Now That's Fresh Water*

Trends in India from the 1992 Demographic and Health Survey

Bilal Haq and Ritvik Puri

02 April 2022

Abstract

Third world countries often struggle with figuring out where the needs of their people lie. Analysis of large datasets can help shed some light on the issue. We obtain a dataset from the National Family Health Survey in India, hosted by the Demographic and Health Surveys program in the U.S. In a reproducible way, we convert this dataset from the 1990s into a usable digital format, and analyze it. In doing so, we conclude that acquiring a large dataset and using it to establish policies should be a primary goal of developing countries that want to improve their economic conditions.

Contents

*Code and data are available at: https://github.com/haqbilal/India-1992-DHS-State-Findings

1 Introduction

The period 1980-2000 featured data science at the forefront of a developing era of statistical modelling, and several third world countries were beginning to jump on the bandwagon. Acknowledging that a large scale survey on the population of India had never been done before, K.B. Pathak (IIPS/India (1995)) was ready to forge the path forward on behalf of his country. He was the director of the International Institute for Population Sciences (IIPS) in Bombay, and his team headed the Project to Strengthen the Research Capabilities of the Population Research Centres in India in 1991. The primary objective of this campaign was to finally provide state-level and national-level estimates of fertility, infant/child mortality, family planning practices, maternal/child health care, and the utilization of services provided for mothers and children. The National Family Health Survey (NFHS) took on this mission, and as a byproduct, provided high quality data to statisticians for analytical research on a range of population and health studies.

The release of this report came with a call from the IIPS for further analyses of the NFHS data, among researches both in and outside of India. The final report was made publicly available to anyone in the word through the United States Agency for International Development's (USAID) Demographic and Health Surveys (DHS) program. The report contains a thorough record of household/individual respondent characteristics. Namely, data on marriage, fertility, family planning, utilization of antenatal services, vaccination, child feeding practices, nutritional status of children, and knowledge of AIDS. Alongside, interstate variations on key indicators were also included, allowing for regional comparisons to be made. As independent researchers and data scientists, we have taken on the IIPS's call to further analyze this data, and report our findings and comments in this paper.

The report onwards features three sections. Firstly, we look at the dataset provided on pages 31-32 of the NFHS 1992 Final Report. Plots and directed acyclic graphs are used to visualize the data and relationship between variables. Moving on, we share our findings and the implications of results and the effect they might have on public policy in India. We also align our own findings with those of the NFHS, and analyze the progress made over the last 20 years, demonstrating the effect that data analysis has had on the development of the country. Finally, we discuss the methodology used by the NFHS and describe how they were effective in achieving their goal. We also propose some critiques that can aid in the future of data analysis in India and other third world countries following in their footsteps. Our report is carried out using R (R Core Team 2020), alongside various libraries cited (see References). R markdown (Allaire et al. 2022) was used for compilation and presentation.

