Distribution of self-reported health in India: The role of income and geography

Ila Patnaik[1], Renuka Sane[2], Ajay Shah[*3], and S. V. Subramaniam[4]

[1,2]*NIPFP*
[3]*xKDR Forum*
[4]*Harvard University*

1st October 2021

[*]Corresponding author, ajayshah@mayin.org

Abstract

Background: We obtain evidence on self-reported health in India using a new large-scale survey database.

Methods: We report summary statistics about the self-reported ill-health rate, and explore relationships with socio-economic parameters through logistic regressions.

Results: The overall average ill health rate is 3.25%. The most important correlates are age, income and location. We find substantial variation across the 102 'homogeneous regions' of the country. Higher income is correlated with better health in 40% of India.

Conclusions: The maps of ill health seen here diverge from conventional wisdom about North vs. South India. Epidemiological studies are required in the hotspots of ill-health and the regions where higher income does not correlate with improved health.

Keywords: Self-reported health, India, geographical variation, SES and health.

Contents

1 Background

Health policy is ultimately about creating conditions in which people are healthy. The wellness of the people is the outcome of interest. Many plausible health outcome measures reflect different dimensions of wellness. Death can be accurately measured, which suggests pathways for health outcomes measurement through death rates and longevity. In the Indian health literature, infant and maternal mortality are the dominant measures which have been employed. However, these measures have limitations, particularly in a less developed country like India. Maternal and child mortality represent a narrow subset of the population, and narrow government programs have influenced these metrics without reshaping the broader health of the population. Measures based on mortality in the overall population are impeded by the limitations of official mortality-related statistics.

Medical researchers are able to use objective metrics from medical tests, such as blood pressure, anemia, diabetes etc. as a measure of the health of an individual. Aggregate measures of disease burden include measurement of Disability Adjusted Life Years (DALYs) which captures the reduction in life expectancy and the diminished quality of life. However, such measurement, done consistently across time and space, involves the construction of complex datasets which is infeasible in less developed countries.

One path to measuring the health status of an individual is to to directly ask individuals to assess their own health. This is termed 'self reported health' (SRH). For example, the WHO has used the question : *In general, how would you rate your health today?*. Responses to such a survey question can be binary, which naturally suggests an *ill-health rate* which is the fraction of a given set of people who report that their self reported health is not good. Alternatively, a five-point scale can be used, e.g. *In general, would you say that you health is excellent, very good, good, fair, or poor?*.

When compared with objective measures based on medical tests, SRH measurement is inexpensive and less intrusive. From the viewpoint of the foundations of human welfare, asking a person if they are feeling well is of essence (B. Singh, 2018). At the same time, SRH suffers from four limitations:

1. SRH inevitably introduces a psychological filter in determining whether there is ill health. For example, there appears to be a gender gap in reporting of health (Boerma et al., 2016; Hirve et al., 2010; L. Singh et al., 2013). The under-reporting seems to hold even after controlling for objective health measures (Dasgupta, 2018). These psychological factors might be correlated

with wealth: the tooth ache or fever that makes an upper class person report ill-health might not trigger the same response from a poor person.

2. The precise phrasing of the SRH question in a household survey matters in interpreting the results. Many different designs of the question can be applied, e.g. *Are you feeling well today?* or *Have you been healthy in the last month?* or *In the last week, were you unwell for atleast one day?*. As a consequence, the numerical values of the SRH rate are not comparable across survey datasets.

3. The SRH approach should not be used for specific morbidities. Disorders like diabetes are not discerned by the person for many years. For example, Onur and Velamuri, 2018 find that the self reported incidence of hypertension and lung disease underestimate the disease burden in India. Other studies find that the prevalence of non communicable diseases is under-reported when viewed through SRH, especially among the poor (Banerjee et al., 2004; Vellakkal et al., 2013).

4. Access to sound health care, and particularly testing, is likely to influence perceived health. This could generate more accurate answers in households with more access to health care.

While SRH has flaws, it contains useful information about the health of an individual. A successful empirical literature has utilised SRH in health research. It correlates with objective health outcome measures (Cullati et al., 2018; Wu et al., 2013). It does reasonably well on predicting mortality, especially among elderly populations in developed countries (Benyamini & Idler, 1999; Heistaro et al., 2001; Idler & Benyamini, 1997; Miilunpalo et al., 1997). In India, Cullati et al., 2018 use the 2002 WHS data and find that SRH is a useful measure. Similarly Subramanian et al., 2009 use the NFHS as well as the NSS data and find that the use of self-rated ill health has validity in relationship to socioeconomic status. Wu et al., 2013 finds a correlation between SRH and disease burden in China.

As a consequence, over the years, SRH measurement has emerged in many survey datasets, such as *World Health Survey 2002 and SAGE 2007* conducted by WHO, the *Health and Retirement study (HRS)* and the *National Health and Nutrition Examination Survey (NHANES)* in the US, the *Survey of Health, Ageing and Retirement in Europe (SHARE)*, the *English Longitudinal Study of Ageing (ELSA)* in the UK.[1] Each of these datasets has

[1]The HRS largely samples the elderly, while the NHANES samples adults and children, and they have used a five-point scale with the categories excellent, very good, good, fair and poor. The ELSA uses a slightly different five point measure: very good, good, fair,

resulted in evidence about SRH which has fed into a substantial downstream literature.

In this paper, we exploit the Consumer Pyramids Household Survey (CPHS), a longitudinal dataset which measures about 170,000 households, three times a year. There is one other panel data set in India is the Indian Human Development Survey (IHDS), where households are observed 10 years apart. While a great deal of the health literature in India is based on the National Sample Survey Organisation (NSSO) and the National Family Health Survey (NFHS) data sets, these are repeated cross-sections, and the existing literature on SRH with these data is relatively limited.

We obtain foundational facts about ill health for individuals in India at one point in time (calendar years 2018 and 2019). We study the variation of ill health with age, and the impact of a variety of socioeconomic factors such as income, gender, caste, religion and education.

We explore the variation of SRH by location. The dataset divides the country into 102 'homogeneous regions' (HRs). In myriad other contexts, the evidence shows substantial heterogeneity within the country, across the HRs. We explore geographical heterogeneity in ill-health, and in the impact of income upon health.

Ultimately, multiple health outcome measures – medical tests, SRH , mortality – need to come together into a rich literature on the causes and consequences of ill health, which can inform the decision making of individuals, health care providers, and public health. The CPHS longitudinal data makes possible a new literature where the causes and consequences of health can be explored. This paper constitutes a first building block for that research program.

