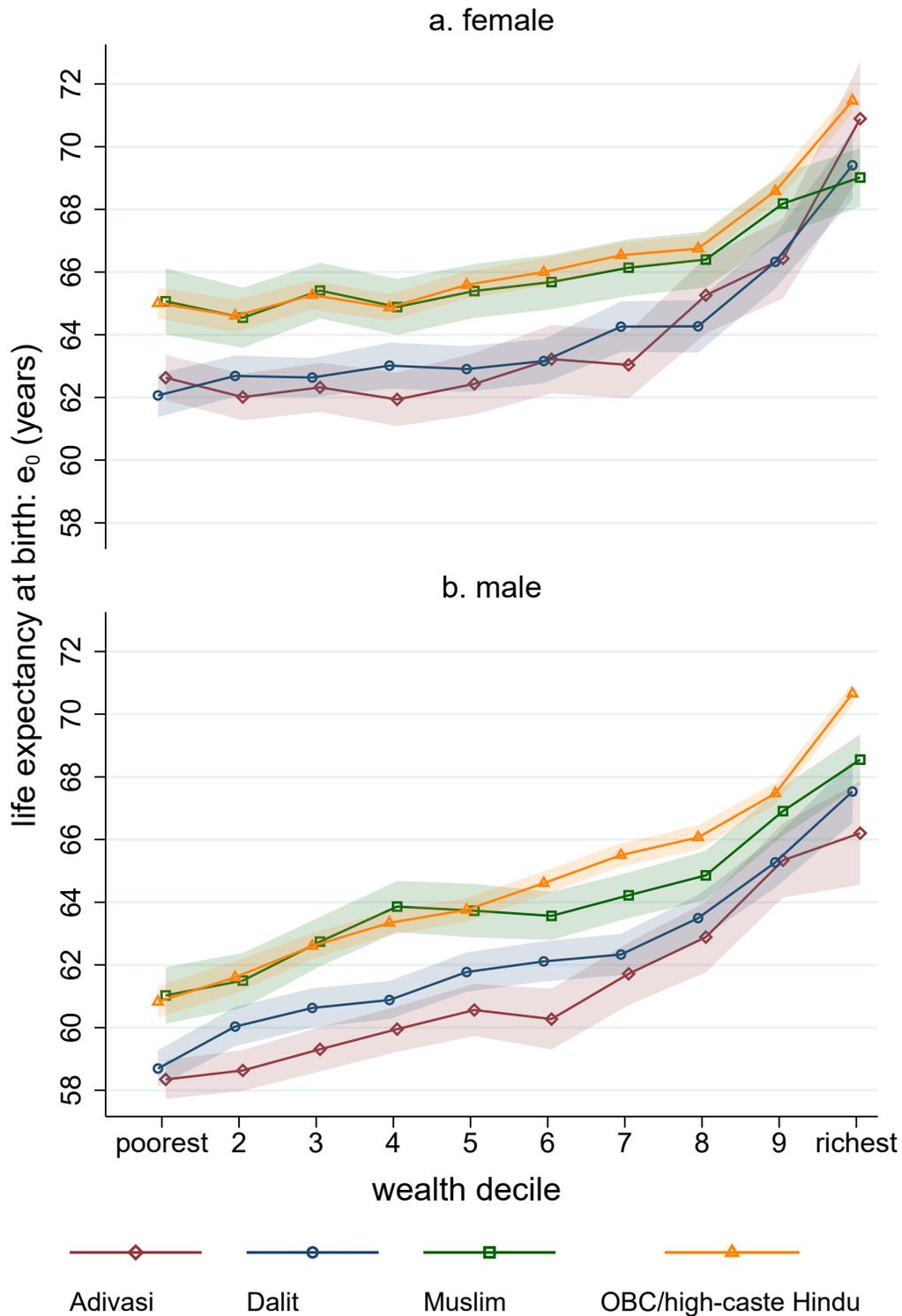
Fig. S6. Life expectancy by household wealth, using state-specific wealth index: Lower life expectancy at birth among Dalits and Adivasis at all levels of wealth. Wealth decile are deciles of a wealth index constructed using a principal component analysis, for each state, of household assets and house infrastructure. Life expectancy estimates for each wealth decile, sex, and social group are calculated using standard life table procedures. Panel (a) shows estimates for females, Panel (b) shows estimates for males. Estimates use sample weights. The shaded areas around the lines represent 95% confidence intervals calculated using a cluster-bootstrap procedure. Source: Annual Health Survey 2010-2011.
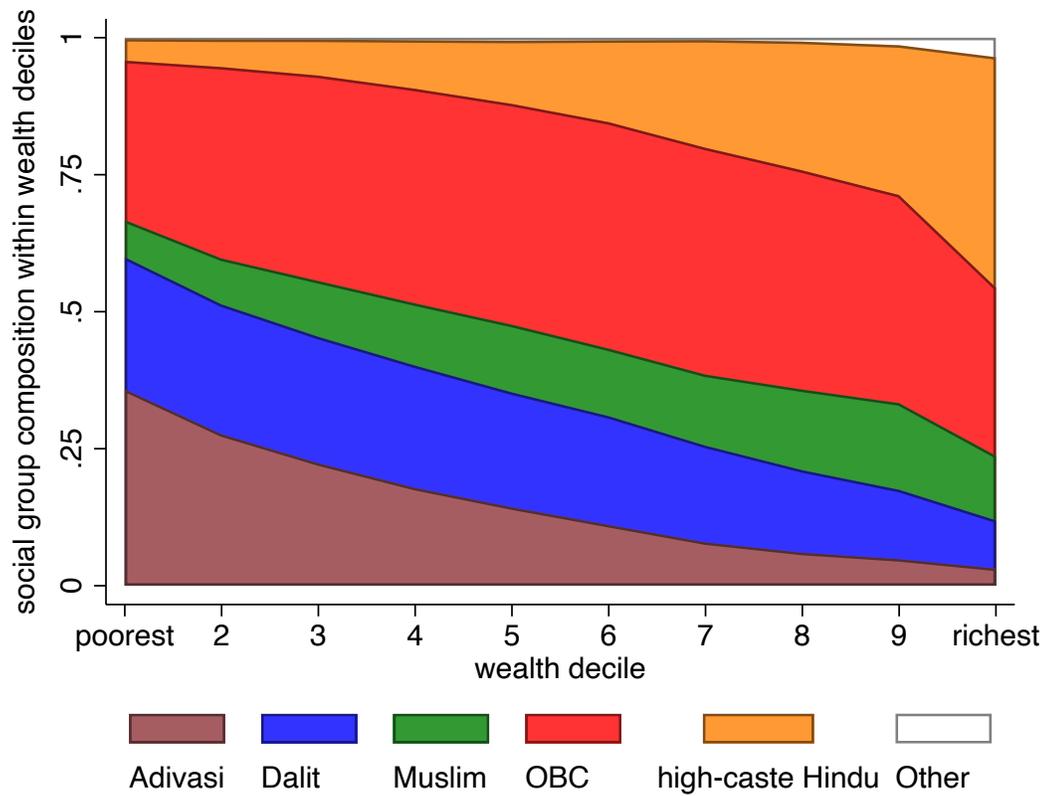
Fig. S7. Compared to poorer deciles, richer deciles have greater representation of high-caste Hindus. The figure shows the fraction of the within-wealth-decile sample comprised by each social group. Our strategy for constructing social groups in the NFHS follows the same approach we use for the AHS (see section 'Supplementary Information: Data preparation'). The NFHS is able to separately identify OBC Hindus and high-caste Hindus. Wealth decile is deciles of a wealth index constructed by the NFHS, based on durable goods and assets that a household owns. We restrict the full NFHS sample to the nine AHS states. Source: National Family Health Survey 2015-2016.