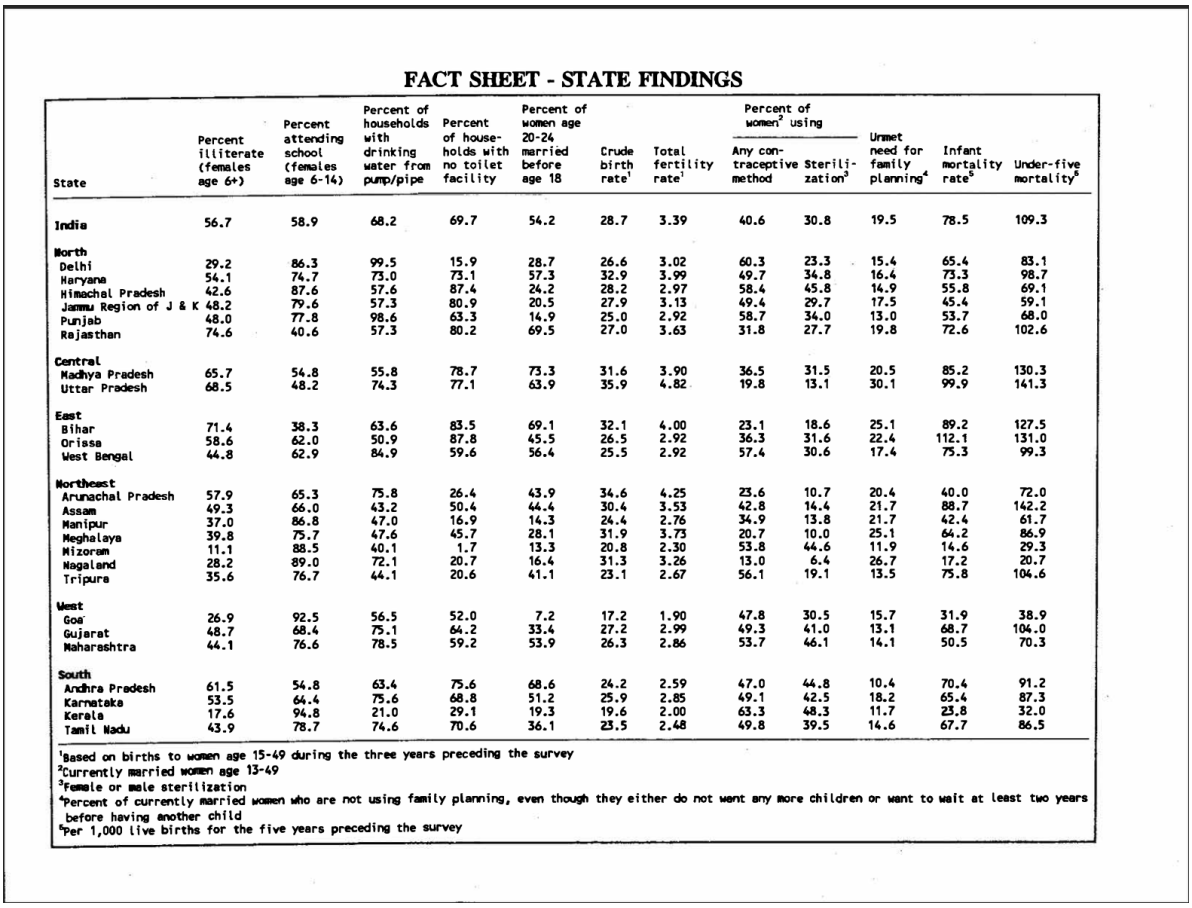
2 Data

2.1 Dataset and Variables

The NFHS's final report from 1992 is publicly hosted on the DHS program's website in PDF format. The data has been sourced in a reproducible way, with a script for downloading and obtaining the relevant datasets. It is an in-depth review of statistics such as access to clean water and amenities, fertility, family planning, child care, and health, on a sample of women and children, with the goal being to summarize the status of women and children, as key factors towards growth of the country, while identifying any present flaws, in hopes of administering aid where possible. In terms of privacy, no underlying information has been kept that could point towards or reveal any individual that took part in the survey. Note that a constraint of this dataset is that it cannot be truly compared to similar reports from other organizations, because of differing methods in categorization as well as geographic stratification. Time constraints may also affect results from different surveys.

Our focus is on the state findings table on pages 31-32 of the report. The full table is shown below in Figures 1 and 2. In particular, we will look at the columns: percent of illiterate females, households with drinking water, mothers receiving antenatal care (that is, special care while pregnant), births delivered in a health facility, and fully immunized children. These variables were selected because they paint the picture of health conditions for those living in India, and the status of pertinent features such as access to water, and

amenities for pregnant women and their children. This information is given for each of the 25 states, as well as in total for all of India. Information from some columns was unavailable or omitted because it involved 25 or less people, in which case they were left out of analysis. A table for the first 5 states of the dataset with filtered variables is shown below in Table 1. All quantities are percentages of the sample for that state, unless otherwise specified.



FACT SHEET - STATE FINDINGS

State	Percent illiterate (females age 6+)	Percent attending school (females age 6-14)	Percent of households with drinking water from pump/pipe	Percent of house-holds with no toilet facility	Percent of women age 20-24 married before age 18	Crude birth rate[1]	Total fertility rate[1]	Percent of women[2] using: Any con-traceptive method	Percent of women[2] using: Sterili-zation[3]	Unmet need for family planning[4]	Infant mortality rate[5]	Under-five mortality[5]
India	56.7	58.9	68.2	69.7	54.2	28.7	3.39	40.6	30.8	19.5	78.5	109.3
North												
Delhi	29.2	86.3	99.5	15.9	28.7	26.6	3.02	60.3	23.3	15.4	65.4	83.1
Haryana	54.1	74.7	73.0	73.1	57.3	32.9	3.99	49.7	34.8	16.4	73.3	98.7
Himachal Pradesh	42.6	87.6	57.6	87.4	24.2	28.2	2.97	58.4	45.8	14.9	55.8	69.1
Jammu Region of J & K	48.2	79.6	57.3	80.9	20.5	27.9	3.13	49.4	29.7	17.5	45.4	59.1
Punjab	48.0	77.8	98.6	63.3	14.9	25.0	2.92	58.7	34.0	13.0	53.7	68.0
Rajasthan	74.6	40.6	57.3	80.2	69.5	27.0	3.63	31.8	27.7	19.8	72.6	102.6
Central												
Madhya Pradesh	65.7	54.8	55.8	78.7	73.3	31.6	3.90	36.5	31.5	20.5	85.2	130.3
Uttar Pradesh	68.5	48.2	74.3	77.1	63.9	35.9	4.82	19.8	13.1	30.1	99.9	141.3
East												
Bihar	71.4	38.3	63.6	83.5	69.1	32.1	4.00	23.1	18.6	25.1	89.2	127.5
Orissa	58.6	62.0	50.9	87.8	45.5	26.5	2.92	36.3	31.6	22.4	112.1	131.0
West Bengal	44.8	62.9	84.9	59.6	56.4	25.5	2.92	57.4	30.6	17.4	75.3	99.3
Northeast												
Arunachal Pradesh	57.9	65.3	75.8	26.4	43.9	34.6	4.25	23.6	10.7	20.4	40.0	72.0
Assam	49.3	66.0	43.2	50.4	44.4	30.4	3.53	42.8	14.4	21.7	88.7	142.2
Manipur	37.0	86.8	47.0	16.9	14.3	24.4	2.76	34.9	13.8	21.7	42.4	61.7
Meghalaya	39.8	75.7	47.6	45.7	28.1	31.9	3.73	20.7	10.0	25.1	64.2	86.9
Mizoram	11.1	88.5	40.1	1.7	13.3	20.8	2.30	53.8	44.6	11.9	14.6	29.3
Nagaland	28.2	89.0	72.1	20.7	16.4	31.3	3.26	13.0	6.4	26.7	17.2	20.7
Tripura	35.6	76.7	44.1	20.6	41.1	23.1	2.67	56.1	19.1	13.5	75.8	104.6
West												
Goa	26.9	92.5	56.5	52.0	7.2	17.2	1.90	47.8	30.5	15.7	31.9	38.9
Gujarat	48.7	68.4	75.1	64.2	33.4	27.2	2.99	49.3	41.0	13.1	68.7	104.0
Maharashtra	44.1	76.6	78.5	59.2	53.9	26.3	2.86	53.7	46.1	14.1	50.5	70.3
South												
Andhra Pradesh	61.5	54.8	63.4	75.6	68.6	24.2	2.59	47.0	44.8	10.4	70.4	91.2
Karnataka	53.5	64.4	75.6	68.8	51.2	25.9	2.85	49.1	42.5	18.2	65.4	87.3
Kerala	17.6	94.8	21.0	29.1	19.3	19.6	2.00	63.3	48.3	11.7	23.8	32.0
Tamil Nadu	43.9	78.7	74.6	70.6	36.1	23.5	2.48	49.8	39.5	14.6	67.7	86.5