2 Methods

2.1 Data sources

We use data from the Consumer Pyramids Household Survey (CPHS) carried out by the Centre for Monitoring Indian Economy (CMIE).[2] The CPHS collects high-quality data through face-to-face interviews with households.

bad and very bad. Another example is the SF-36, a short-form 36-item questionnaire developed by the Rand Corporation for medical outcomes measurement in in the US. It is a patient-reported survey of health as measured along eight multi-item categories.

[2]This is available to all researchers upon payment of a subscription fee.

Answers provided by respondents are captured in a mobile phone through a specially developed app. CPHS measures a panel of households at three points in each year. Households are met with in three waves a year; Wave 1 runs from January - April, Wave 2 from May - August, and Wave 3 from September - December.

The sample is nationally representative and selected through a multistage stratified design.[3] The broadest level of stratification is the "Homogeneous Region (HR)", a set of neighbouring districts that are similar in three dimensions: agro-climatic conditions, urbanisation levels and female literacy. India is partitioned into 102 such HRs in the CPHS. The HR is further divided into stratum – which is either rural (all villages in the HR) or urban (towns which are further classified into four different stratum based on their size) region within a HR.[4] Thus, each HR is further disaggregated into 5 stratums, 1 rural and 4 urban. The primary sampling units (PSUs) in the survey are villages and towns from the 2011 Census of India. The ultimate sampling units are the households from these primary sampling units. CPHS questions are of two broad categories: those that are asked at the household level (such as expenditure, income and asset ownership) and those that are asked of each member of the household (such as demographics, religion, caste, education, and health).

In any narrow period of time, there can be special problems like a local epidemic or a natural disaster, which can generate higher ill-health in one particular region. Some aspects of the disease burden can be seasonal; e.g. reduced water quality in summer or respiratory ailments in North India in the winter. In the year 2017, there is the possibility of an adverse impact upon health of the Demonetisation event of November 2016. In the year 2020, there was the pandemic. Hence, in this paper, we use data from the six waves of 2018 and 2019. Averaging over six waves helps remove isolated episodes of ill health and seasonal factors. The results in this paper can be viewed as a characterisation of baseline conditions in India, against which special periods such as the pandemic can be compared in future research.

The CMIE measurement process thus involves one member – the respondent – who responds on behalf of all members in the household. The text of the

[3] The count of sample households increased from 166,744 households in the January - April, 2014 wave to 174,405 households in the September - December, 2019 wave. Over the six years a total of 65,892 households were added to the original sample of 166,744 households and 58,231 households were dropped for various reasons.

[4] Towns with more than 200,000 households are classified as *Very Large* stratum, between 60,000-200,000 are classified as *Large* stratum, between 20,000-60,000 households are *medium* stratum and below 20,000 households are *small* stratum.

question is *"Does the member feel healthy as of today?"*. To that extent, this measure is not *self*-reported health, but the state of health of an individual, as observed by the survey respondent in the same household. This may help ensure that relatively minor conditions, which are known to the person but not the respondent, do not influence the answer. Similarly, mental health issues could influence SRH as reported by an individual but not the information obtained from an observer in the household.

2.2 Variables of interest

Age We form six discrete age bins: 0-4, 5-9, 10-34, 35-49, 50-59 and 60+. We expect a U-shaped curve where the very young and the very old are more unwell.

Income Income could influence health through four pathways: nutrition, housing quality (which would influence hygiene and physical access for disease vectors), knowledge and health care. In this paper, we seek to measure the overall correlation between income and ill-health, summing across these four factors. In order to avoid endogeneity bias (where ill-health triggers off reduced income), income in month t is expressed as an average of income over the previous 12 months. Household income as reported in CPHS is converted into real rupees using the Consumer Price Index.[5]

Socio-economic characteristics Aggregate facts about the incidence of ill health will be shaped by demographic structure as well as socio-economic characteristics. Towards this, we examine the age-specific SRH rate, as well as the variation in this rate across different social and income categories.

Location We examine how ill health varies by location. This is interesting, in and of itself, as it shows where the unwell persons are. This shows the spatial distribution of health care requirements. In addition, this can potentially yield insights on public health interventions that can improve health.

2.3 Odds ratios of ill health

We estimate a pooled logit model explaining ill-health. Standard errors are clustered at the level of the "primary sampling unit (PSU)".[6] We estimate three models:

1. We first model the log odds of reporting poor health using binary regression

[5] The base year of CPI is 2012-13.

[6] These are the towns for the urban sample and villages for the rural sample.

with a logit link function and robust error variance, given as:

$$log\frac{\pi_i}{(1-\pi_i)} = \beta_0 + \beta X_i + T$$

where $\frac{\pi_i}{(1-\pi_i)}$ is the odds that self reported poor health for individual $i = 1, 0$ otherwise. β_0 represents the log odds of reporting poor health for the reference category. βX_i represents the change in the log odds of reporting poor health for a one unit change in a vector of independent variables (age, sex and education, income quintiles). Here, the odds of an event is the ratio of the probability of the event happening to the probability of not happening (i.e $\frac{p}{1-p}$).

In this model, we have time fixed effects, T, to permit systematic change of the overall ill-health rate by wave.

2. The location may influence ill health. This could derive from differences in state performance on public health, e.g. on problems such as pollution control. To measure the variation of ill-health by location, after controlling for individual characteristics, we allow the intercept to vary by location (k). We estimate:

$$log(\frac{\pi_i}{1-\pi_i}) = \beta_0 + \beta X_i + k + T$$

This HR fixed effects estimate effectively has a distinct intercept k for each homogeneous region. This yields estimates for 102 coefficients for the HRs of India, which can be viewed as properties of these locations. A sorted list of these coefficients represents a sorted list of the places in India where certain features of these locations are associated with the best or worst health, after controlling for household characteristics.

3. Finally, we also allow the slope in income to vary by location. We estimate:

$$log(\frac{\pi_i}{1-\pi_i}) = \beta_0 + \beta X_i + k + k * log.income + T$$

With this in hand, there is an overall model that applies for the whole country, but the slope and intercept for log income are measured on a per-HR basis.

With these estimates in hand, we perform certain counter-factual calculations for one canonical individual. We focus on a Hindu male, who is of the SC/ST caste, who is in the 35-49 age group, and is educated upto Class12/Diploma level. The person is in a rural household which has a real income of Rs.13,000 per month (about USD 266 per month). This is approximately the modal individual in the dataset. Holding these individual and household characteristics constant, we explore the extent to which the predicted ill health probability changes when this person is moved across each HRs. This gives us a way to visualise the impact of location on SRH. It shows how ill health varies within India, after removing non-comparability owing to differences in age and income.