Sangita Vyas, Payal Hathi, Aashish Gupta

Table S1. Data descriptions: AHS, SRS, and NFHS

Data source (1)	Survey design (2)	Death questions used (3)	Sample size (4)	Reference period (5)
AHS	sample survey: multi-stage stratified random sampling	household deaths question, instruction to enumerators: "For deaths occurred to the usual residents of the sample unit during January 1, 2007 to December 31, 2009. Still births not to be included." No further instruction beyond this.	4.1 million households in 9 states	Jan 2007 - Dec 2009
SRS	dual record: complete enumeration of births and deaths in a panel of villages, and semi-annual independent survey	we use published state-level age-specific mortality rates for years 2007, 2008, 2009	1.3 million households nationally	2007, 2008, 2009
NFHS	sample survey: multi-stage stratified random sampling	for individuals age 5 and older: household deaths question, "Did any usual member of this household die since Jan 2012?" for children under age 5: birth history recording information on all live births	0.6 million households nationally	Jan 2012 - Dec 2016

Mortality and life expectancy estimates generated by the SRS are used by both the UN-Inter-Agency Group for Child Mortality Estimation (UN-IGME) and the WHO Life Tables (7). Following (8), in analyses that use data from the NFHS, we use the birth history to estimate child mortality and the mortality roster for estimating mortality at ages five and older. Sources: Annual Health Survey 2010-2011; Sample Registration System 2007, 2008, 2009; and National Family Health Survey 2015-2016.