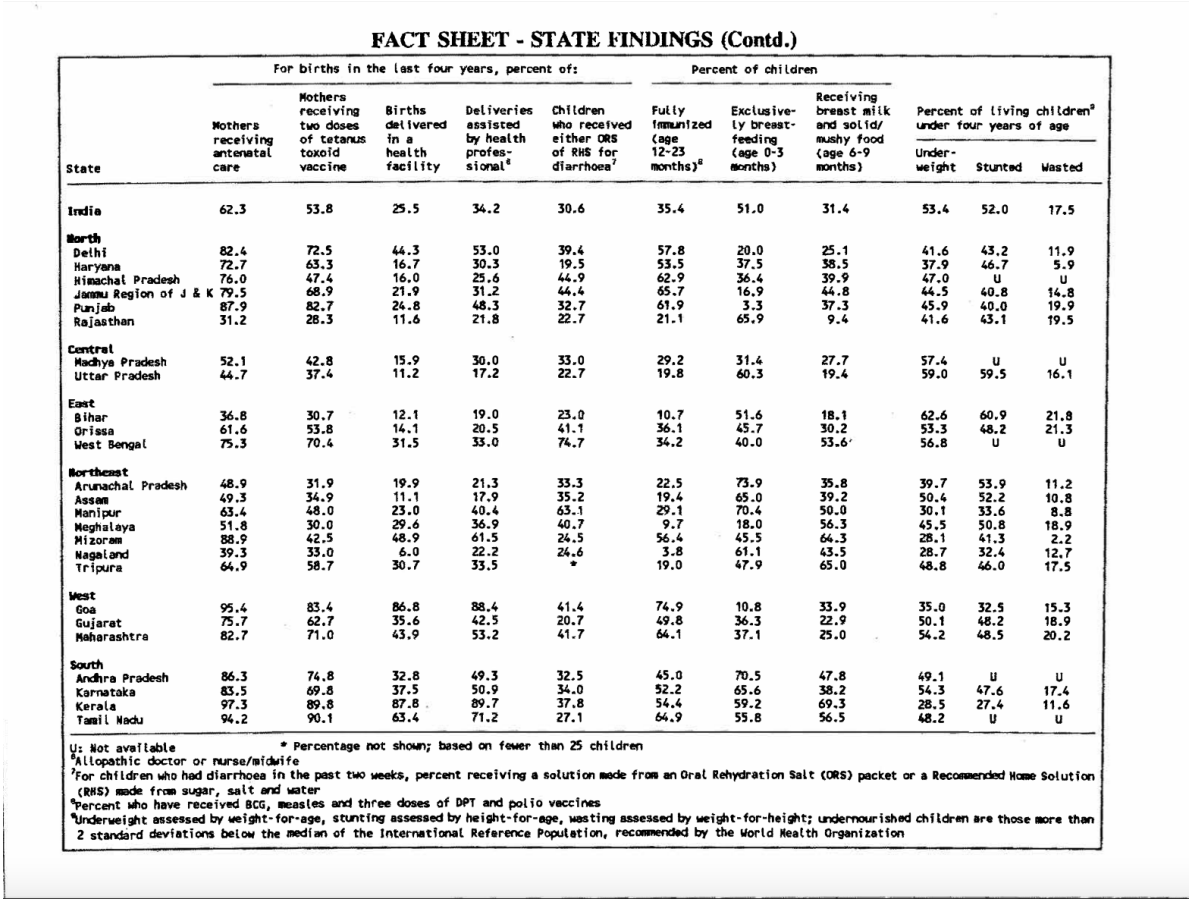
[1]Based on births to women age 15-49 during the three years preceding the survey
[2]Currently married women age 13-49
[3]Female or male sterilization
[4]Percent of currently married women who are not using family planning, even though they either do not want any more children or want to wait at least two years before having another child
[5]Per 1,000 live births for the five years preceding the survey

Figure 1: The Table in the NFHS Final Report

2.2 Methodology

In formulating the survey questionnaires (see Appendix F of the NFHS final report), a Questionnaire Design Workshop was held in Pune, September 1991. There, representatives from IIPS and other organizations such as USAID, designed the contents, which were designed with India's low contraceptive prevalence in mind. Modifications were made to the questionnaire that kept in consideration the Indian social and cultural environment as well as the NFHS objectives. Individual states also included state-specific questions pertaining to issues of importance in those regions. Three questionnaires were created; the Household Questionnaire, Woman's Questionnaire, and Village Questionnaire (IIPS/India (1995)).

The Household Questionnaire listed counts of residents in each sample household, and visitors who stayed from the night before. Information collected included age, sex, marital status, education, occupation, and household conditions such as access to water and toilets. It's main purpose was to identify women eligible to participate in the Woman's Questionnaire (women in between 13-49 years of age). The Woman's Questionnaire then gathered information on the respondents' background, reproduction (i.e. births given), contraception, health of children, fertility preferences (i.e. desire for children), and nutritional status of chil-



FACT SHEET - STATE FINDINGS (Contd.)