3 Results

The data contains information on 736,945 unique individuals from 170,804 unique households across the six waves from Wave 1 2018 till Wave 3 2019. The overall sample size is 3.5 million observations: on average, each unique individual is measured 4.5 times. Table A.1 in the appendix shows the sample size across each wave used in the dataset. The panel is unbalanced: some households are present in less than three waves.

Table A.2 in the appendix presents summary statistics about the dataset. There are 47.18% females. About 69.5% of the sample is in the 10-49 age group. While 84% of the sample is Hindu, 11% is Muslim. The education measure that we focus on in this paper is the education level of the most educated person in the household: this person may be expected to process information and help make health-related decisions for everyone in the household.

The overall estimate of self-reported ill health rate in the dataset is 3.25%. This overall average SRH rate implies an estimated 44.4 million individuals in India were reported as unhealthy on any one given date. It also implies that, on an average, individuals are unwell for about 12 days a year.

Figure 1 presents the share of unhealthy people in every HR. The map shows many intriguing facts that invite greater exploration. A few regions stand out as having a greater ill-health rate: Uttarakhand, West Haryana, East Uttar Pradesh, Bengal, Assam, Telangana and Andhra Pradesh, and Kerala. Both Himachal Pradesh and Uttarakhand are mountainous regions, but ill health in Uttarakhand is much worse.

One concern about SRH measurement is the extent to which psychological biases are systematically present in certain cultures. When we look at the map, there are many states within which we may expect a certain degree of cultural homogeneity, but the SRH values are heterogeneous. Haryana, Kerala, Uttar Pradesh and West Bengal show such features. This helps increase our confidence in the extent to which the psychological aspects of SRH measurement are not primarily shaped by culture.

Figure 1 Share of unhealthy people across HRs

The figure presents the geographical variation in the share of unhealthy persons across the HRs of India. The white parts of the map are the areas where CMIE does not conduct survey operations.

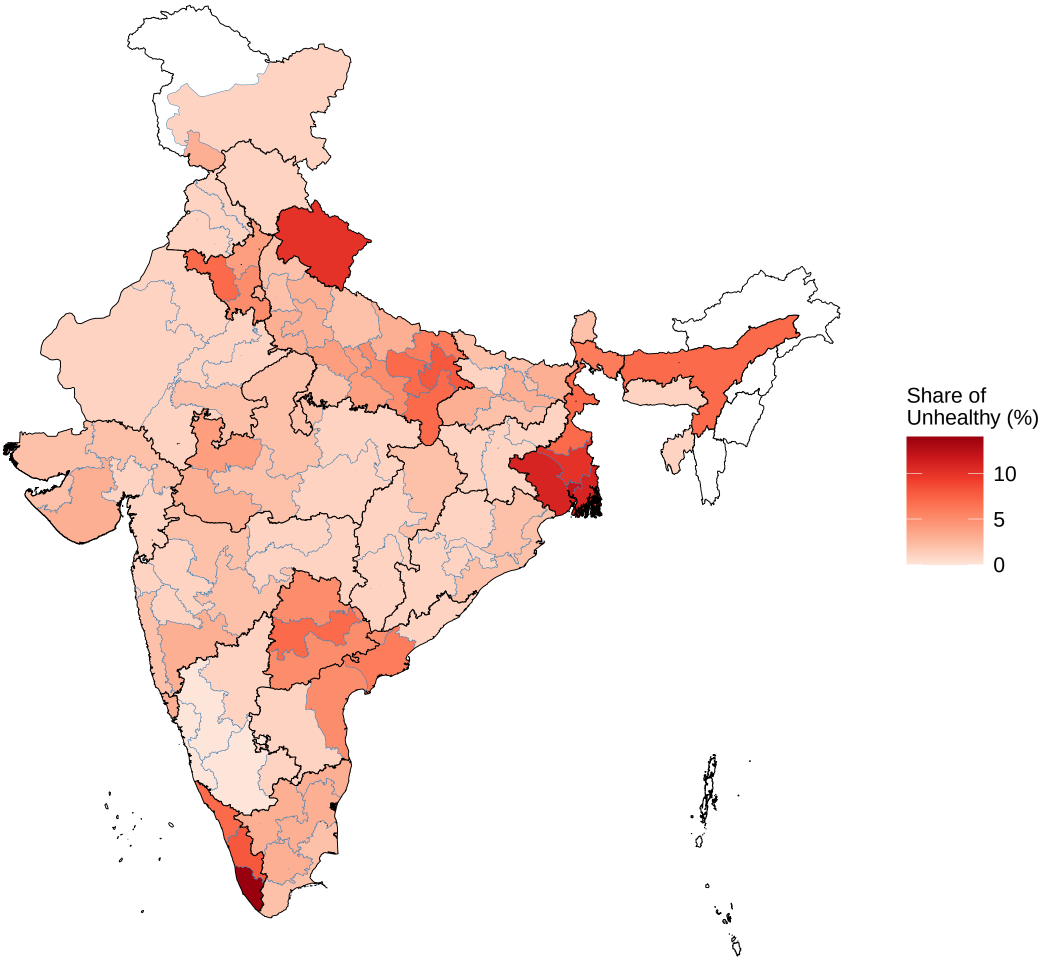


Figure 2 Age specific unhealthy curve

We show the share of persons who self-report ill health, across age groups. The y axis is in log scale.

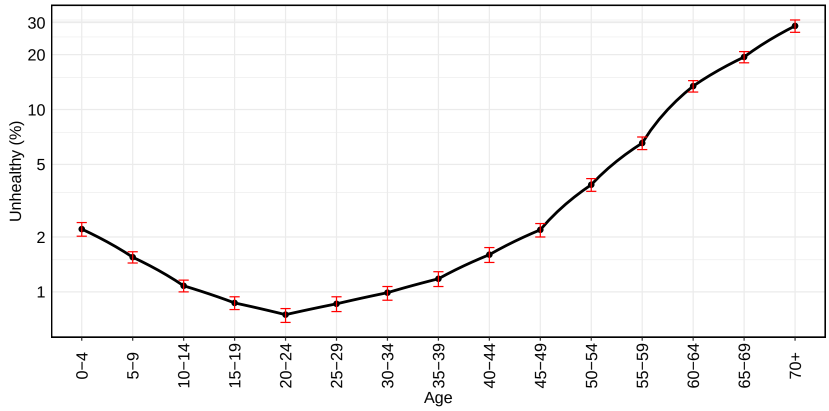


Figure 1 shows the rates of self-reported ill health across the country, regardless of the reason why this might be happening. This map directly illuminates health *care* requirements. From the viewpoint of understanding the health of the people, of course, we must recognise that many factors are at work in generating this heterogeneity. As an example, high levels of ill health in Kerala may reflect the greater share of the elderly in Kerala.