Table S2. Description of Annual Health Survey sample

	Uttarakhand	Rajasthan	Uttar Pradesh	Bihar	Assam	Jharkhand	Odisha	Chhattisgarh	Madhya Pradesh	All states
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Panel A: Sample PSUs and households										
# of PSUs	2,501	1,842	3,927	2,356	1,784	2,109	2,365	1,255	2,557	20,696
# of households	367,177	352,744	847,675	594,834	383,413	378,919	457,725	273,609	494,616	4,150,712
PSU mean households	147	192	216	252	215	180	194	218	193	201
Panel B: Sample individuals alive at any time between Jan 2007 and Dec 2009										
Adivasi	0.048	0.145	0.017	0.022	0.129	0.250	0.214	0.306	0.209	0.102
Dalit	0.214	0.173	0.232	0.183	0.098	0.136	0.186	0.128	0.166	0.188
Muslim	0.125	0.094	0.190	0.168	0.313	0.155	0.020	0.017	0.067	0.144
OBC/high-caste Hindu	0.591	0.570	0.557	0.626	0.437	0.453	0.573	0.542	0.547	0.558
Other	0.022	0.018	0.005	0.001	0.022	0.006	0.008	0.007	0.010	0.008
n	1,652,464	1,826,710	4,662,156	3,155,938	1,772,163	1,960,716	1,981,435	1,246,948	2,353,231	20,611,761

Estimates use sample weights. PSU is primary sampling unit, which are Census Enumeration Blocks (CEBs) in urban areas, and villages in rural areas. Source: Annual Health Survey 2010-2011.

Sangita Vyas, Payal Hathi, Aashish Gupta

Table S3. Social group composition

| | | social_group | | | |
	'ST'	'SC'	'Other' Social Group	Missing Response	Total
religion					
'Hindu'	1,877,130	3,791,210	11,489,094	48	17,157,481
	9.11%	18.39%	55.74%	0.00%	83.24%
	(Adivasi)	(Dalit)	(OBC/high-caste Hindu)	(Randomly assigned)	
'Muslim'	46,752	48,357	2,976,676	5	3,071,790
	0.23%	0.23%	14.44%	0.00%	14.9%
	(Adivasi)	(Dalit)	(Muslim)	(Randomly assigned)	
Other Answers	184,509	31,779	161,220	2	377,509
	0.9%	0.15%	0.78%	0.00%	1.83%
	(Adivasi)	(Dalit)	(Other)	(Randomly assigned)	
Missing Response	0	2	7	4,971	4,981
	0.00%	0.00%	0.00%	0.02%	0.02%
	(Adivasi)	(Dalit)	(Randomly assigned)	(Randomly assigned)	
Total	2,108,390	3,871,348	14,626,997	5,026	20,611,761
	10.23%	18.78%	70.96%	0.02%	100%

This table shows the breakdown of the total sample by answers to the separate questions in the AHS on *social_group* and *religion*. The first number in each cell is the number of individuals and the percentage is the fraction of the total sample, in each category. The categorization into Adivasi, Dalit, Muslim, OBC/high-caste Hindu, and Other, shown in parenthesis are our assignment to different caste and religion categories used in the analysis. Randomly assigned indicates that individuals were randomly assigned to one of the five caste and religion categories used in the analysis. Estimates use sample weights. Source: Annual Health Survey 2010-2011.