State	For births in the last four years, percent of: Mothers receiving antenatal care	Mothers receiving two doses of tetanus toxoid vaccine	Births delivered in a health facility	Deliveries assisted by health professional[6]	Children who received either ORS of RHS for diarrhoea[7]	Percent of children Fully immunized (age 12-23 months)[8]	Exclusively breast-feeding (age 0-3 months)	Receiving breast milk and solid/mushy food (age 6-9 months)	Percent of living children[9] under four years of age Underweight	Stunted	Wasted
India	62.3	53.8	25.5	34.2	30.6	35.4	51.0	31.4	53.4	52.0	17.5
North											
Delhi	82.4	72.5	44.3	53.0	39.4	57.8	20.0	25.1	41.6	43.2	11.9
Haryana	72.7	63.3	16.7	30.3	19.5	53.5	37.5	38.5	37.9	46.7	5.9
Himachal Pradesh	76.0	47.4	16.0	25.6	44.9	62.9	36.4	39.9	47.0	U	U
Jammu Region of J & K	79.5	68.9	21.9	31.2	44.4	65.7	16.9	44.8	44.5	40.8	14.8
Punjab	87.9	82.7	24.8	48.3	32.7	61.9	3.3	37.3	45.9	40.0	19.9
Rajasthan	31.2	28.3	11.6	21.8	22.7	21.1	65.9	9.4	41.6	43.1	19.5
Central											
Madhya Pradesh	52.1	42.8	15.9	30.0	33.0	29.2	31.4	27.7	57.4	U	U
Uttar Pradesh	44.7	37.4	11.2	17.2	22.7	19.8	60.3	19.4	59.0	59.5	16.1
East											
Bihar	36.8	30.7	12.1	19.0	23.0	10.7	51.6	18.1	62.6	60.9	21.8
Orissa	61.6	53.8	14.1	20.5	41.1	36.1	45.7	30.2	53.3	48.2	21.3
West Bengal	75.3	70.4	31.5	33.0	74.7	34.2	40.0	53.6	56.8	U	U
Northeast											
Arunachal Pradesh	48.9	31.9	19.9	21.3	33.3	22.5	73.9	35.8	39.7	53.9	11.2
Assam	49.3	34.9	11.1	17.9	35.2	19.4	65.0	39.2	50.4	52.2	10.8
Manipur	63.4	48.0	23.0	40.4	63.1	29.1	70.4	50.0	30.1	33.6	8.8
Meghalaya	51.8	30.0	29.6	36.9	40.7	9.7	18.0	56.3	45.5	50.8	18.9
Mizoram	88.9	42.5	48.9	61.5	24.5	56.4	45.5	64.3	28.1	41.3	2.2
Nagaland	39.3	33.0	6.0	22.2	24.6	3.8	61.1	43.5	28.7	32.4	12.7
Tripura	64.9	58.7	30.7	33.5	*	19.0	47.9	65.0	48.8	46.0	17.5
West											
Goa	95.4	83.4	86.8	88.4	41.4	74.9	10.8	33.9	35.0	32.5	15.3
Gujarat	75.7	62.7	35.6	42.5	20.7	49.8	36.3	22.9	50.1	48.2	18.9
Maharashtra	82.7	71.0	43.9	53.2	41.7	64.1	37.1	25.0	54.2	48.5	20.2
South											
Andhra Pradesh	86.3	74.8	32.8	49.3	32.5	45.0	70.5	47.8	49.1	U	U
Karnataka	83.5	69.8	37.5	50.9	34.0	52.2	65.6	38.2	54.3	47.6	17.4
Kerala	97.3	89.8	87.8	89.7	37.8	54.4	59.2	69.3	28.5	27.4	11.6
Tamil Nadu	94.2	90.1	63.4	71.2	27.1	64.9	55.8	56.5	48.2	U	U

U: Not available * Percentage not shown; based on fewer than 25 children

[6]Allopathic doctor or nurse/midwife

[7]For children who had diarrhoea in the past two weeks, percent receiving a solution made from an Oral Rehydration Salt (ORS) packet or a Recommended Home Solution (RHS) made from sugar, salt and water

[8]Percent who have received BCG, measles and three doses of DPT and polio vaccines

[9]Underweight assessed by weight-for-age, stunting assessed by height-for-age, wasting assessed by weight-for-height; undernourished children are those more than 2 standard deviations below the median of the International Reference Population, recommended by the World Health Organization

Figure 2: The Table in the NFHS Final Report

Table 1: The first 10 rows of our filtered dataset (values are percentages)

State	Illiteracy	Access to Water	Antenatal	Births in Facility	Immunization
India	56.7	68.2	62.3	25.5	35.4
Delhi	29.2	99.5	82.4	44.3	57.8
Haryana	54.1	73.0	72.7	16.7	53.5
Himachal Pradesh	42.6	57.6	76.0	16.0	62.9
Jammu & Kashmir	48.2	57.3	79.5	21.9	65.7
Punjab	48.0	98.6	87.9	24.8	61.9
Rajasthan	74.6	57.3	31.2	11.6	21.1
Madhya Pradesh	65.7	55.8	52.1	15.9	29.2
Uttar Pradesh	68.5	74.3	44.7	11.2	19.8
Bihar	71.4	63.6	36.8	12.1	10.7

dren (i.e. height and weight). Lastly, the Village Questionnaire collected data on the amenities accessible to villages, such as electricity, water, transport, and educational and health facilities (IIPS/India (1995)).

A three-stage sample design was used in each state: first selecting cities, then blocks, and finally households. These selections were based on enumerations of the cities and blocks from India's 1991 census. The cities were divided into three strata: (1) self-selecting cities (i.e. volunteers), (2) district headquarters, (3) all other cities. In each strata, selection of the required number of blocks was followed by selection of an average of 20 households. This accounts for possible bias in population densities of different cities, as well as geographical/regional bias from different locations (IIPS/India (1995)).