We turn to examining the variation of the ill-health rate by individual and household characteristics. Figure 2 shows the age variation of the ill-health rate. It is useful to recall that an ill health rate of 1% corresponds to 3.65 unhealthy days per year. The graph, where the y axis is in log scale, shows a U shaped pattern. A little over 2% of infants (0-4 age group) are reported as unhealthy. This declines to the lowest ill-health at the 20-24 age group, and then degrades with age. For the entire age range from 10 to 39 years, the ill health rate is relatively low, with a peak value that is slightly above 1%.

As we move to the elderly, there are large jumps in the share of unhealthy people. In the Indian health literature and in health policy discussions, there has been considerable interest in the left edge (e.g. with problems of infant mortality) but not at the right edge (the health of the elderly).

Table 1 shows the fraction of people who reported being unhealthy in each age group, interacted with socio-economic characteristics. The first row – Overall – represents a useful comparison point for the other rows.

Table 1 Fraction of persons who report ill-health

The table presents the ill health rate in different age groups, with the variation across socio-economic characteristics, in India.

	Age					
	0-4	5-9	10-34	35-49	50-59	60+
Overall	2.21	1.55	0.91	1.66	5.1	20.24
Region						
Rural	2.26	1.59	0.98	1.82	5.32	20.67
Urban	2.04	1.44	0.74	1.36	4.70	19.52
Gender						
Male	2.22	1.53	0.86	1.43	4.60	20.59
Female	2.20	1.56	0.96	1.89	5.66	19.82
Religion						
Hindu	2.21	1.55	0.90	1.59	5.07	20.22
Muslim	2.29	1.57	1.06	2.73	6.69	22.72
Others	1.90	1.45	0.54	0.60	2.37	16.98
Caste Category						
Upper Caste	2.15	1.54	0.94	2.22	6.57	24.21
OBC & Intermediate Caste	2.22	1.60	0.88	1.41	4.43	19.00
SC & ST	2.23	1.48	0.95	1.79	5.42	18.92
Not Stated	2.61	1.53	0.58	0.57	1.98	22.02
Maximum household education						
None or Primary	2.73	1.70	1.23	2.08	6.02	21.70
Class 10	2.10	1.50	1.02	1.74	5.26	19.59
Class 12/Diploma	2.20	1.46	0.81	1.67	4.82	19.51
College & above	2.21	1.69	0.82	1.45	4.94	20.90
Income Quintiles						
Lowest	2.21	1.60	1.23	2.49	6.87	21.82
Second	2.12	1.47	0.94	1.61	4.91	19.46
Middle	2.26	1.33	0.81	1.34	4.92	20.88
Fourth	2.23	1.66	0.69	1.24	4.52	20.42
Fifth	2.31	1.78	0.63	1.00	3.85	17.65
Region						
Central	2.46	1.47	0.56	0.73	2.77	13.93
East	2.15	1.61	1.16	3.34	8.89	23.73
North	2.26	1.53	0.93	1.71	5.32	24.06
North-East	2.02	0.83	0.95	4.13	12.95	39.05
South	1.61	1.41	0.59	0.51	3.68	24.74
West	2.69	1.87	1.02	1.04	1.61	4.27
Sample Size	83,135	192,377	1,569,556	852,713	450,194	333,305

In the 35–49 age range, there is greater ill-health with the rural population, for women and for Muslims. The upper caste shows the highest ill-health rate in the 35–49 age range. When the most educated person in the household is in the lowest education category (none or primary), the ill-health rate is much higher, at 2.08%, in the 35–49 age range and 1.23% in the 10–34 age range.

Higher household income is generally associated with lower ill health rates. For all the age ranges from 10 to 59, there is a monotonic relationship: richer households have reduced ill health. But this is not the case below age 10 and above age 59.

A survivorship bias may be at work. If poor people are more prone to die, when prosperity arrives, mortality may improve, so that persons are more likely to be alive and report they are unwell. The ill-health rate for age 0-4 is at 2.21% in the lowest income quintile and actually worsens to 2.31% at the top quintile. Similarly, when it comes to the elderly, the ill health rate falls to 19.46% in the second income quintile, but rises to 20.88% for the middle and then the best value of 17.65% is obtained at the top income quintile.

Finally, there are strong location effects visible in this table. In the age 35–49 range, we get a large variation by region. Two regions are well above the overall average of 1.66% (4.13% (North-East), 3.34% (East)) and three are well below (1.04% (West), 0.73% (Central) and 0.51% (South)).

While the patterns in Table 1 are quite revealing, all these covariates are correlated with each other. It is, therefore, useful to look at the adjusted odds ratios from the logit regression where these covariates are all present at once in the model. These are presented in Table 2. Columns (1) presents the odds-ratio and the 95% confidence intervals and clustered standard errors for a single model that covers all households, with only wave fixed effects. Column (2) presents the same, with HR fixed effects. The single intercept for the whole country is broken out into 102 intercepts for the 102 HRs.

Finally, Column (3) presents the results with HR controls and HR and income interaction effects. Instead of a single coefficient for log income for all households, this coefficient is permitted to vary by HR.

The variation of the odds-ratio in Model (2), by age, is consistent with the variation seen in the simple summary statistics of Figure 2. The coefficient of log income is statistically and economically significant. A striking feature of Model (2) when compared with Model (1) is that once we control for age, income and location, the importance of religion and caste subsides. The most important sources of variation of SRH, in Model (2) and (3) are age,

Table 2 Odds ratios from logistic regressions that predict self reported (ill)Health

The table presents the results from a logit regression of individual characteristics on self reported health. We report the odds-ratio and the 95% confidence intervals. Standard errors have been clustered at the PSU level.