Table S4. Summary statistics by social group: socioeconomic disadvantage for marginalized social groups

	Adivasi	Dalit	Muslim	OBC/high-caste Hindu	Other	All
	(1)	(2)	(3)	(4)	(5)	(6)

Panel A: Sample entries and exits between Jan 2007 and Dec 2009

	Adivasi	Dalit	Muslim	OBC/high-caste Hindu	Other	All
n (alive in Jan 2007)	2,321,293	3,410,534	2,423,716	10,916,863	165,691	19,238,097
births during period	175,871	266,052	196,687	726,854	9,644	1,375,108
deaths during period	59,775	85,842	50,382	246,425	3,661	446,085
n (alive at any time during period)	2,496,993	3,676,303	2,620,191	11,642,953	175,321	20,611,761
n person years	7,152,510	10,517,483	7,506,547	33,506,363	506,507	59,189,409

Panel B: Individuals alive at any time between Jan 2007 and Dec 2009

	Adivasi	Dalit	Muslim	OBC/high-caste Hindu	Other	All
died between Jan 2007 and Dec 2009	0.024	0.024	0.020	0.022	0.021	0.022
female	0.490	0.482	0.490	0.484	0.480	0.485
age (on Jan 2007)	23.269	22.884	21.357	24.914	27.643	23.872
household:						
in rural area	0.900	0.822	0.685	0.767	0.550	0.779
in bottom 40% of wealth index	0.673	0.541	0.439	0.359	0.228	0.437
owns land	0.633	0.461	0.336	0.545	0.336	0.510
has radio	0.132	0.155	0.182	0.224	0.291	0.196
has television	0.159	0.226	0.264	0.356	0.591	0.300
has computer	0.037	0.048	0.061	0.080	0.149	0.067
has phone	0.360	0.542	0.641	0.690	0.857	0.621
has washer	0.015	0.018	0.050	0.057	0.222	0.046
has refrigerator	0.036	0.051	0.110	0.129	0.401	0.103
has sewing machine	0.070	0.105	0.213	0.196	0.402	0.169
has bicycle	0.470	0.506	0.497	0.557	0.536	0.530
has motorcycle	0.088	0.093	0.137	0.203	0.385	0.162
has car	0.015	0.014	0.023	0.042	0.126	0.032
has tractor	0.012	0.011	0.015	0.032	0.084	0.024
has water pump	0.051	0.054	0.056	0.108	0.137	0.085
has >1 room in dwelling	0.544	0.467	0.486	0.624	0.755	0.568
has a kitchen	0.454	0.343	0.506	0.508	0.769	0.472
is kaccha	0.627	0.350	0.241	0.247	0.176	0.306
uses solid fuel for cooking	0.933	0.895	0.803	0.785	0.530	0.823
in PSU with >50% solid fuel use	0.940	0.900	0.815	0.822	0.604	0.847
defecates in the open	0.834	0.768	0.444	0.614	0.272	0.643
in PSU with >50% open defecation	0.851	0.780	0.482	0.694	0.333	0.697

Sample entries and exits are estimated from information provided in the mortality and household rosters. Wealth decile are deciles of a wealth index constructed using a principal component analysis of household assets and house infrastructure, including radio, television, landline, mobile phone, washing machine, refrigerator, sewing machine, bicycle, motorcycle, car, tractor, water pump, number of rooms in house, presence of a kitchen, and materials of house. Kaccha means households built using mud, straw, bamboo, and other temporary materials. Fraction open defecation and fraction solid fuel use in the PSU represent the PSU-level fraction of individuals defecating in the open and eating food cooked on solid fuels. Panel A shows raw unweighted counts. Estimates in Panel B use sample weights. Source: Annual Health Survey 2010-2011.

Sangita Vyas, Payal Hathi, Aashish Gupta

Table S5. Regressions predicting death, for individuals < age five on Jan 2007

Dependent variable:	Probability of death (per thousand) between Jan 2007 and Dec 2009			
	(1)	(2)	(3)	(4)
Social group (omitted category is OBC/high-caste Hindu)				
Dalit	6.463**	6.119**	4.721**	4.280**
	(0.451)	(0.444)	(0.462)	(0.469)
Adivasi	7.459**	6.341**	3.693**	3.426**
	(0.479)	(0.478)	(0.473)	(0.474)
Muslim	2.765**	3.734**	2.973**	3.961**
	(0.494)	(0.488)	(0.452)	(0.436)
Rural		10.87**	6.661**	1.848*
		(0.743)	(0.777)	(0.829)
Wealth decile			-1.452**	-1.008**
			(0.0530)	(0.0612)
Household owns land			1.831**	0.884**
			(0.323)	(0.325)
Fraction open defecation in PSU				2.681**
				(0.983)
Fraction solid fuel in PSU				1.057
				(1.317)
Household defecates in the open				4.403**
				(0.351)
Household uses solid fuel				4.170**
				(0.535)
Potential exposure in months	x	x	x	x
Age × sex fixed effects	x	x	x	x
N	3,709,562	3,709,562	3,709,562	3,709,562
Mean of dependent variable	32.398	32.398	32.398	32.398
Adjusted R-squared	0.0202	0.0208	0.0212	0.0214