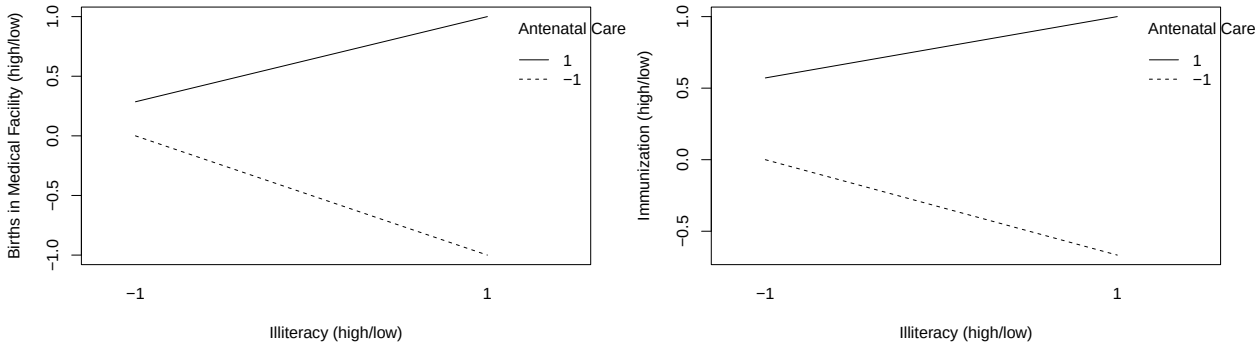
The methodology was prone to one major weakness, accessibility of the survey. Data collection teams were noted to struggle in reaching sampling units located in hilly regions, due to lack of roads and other transportation. For example, in the state Uttar Pradesh, teams covered the sampling units on foot. Further, security was another concern. During December 1992, communal riots caused the presence of bandits in some villages. In this case, sampling units in Madhya Pradesh were avoided, and the data collection could not be fully completed. Moreover, unseasonal rains that were not forecast delayed data collection in Karnataka, Kerala, and other Northeastern states. Lastly, teams failed to establish contact with some households, and as many as possible were attempted a second time at a later date to minimize the nonresponse rate, however in some cases it was unavoidable (IIPS/India (1995)).

2.3 Visualization

For brevity, we will refer to the percentages of illiterate females, households with drinking water, mothers receiving antenatal care, births delivered in a health facility, and fully immunized children, as illiteracy rate, drinking water rate, antenatal care rate, births in health facility rate, and immunization rate.

Now we study is the interaction between our variables. That is, if a change in one is affecting a change in another, on some outcome. To aid in this process, we define our two yields, or outcomes, as the immunization rate and birth in health facility rate of children. This makes sense as these two variables serve as indicators of the wellbeing of children in the country. Then our 3 predictors are the illiteracy rate, access to water rate, and antenatal care rate.

Moving forward, we will define the 5 rates above as being high, if they are greater than the grand mean (corresponding to the row for India), and low if they are less. We will create 5 new indicator variables for each rate, taking the value 1 if the rate is high, and -1 if the rate is low. This will allow us to treat the data as coming from a 2^k factorial design, where $k = 3$ is the number of levels, with 2 outcomes. The interaction plots for each of the $k! = 3! = 6$ combinations below in Figure 3 show that we have interactions between the following of variables (because the lines are not parallel). This is represented by the directed acyclic graph in figure 4, where a variable points to a variable that it influences, treating the outcome variables as influenced by the predictor variables.