	(1)	(2)	(3)
Residence *(Ref:Rural)*			
Urban	1.064	1.039	1.025
	(0.936, 1.210)	(0.975, 1.108)	(0.961, 1.093)
Age *(Ref: age 10-34)*			
0-4	2.771***	2.595***	2.587***
	(2.525, 3.042)	(2.346, 2.871)	(2.337, 2.863)
5-9	1.814***	1.814***	1.826***
	(1.704, 1.932)	(1.695, 1.943)	(1.706, 1.955)
35-49	1.754***	1.780***	1.788***
	(1.612, 1.909)	(1.621, 1.955)	(1.629, 1.963)
50-59	5.859***	6.065***	6.043***
	(5.303, 6.472)	(5.457, 6.740)	(5.438, 6.715)
60+	28.071***	31.974***	31.639***
	(25.269, 31.185)	(28.710, 35.610)	(28.423, 35.219)
Gender *(Ref: Male)*			
Female	1.053***	1.051***	1.048***
	(1.022, 1.086)	(1.019, 1.084)	(1.016, 1.081)
Religion *(Ref: Hindu)*			
Muslim	1.062	1.058	1.051
	(0.953, 1.183)	(0.989, 1.131)	(0.984, 1.122)
Other	0.725***	1.038	1.029
	(0.595, 0.884)	(0.952, 1.131)	(0.943, 1.123)
Caste *(Ref: Upper Caste)*			
Not Stated	0.747	1.031	1.049
	(0.531, 1.050)	(0.915, 1.162)	(0.927, 1.186)
OBC/Intermediate	0.705***	0.989	0.995
	(0.641, 0.776)	(0.944, 1.036)	(0.951, 1.040)
SC/ST	0.730***	0.974	0.983
	(0.666, 0.800)	(0.927, 1.023)	(0.936, 1.032)
Education *(Ref: None or Primary)*			
Class 10	0.976	1.016	1.070***
	(0.906, 1.050)	(0.966, 1.070)	(1.017, 1.125)
Class 12/ Diploma	0.964	0.981	1.037
	(0.879, 1.056)	(0.924, 1.041)	(0.977, 1.101)
College and above	1.070	1.014	1.063
	(0.955, 1.200)	(0.945, 1.089)	(0.991, 1.140)
Log.income	0.788***	0.839***	1.162
	(0.730, 0.851)	(0.798, 0.883)	(1.006, 1.342)
Intercept	0.094***	0.008***	0.0004***
	(0.047, 0.186)	(0.005, 0.014)	(0.0001, 0.002)
HR controls	No	Yes	Yes
HR*log income controls	No	No	Yes
Wave (Time) controls	Yes	Yes	Yes
Observations	3,481,280	3,481,280	3,481,280
Log Likelihood	−402,203.200	−364,715.000	−363,194.600
BIC	804737.8	731282.7	729763.2
Note:			*p<0.1; **p<0.05; ***p<0.01

income and location. In Model (3), there are small enhanced ill health rates for women (OR=1.048) and one education categories (OR=1.070 for Class 10).

When we compare Model (1) versus Model (2) there is a large gain in the Bayesian Information Criterion (from 804,737 to 731,282). This suggests that the HR fixed effect (variation of the intercept of the model by HR) is particularly important. In comparison, there is a reduced gain in going from Model (2) to Model (3) (from 731,282 to 729,763).

Figure 1 had shown us the unconditional geographical variation of ill health. With the models of Table 2 in hand, we can control for some explanatory variables and focus on geographical variation. In order to do this, we make predictions for the ill-health rate for the canonical individual: a Hindu, male, in the 35-49 age group, of the SC/ST caste category, with a Class 12/Diploma education, in a rural household with a real income of Rs.13,000 per month. This is approximately the modal person in the dataset. Within each homogeneous region, we compute the predicted probability of ill health, using Model (3). These predicted probabilities are mapped in Figure 3.

The comparison between the unconditional geographical variation of ill health (Figure 1) and the geographical variation of ill-health after controlling for individual characteristics (Figure 3) is revealing. The range of values, in the unconditional map, is much higher, with values ranging from 0 to 15%. Once we narrow down to the canonical person, important sources of variation (age and income) are removed from the picture. Persons in the 35–49 age range are healthier than the overall population. Hence, the range of values seen in Figure 3 is smaller: from 0 to 4%.

When a hot spot like south Kerala is visible in Figure 1, this could be associated with age structure, income or location. When it shows up (more weakly) in Figure 3, this suggests there is some feature of the *location* which is associated with greater ill health, which is not merely induced by age structure and income.

A striking feature of the two figures is the extent to which they look similar (though of course the scale is quite different). This is consistent with the statistical findings in Table 2, that location has a powerful impact upon ill health.

The overall result in Model 2 and Model 3 is that health improves with log income. Model (3) permits the slope in income to also vary by location, thus yielding 102 coefficients, for the slope on log income in each HR. Figure 4 helps us visualise the estimated slopes. Regions that are coloured blue are

Figure 3 Model-based predictions of the probability of ill health for a fixed individual

This figure shows the predicted probability, from Model (2) of Table 2, for an approximately modal person: a male, Hindu, age group 35-49, SC/ST, Class 12/Diploma, real monthly income of Rs.13,000, residing in a rural region.