Wealth decile are deciles of a wealth index constructed using a principal component analysis of household assets and house structure, including radio, television, landline, mobile phone, washing machine, refrigerator, sewing machine, bicycle, motorcycle, car, tractor, water pump, number of rooms in house, presence of a kitchen, and materials of house. Fraction open defecation and fraction solid fuel use in the PSU represent the PSU-level fraction of individuals defecating in the open and eating food cooked on solid fuels. Potential exposure in months is the total number of months the child could have lived from Jan 2007 through Dec 2009. For children who were born before Jan 2007, potential exposure is 36 months. For children born after Jan 2007, potential exposure is the number of months the child could have lived before Dec 2009. Age groups are abridged life table age-groups (years). Regressions use sample weights and cluster standard errors by PSU. Source: Annual Health Survey 2010-2011.

Table S6. Regressions predicting death, for individuals $\geq$ age five on Jan 2007

Dependent variable:	Probability of death (per thousand) between Jan 2007 and Dec 2009			
	(1)	(2)	(3)	(4)
Social group (omitted category is OBC/high-caste Hindu)				
Dalit	3.417**	3.363**	2.698**	2.596**
	(0.148)	(0.149)	(0.142)	(0.138)
Adivasi	4.416**	4.249**	3.345**	3.332**
	(0.237)	(0.235)	(0.237)	(0.235)
Muslim	0.689**	0.830**	0.350*	0.251
	(0.160)	(0.164)	(0.166)	(0.165)
Rural		1.293**	0.391	-0.506+
		(0.227)	(0.249)	(0.283)
Wealth decile			-0.502**	-0.413**
			(0.0244)	(0.0234)
Household owns land			-0.609**	-0.797**
			(0.123)	(0.120)
Fraction solid fuel in PSU				-0.117
				(0.403)
Household uses solid fuel				2.400**
				(0.159)
Age $\times$ sex fixed effects	x	x	x	x
N	16,886,515	16,886,515	16,886,515	16,886,515
Mean of dependent variable	19.428	19.428	19.428	19.428
Adjusted R-squared	0.0914	0.0915	0.0916	0.0916

Wealth decile are deciles of a wealth index constructed using a principal component analysis of household assets and house infrastructure, including radio, television, landline, mobile phone, washing machine, refrigerator, sewing machine, bicycle, motorcycle, car, tractor, water pump, number of rooms in house, presence of a kitchen, and materials of house. Fraction open defecation and fraction solid fuel use in the PSU represent the PSU-level fraction of individuals defecating in the open and eating food cooked on solid fuels. Age groups are abridged life table age-groups (years). Regressions use sample weights and cluster standard errors by PSU. Source: Annual Health Survey 2010-2011.

References

1. S Preston, P Heuveline, M Guillot, Demography: measuring and modeling population processes. 2001. *Malden, MA: Blackwell Publ.* (2000).
2. SB Desai, et al., Human development in india. *New York: Oxf. Univ.* (2010).
3. D Filmer, LH Pritchett, Estimating wealth effects without expenditure data—or tears: an application to educational enrollments in states of india. *Demography* **38**, 115–132 (2001).
4. Registrar General and Census Commissioner of India, *Compendium of India's Fertility and Mortality Indicators ,1971 - 2013.* (Government of India), (2013).
5. International Institute for Population Sciences - IIPS/India and ICF., *India National Family Health Survey NFHS-4 2015-16.* (Mumbai, India: IIPS and ICF), (2017).
6. TH Tai, A Noymer, Models for estimating empirical gompertz mortality: With an application to evolution of the gompertzian slope. *Popul. Ecol.* **60**, 171–184 (2018).
7. UN-IGME, Levels and trends in child mortality: 2017 report, (UNICEF-WHO-WB-UNPD. http://www. childmortality. org. Accessed 22 Apr), Technical report (2018).
8. A Gupta, N Sudharsanan, Large and persistent life expectancy disparities among India's social groups. (2020).