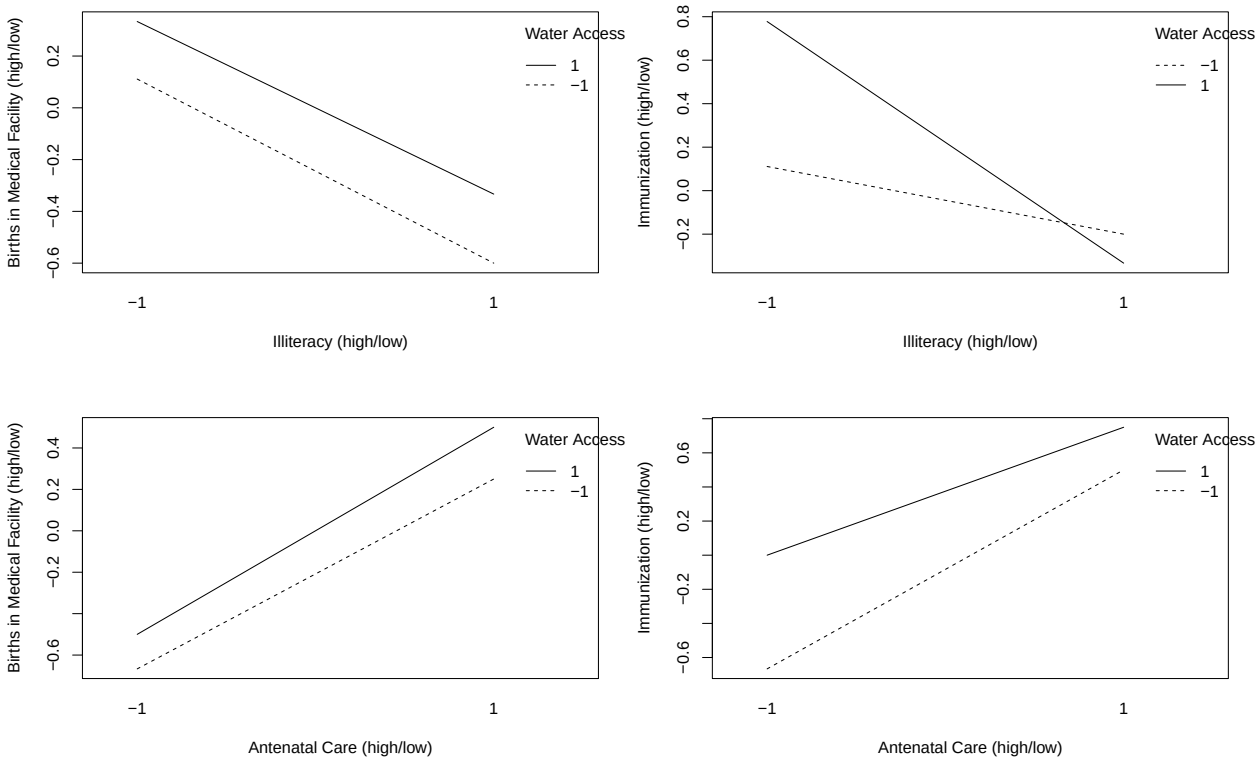


Figure 3: Interaction plots for each combination of predictors on outcomes

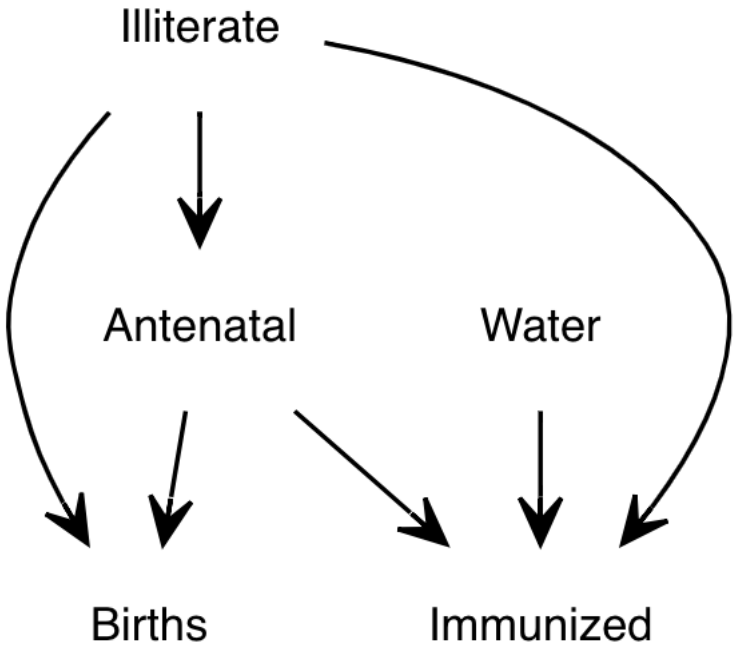


Figure 4: A directed acyclic graph showing the interaction between the 5 variables

3 Results

For all comparisons made in this section, we look at each state as a different case, i.e. an example of the conditions met in that state. Thus, we can compare between states, but not within a state, because of the nature of the data.

The first comparison we make is between the percentage of illiterate females and the percentage of vaccinated children. This will tell us the effect education of the mother has on the benefit of their child. The graph below in Figure 5 shows that as illiteracy rate goes up, the rate of vaccination goes down. This seems indicative that educated women are more likely to vaccinate their children.

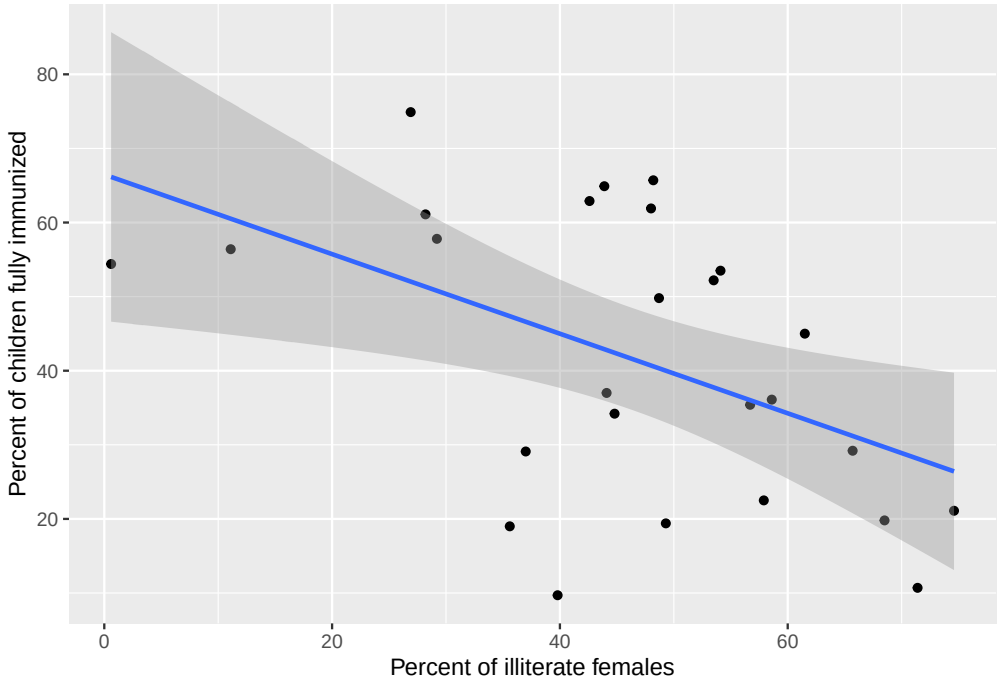


Figure 5: Graph of Illiteracy vs Immunization Rate

Furthermore, we also check the relation between percentage of mothers receiving antenatal care and percentage of births delivered in a health facility. After plotting the two variables together and fitting a linear model in Figure 6, we see that states with a high rate of antenatal care also have a high rate of births in a health facility. From this we can interpret that states with a priority on health care take better care of their pregnant women and children. Thus it seems likely, and is indeed the case as demonstrated by Figure 7, that states with a high illiteracy rate also have a low rate of births in a health facility. Note that the outcomes also appear to have an affect on each other, where as shown by the graph Figure 7, as states with a higher birth in a health facility rate, also have a high vaccination rate.

The overall findings from the survey are summarized in this section, as they pertain to our variables of interest. 57% of all females age 6 and above are illiterate, with about 9% having a high school education or more. 68% of households have access to some form of drinking water, from a pump or pipe. We have that 62% of mothers received special care while they were pregnant, but only 26% of mothers gave birth in medical facilities. Finally, only 35% of children were immunized (IIPS/India (1995)).