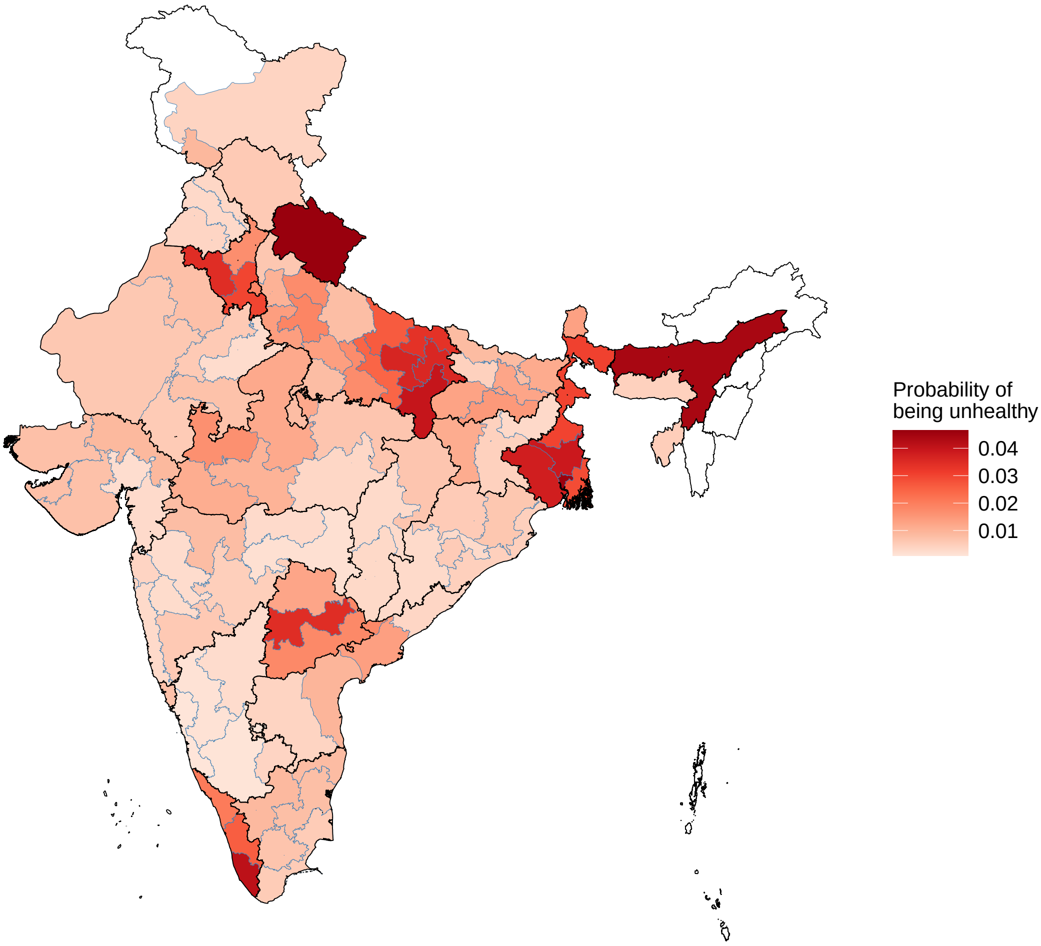
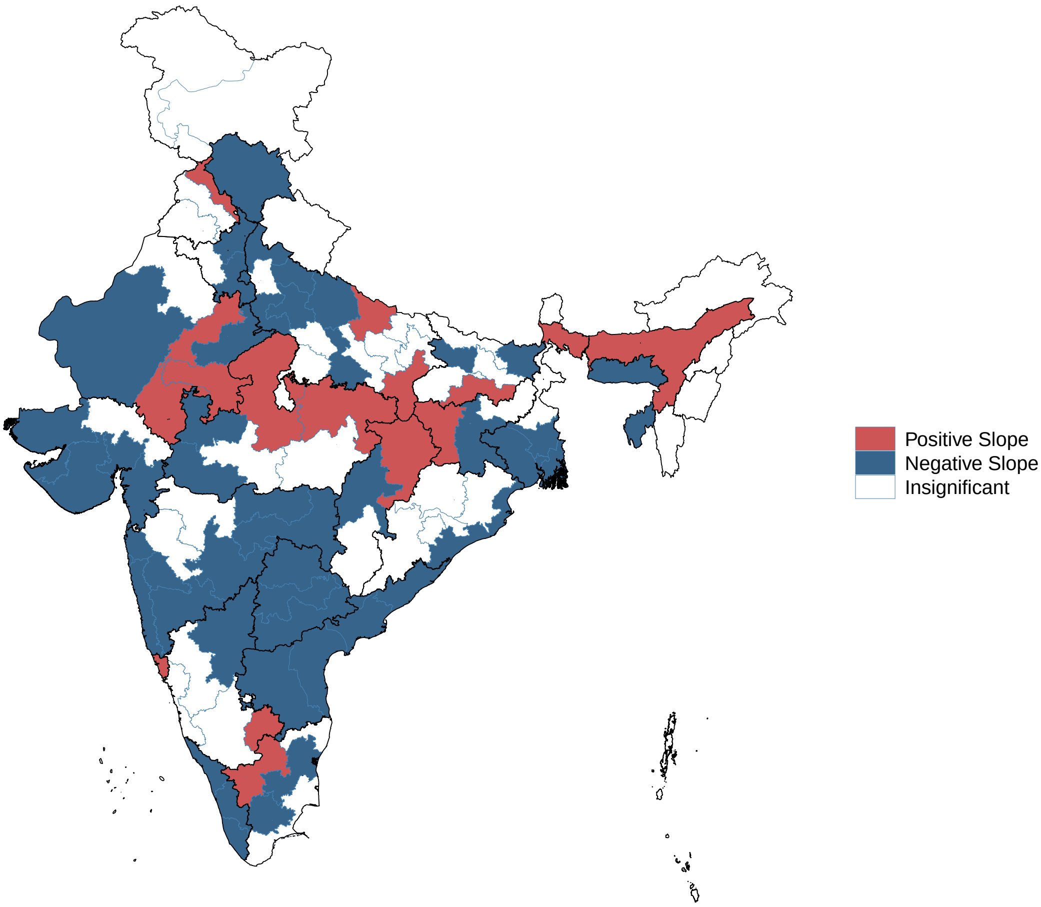


Figure 4 Variation in income and HR slope coefficients

The figure shows the variation in the coefficients on the slope of the location and income interaction effect.



those where higher income is associated with reduced ill health, where the null hypothesis of a slope of 0 is rejected at a 95% level of significance. This property is found in only 49% of the HRs of India. With 35% of the HRs, H_0 is not rejected. In 16% of the locations, shown in red, ill health is *greater* for households with higher income.

4 Discussion

Table 3 shows the names of the 10 HRs of India with the highest and lowest ill-health rates. There is a need for further research in understanding the sources of this variation, particularly the variation seen in the predicted ill-health rates for the modal person which reflect characteristics of these *locations* that can potentially be influenced by modified strategies in public health.

We may offer some conjectures about the epidemiological phenomena at work in Figure 3. The region of eastern Uttar Pradesh, Assam and Bengal is known to face high arsenic contamination of water (Chakraborti et. al., 2018; Sengupta et. al., 2003). The adverse health outcomes of arsenic contamination in the form of skin lesions, neurological effects, obstetric problems, cardiovascular effects and cancers typically involving the skin, lung, and bladder (Ahamed et. al., 2006; Chakrabarty & Sarma, 2011; Mahanta et al., 2016). Our reported measures of poor health in Eastern India may reflect this phenomenon.

Similarly, the high ill-health in Uttarakhand, as opposed to the neighbouring hilly state of Himachal Pradesh, might be associated with religious pilgrimage and festivals (David & Roy, 2016; Sridhar et al., 2015) where pilgrims from all across India bring novel pathogens to Uttarakhand, and unhygenic crowding of pilgrims within Uttarakhand create conditions where more virulent strains succeed.

The conventional wisdom of health economics expects that higher income generates better health, through a combination of improved nutrition, housing (which impacts on hygiene and disease vectors), knowledge and health care. In the aggregate, in our results, an increase in log income is indeed associated with reduced ill health. While this paper makes no causal claims, ultimately, if there is a causal and positive link between income and health, then mere economic growth would help improve health (while recognising that some of this impact flows through higher public and private expenses on health).