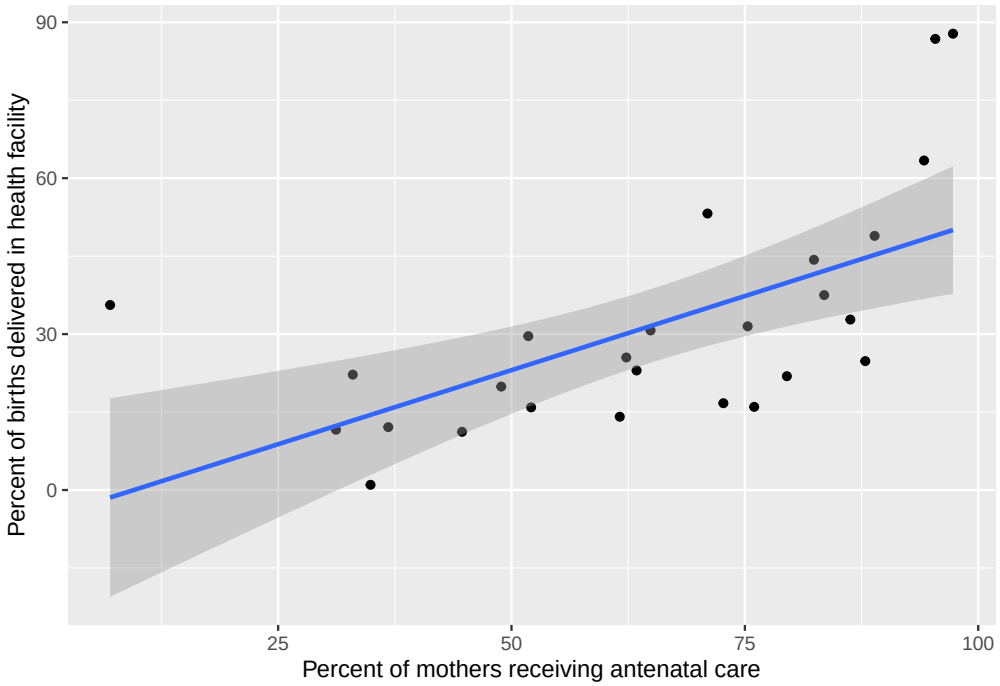


Figure 6: Graph of Antenatal Care vs Births in a Health Facility

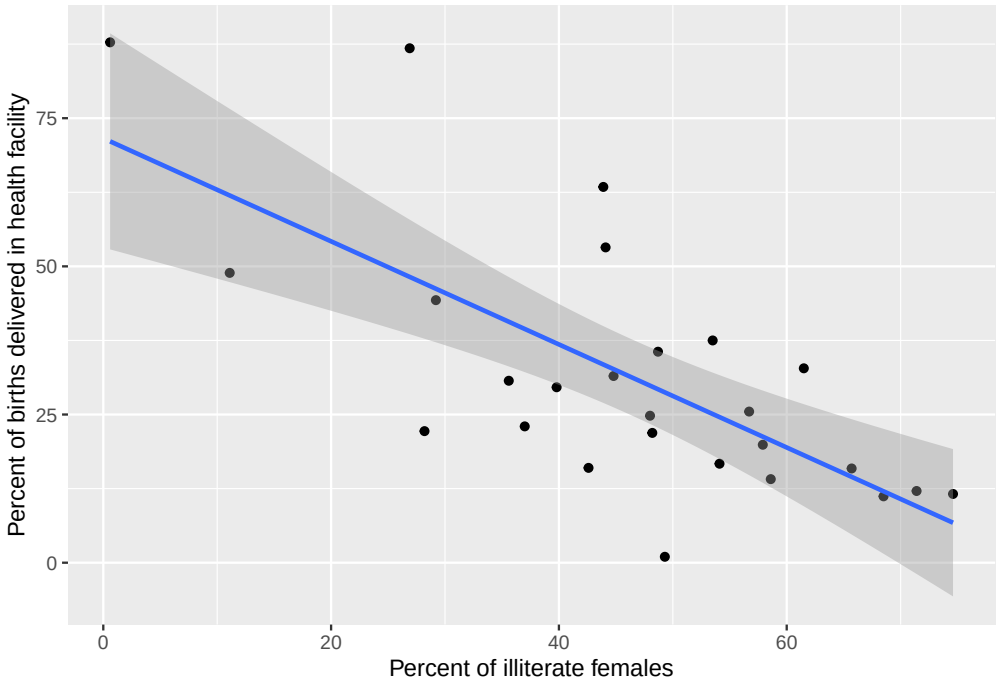


Figure 7: Graph of Illiteracy vs Births in a Health Facility

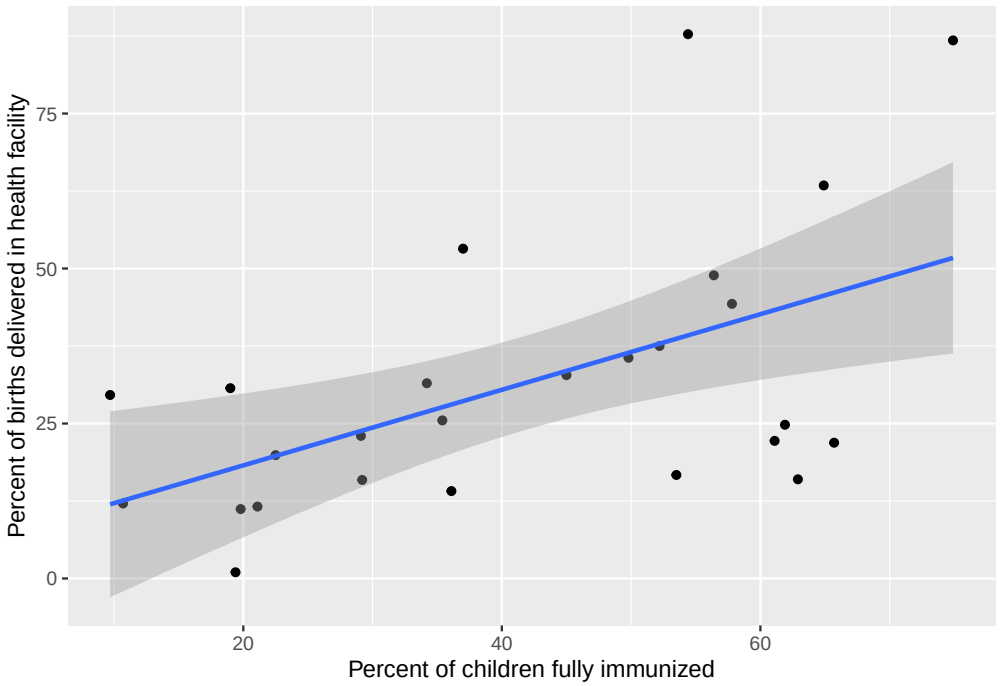


Figure 8: Graph of Immunization Rate vs Births in a Health Facility

4 Discussion

Using the above findings, we observe that key areas for improvement in India's health sector are availability of medical facilities, as we know from above that those receiving antenatal care would give birth in those facilities if they were accessible. Further, the low vaccination rate can be combatted by spreading awareness of the importance of vaccines, outside of schools, because the majority of females could not attend school to learn about them. A core, systemic, issue in India is the high poverty rate, which likely is responsible for many of the above issues alone. However, the NFHS survey failed to account for this variable. If they had recorded the economic status of their sampling units, another form of bias could be eliminated, and the results purified. Thus, in the future, accounting for this source of variation (both within and between states) can lead to more reliable results.

As the study was conducted in 1992, we can compare our derived results with the actual implementations used by the Indian government since the survey took place. From (water.org (2020)), just over 6% of the population is left without access to water, which is a significant improvement from 100-68=32% without access in 1992. Furthermore, a huge step forward has been taken in literacy, with 65% of females now being literate, as compared to over half the population of females being illiterate previously (India (2020a)). Another successful indicator is the advancements made in the health sector. In terms of immunizations, India became polio-free in 2014, and eliminated maternal tetanus in 2015 (UNICEF (2017)). Furthermore, 65% of children in India now receive vaccination during their first year of life, as compared to 35% in 1992. This was accomplished through the aid of organizations like UNICEF, whose primary goal is to ensure the health and safety of children in the country. This also demonstrates the substantial affect that international aid has on the development of a third world country. India has been continuously renewing the NFHS, with 3 more since the first one, conducted in 1998, 2005, and 2015, with further expansion on the horizon (IIPS (2009)).

An issue with the survey that should be considered in the future is the lack of information collected on men. The survey primarily focused on women and children, as they are the key population indicators, but this leads to any implementations based on this survey being biased towards the demographics for who information was collected on, omitting a significant chunk of the population, males. In the twentieth century

(1900s), the population ratio of females to males was 972 females to 1000 males, and that ratio increased by 10% as measured in 2011 (India (2020b)). Another major holdback in conducting this survey was the large cost associated with all the necessary procedures, which also came with several hurdles as discussed in (2.2 Methodology).