The surprising result is that there is considerable geographical heterogeneity

Table 3 The HRs with the highest and lowest ill-health rates

The first column shows the 10 HRs of India with the lowest and highest observed ill-health rates (i.e. from the picture seen in Figure 1). This reflects a combination of income, age structure and location characteristics.
The second column shows the 10 HRs of India with the lowest and highest predicted ill-health rates for the modal person (i.e. predicted using Model (3) of Table 2, which is the same as the geographical variation depicted in Figure 3). This reflects location characteristics.
As an example, Uttarakhand shows up as the 98th most unhealthy HR based on the raw ill-health rate. After controlling for income and age structure, it proves to have the highest ill-health.

Rank	By raw ill-health rate	Predicted modal person
1.	Chitradurga - Mysore	Uttara Kannada - Dakshin Kannada
2.	Uttara Kannada - Dakshin Kannada	Chitradurga - Mysore
3.	Belgaum - Shimoga	Belgaum - Shimoga
4.	Bharatpur - Tonk	Hingoli - Gadchiroli
5.	Hingoli - Gadchiroli	Ahmadabad - Kheda
6.	Giridih - Dumka	Nashik - Ahmadnagar
7.	Bidar - Bellary	Bidar - Bellary
8.	Ahmadabad - Kheda	Bharatpur - Tonk
9.	Pune	Pune
10.	Rajasamand - Banswara	Amravati
93.	Assam	Azamgarh - Gorakhpur
94.	Sirsa - Bhiwani	Faizabad - Jaunpur
95.	Azamgarh - Gorakhpur	Sirsa - Bhiwani
96.	Palakkad - Idukki Azamgarh - Gorakhpur	Puruliya - Medinipur
97.	Uttarakhand	Mau - Sonbhadra
98.	Barddhaman - Nadia	Barddhaman - Nadia
99.	Puruliya - Medinipur	Kottayam - Thiruvananthapuram
100.	24 Parganas	Assam
101.	Kolkata - Haora	Kolkata - Haora
102.	Kottayam - Thiruvananthapuram	Uttarakhand

in this slope. In about 40% of India, the relationship is negative (the rich have less ill health). But in the rest of India, this positive association is not observed. There is a significant area where health is *worse* for higher income households.

The impact of increased income upon health through improved nutrition and housing is likely to operate all across the country in relatively uniform ways. If health care is completely absent, there should be a negative slope in income as higher income is likely to generate better nutrition, housing and knowledge. If government-supplied health care works well, all income groups would get access to comparable health care, and a negative slope in income would still be generated through the impact of nutrition, housing and knowledge. If private health care works well, higher income would generate better purchases of health care services, and a negative slope in income would arise.

It is thus a puzzle, to explain the white areas (rich and poor are similarly healthy) and red areas (the rich have higher ill-health). We conjecture that this is related to health care and survivorship bias. High income households in certain regions may get low quality care, to a point where it overshadows the other channels of impact (improved nutrition, housing, knowledge). If there is a large income gap in mortality, for the young and the old, then high income may generate more survivors who are in a state of ill health, thus generating a zero or negative slope for health in income.

The two maps (Figures 3 and 4) do not fit within a simple North India vs. South India stereotype. While the population centres of Uttar Pradesh and Bihar are considered to be poverty traps, the canonical person is only particularly unhealthy in Eastern Uttar Pradesh. There are some parts of UP and Bihar where the rich are healthier than the poor. The HRs which have greater difficulties are present in many parts of the country and call for a revision of our priors about where ill health in India is found.

The phenomena seen in the two maps (Figures 3 and 4) are not the same. The regions where ill health worsens for high income people are not the same as the regions with high ill health. These are distinct phenomena that require distinct explanations.

There are two limitations of this work. The standard difficulties of SRH measurement are present here: where health is observed through an intermediating psychological filter which introduces a certain degree of imprecision. To the extent that these psychological characteristics are correlated with the variables of interest, the statistical estimates are biased.

The second limitation of this study is that the logistic regressions shown here lack a causal interpretation. These calculations should be viewed as merely describing features in the data. They do not guide interventions where certain features are changed in order to impact upon SRH.

5 Conclusions

In this paper, we have undertaken a novel examination of a health outcome measure in India in a large-scale household survey database pertaining to calendar years 2018 and 2019, drawing on the health status seen in 3.2 million records for individuals. In the existing health literature, there is a substantial analysis of the health of infants, children and mothers. In this paper, we analyse the entire population.

The overall aggregate ill-health rate is about 3.25%, which maps to about 44 million persons being unwell on any one day. The average individual is unwell for about 12 days a year. There is a U-shaped curve, with a long zone of reduced ill health rates from age 10 to age 39.

The important correlates of SRH are age, income and location. Once these are controlled for, other individual characteristics that were explored here (education, caste, religion, gender) are relatively unimportant.

Location is remarkably important. Health researchers, and health policy makers, need to look beyond all-India averages, and recognise the immense variation within the country. The three maps constructed in the paper – the raw ill-health rate, the ill-health rate of the modal person and the regions where higher income is correlated with improved health – are novel and go beyond the simple preconceptions of North India vs. South India. The geographical variation shown here merits further exploration.

6 Declarations

6.1 Ethics approval and consent to participate

This research does not constitute regulated human subjects research requiring review. CMIE has obtained necessary participant consent.

6.2 Consent for publication

Not applicable.

6.3 Availability of data and materials

The CPHS database, which is the foundation of this paper, is available to all researchers from CMIE upon payment of a subscription fee. This data was used under license for the current study and is hence not publicly available.

6.4 Competing interests

The authors declare that they have no competing interest.

6.5 Funding

We acknowledge support from the Thakur Foundation.

6.6 Authors' contributions

IP and AS initiated the project. The research strategy was created by all the four authors. The implementation was done by IP, AS, and RS. All authors read and approved the final manuscript.

6.7 Acknowledgements

We thank Subhamoy Chakraborty, Moumita Das and Mithila Sarah for research assistance. We are grateful to Jeff Hammer, Indu Bhushan, and Amrita Agarwal for useful discussions. We acknowledge funding support from the Thakur Foundation. All errors are our own.

References

Ahamed et. al. (2006). Arsenic groundwater contamination and its health effects in the state of Uttar Pradesh (UP) in upper and middle Ganga plain, India: a severe danger. *Science of the Total Environment*, *370*(2-3), 310–322. https://doi.org/10.1016/j.scitotenv.2006.06.015

Banerjee, A., Deaton, A. & Duflo, E. (2004). Health, health care, and economic development. *American Economic Review*, *94*(2), 326–330. https://doi.org/10.1257/0002828041301902

Benyamini, Y. & Idler, E. L. (1999). Community studies reporting association between self-rated health and mortality: Additional studies, 1995 to 1998. *Research on aging*, *21*(3), 392–401.

Boerma, T., Hosseinpoor, A. R., Verdes, E. & Chatterji, S. (2016). A global assessment of the gender gap in self-reported health with survey data from 59 countries. *BMC public health*, *16*(1), 675.

Chakrabarty, S. & Sarma, H. P. (2011). Heavy metal contamination of drinking water in Kamrup district, Assam, India. *Environmental monitoring and assessment*, *179*(1-4), 479–486. https://doi.org/0.1007/s10661-010-1750-7

Chakraborti et. al. (2018). Groundwater arsenic contamination in the Ganga river basin: A future health danger. *International Journal of Environmental Research and Public Health*, *15*(180). https://doi.org/doi:10.3390/ijerph15020180

Cullati, S., Mukhopadhyay, S., Sieber, S., Chakraborty, A. & Burton-Jeangros, C. (2018). Is the single self-rated health item reliable in India? A construct validity study. *BMJ global health*, *3*(6), e000856.

Dasgupta, A. (2018). Systematic measurement error in self-reported health: Is anchoring vignettes the way out? *IZA Journal of Development and Migration*, *8*(1), 12.

David, S. & Roy, N. (2016). Public health perspectives from the biggest human mass gathering on earth: Kumbh Mela, India [Mass Gathering Medicine]. *International Journal of Infectious Diseases*, *47*, 42–45. https://doi.org/https://doi.org/10.1016/j.ijid.2016.01.010

Heistaro, S., Jousilahti, P., Lahelma, E., Vartiainen, E. & Puska, P. (2001). Self rated health and mortality: A long term prospective study in eastern finland. *Journal of Epidemiology & Community Health*, *55*(4), 227–232.

Hirve, S., Juvekar, S., Lele, P. & Agarwal, D. (2010). Social gradients in self-reported health and well-being among adults aged 50 years and over in Pune district, India. *Global Health Action*, *3*(1), 2128. https://doi.org/10.3402/gha.v3i0.2128

Idler, E. L. & Benyamini, Y. (1997). Self-rated health and mortality: A review of twenty-seven community studies. *Journal of Health and Social Behavior*, *38*(1), 21–37. http://www.jstor.org/stable/2955359

Mahanta, R., Chowdhury, J. & K.Nath, H. (2016). Health costs of arsenic contamination of drinking water in Assam, India. *Economic Analysis and Policy*, *49*, 30–42. https://doi.org/10.1016/j.eap.2015.11.013

Miilunpalo, S., Vuori, I., Oja, P., Pasanen, M. & Urponen, H. (1997). Self-rated health status as a health measure: The predictive value of self-reported health status on the use of physician services and on mortality in the working-age population. *Journal of clinical epidemiology*, *50*(5), 517–528.

Onur, I. & Velamuri, M. (2018). The gap between self-reported and objective measures of disease status in India. *PloS one*, *13*(8), e0202786.

Sengupta et. al. (2003). Groundwater arsenic contamination in the Ganga-Padma-Meghna-Brahmaputra plain of India and Bangladesh. *Archives of Environmental Health: An International Journal*, *58*(11), 701–702. https://doi.org/10.3200/AEOH.58.11.701-702

Singh, B. (2018). Anthropological investigations of vitality. *Journal of Ethnographic Theory*, *8*(3), 550–565.

Singh, L., Arokiasamy, P., Singh, P. K. & Rai, R. K. (2013). Determinants of gender differences in self-rated health among older population: Evidence from India. *SAGE Open*, *3*(2), 2158244013487914. https://doi.org/10.1177/2158244013487914

Sridhar, S., Gautret, P. & Brouqui, P. (2015). A comprehensive review of the Kumbh Mela: Identifying risks for spread of infectious diseases. *Clinical Microbiology and Infection*, *21*(2), 128–133. https://doi.org/https://doi.org/10.1016/j.cmi.2014.11.021

Subramanian, S., Subramanyam, M. A., Selvaraj, S. & Kawachi, I. (2009). Are self-reports of health and morbidities in developing countries misleading? Evidence from India. *Social science & medicine*, *68*(2), 260–265.

Vellakkal, S., Subramanian, S. V., Millett, C., Basu, S., Stuckler, D. & Ebrahim, S. (2013). Socioeconomic inequalities in Non-Communicable Diseases prevalence in India: Disparities between self-reported diagnoses and standardized measure. *PLoS ONE*, *8*((7): e68219).

Wu, S., Wang, R., Zhao, Y., Ma, X., Wu, M., Yan, X. & He, J. (2013). The relationship between self-rated health and objective health status: A population-based study. *BMC public health*, *13*(1), 320.

Table A.1 Sample size across the three waves of 2018 and 2019

This table presents summary statistics about the CPHS data for 2018 and 2019.

Months	Wave	Total Households	Total Members
Jan - Apr, 2018	W1 2018	143151	575082
May - Aug, 2018	W2 2018	149101	594655
Sep - Dec, 2018	W3 2018	147123	584109
Jan - Apr, 2019	W1 2019	146292	582180
May - Aug, 2019	W2 2019	147840	575044
Sep - Dec, 2019	W3 2019	147291	570210

Appendix

Tables and Figures

Table A.2 Data Summary: 2018 and 2019

This table presents summary statistics about the CPHS data for 2018 and 2019.

Variable		Sample Size
Unique Households		170,804
Unique Individuals		736,945
Total Sample size		3,481,280
Self Reported Health (%)		
Healthy	96.77	3,368,947
Unhealthy	3.23	112,333
Gender (%)		
Male	52.82	1,838,657
Female	47.18	1,642,623
Residence (%)		
Rural	36.86	1,283,068
Urban	63.14	2,198,212
Age Group (%)		
0-4	2.39	83,135
5-9	5.53	192,377
10-34	45.09	1,569,556
35-49	24.49	852,713
50-59	12.93	450,194
60+	9.57	333,305
Religion (%)		
Hindu	83.83	2,918,231
Muslims	10.55	367,321
Others	5.62	195,728
Caste Category (%)		
Upper Caste	23.51	818,278
OBC/Intermediate	48.36	1,683,578
SC/ST	26.66	928,222
Not Stated	1.47	51,202
Max HH Education (%)		
None or Primary	4.54	158,205
Class 10	32.21	1,121,262
Class 12/ Diploma	31.36	1,091,790
College and above	31.89	1,110,023
Income quintile (average household monthly income)		
Lowest	6273	695,718
Second	9465	696,396
Middle	12889	696,395
Fourth	18674	696,397
Highest	38968	696,374
Region (%)		
Central	8.02	279,171
East	18.53	645,021
North	34.56	1,203,178
North-East	2.39	83,115
South	19	661,390
West	17.51	609,405