To facilitate the delivery of the survey, the NFHS team may consider administering it online. When the survey was initially conducted in 1992, nearly the entire population did not have access to any form of internet communication, but now, as of 2019, over 41% of the population is online (Bank (2019)). As demonstrated by organizations like Statistics Canada, it's the least expensive mode, and it can be used to reach a very large number of people (Government of Canada (2021)). The respondent can fill the survey at the moment of his choice. Self-completed (i.e. online) questionnaires are also easier when the survey questions are sensitive, such as some of the topics covered by the NFHS. The downside of this mode is that response rates are lower than for other collection mode and the quality of collected data can suffer (Government of Canada (2021)). Despite the detriments, it is still advisable out of the interest of safety, especially to avoid key areas that were missed because of uncharted terrain and potential threats in various villages.

In conclusion, we began with the goal to demonstrate how data analysis can and should be used to gain an understanding of the overall population status of a country. Then we discuss how to use this information to make decisions that better the lives of the citizens. Third world countries can and should acquire large datasets representative of their population, and employ the techniques discussed in this paper, and the NFHS Final Report, to induce modifications to their public policies in such a way that the needs of the many are satisfied. Thus illuminating the path towards further development of that country with the goal of economic prosperity. In short, the inductive application of data deduction fosters the grounds for which a country may stem its roots, and make a plan to grow.

Appendix

A Datasheet

Motivation

1. *For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.*
   - The dataset was created to strengthen the capabilities of the population research centres in India.
2. *Who created the dataset (for example, which team, research group) and on behalf of which entity (for example, company, institution, organization)?*
   - The dataset was created by the Ministry of Health and Family Welfare, alongside the International Institute for Population Sciences.
3. *Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.*
   - The creation was funded by the Government of India.
4. *Any other comments?*
   - The NFHS was one of the most complete surveys of its kind ever conducted in India, at the time (1992). As a result, it provided a huge amount of population-wise information that could otherwise not be determined.

Composition

1. *What do the instances that comprise the dataset represent (for example, documents, photos, people, countries)? Are there multiple types of instances (for example, movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.*
   - The instances represent the states of India. The types are: North, Northeast, West, East, South, and Central, corresponding to different regions of the country.
2. *How many instances are there in total (of each type, if appropriate)?*
   - There are 26 instances
3. *Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (for example, geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (for example, to cover a more diverse range of instances, because instances were withheld or unavailable).*
   - Since every state, at the time of this dataset, was included, the dataset does contain all possible instances.
4. *What data does each instance consist of? "Raw" data (for example, unprocessed text or images) or features? In either case, please provide a description.*
   - Each instance consists of 23 continuous variables (raw data), and 1 feature (the state represented). TODO: DEFINE ALL THE COLUMNS AND GIVE EXPLANATIONS?
5. *Is there a label or target associated with each instance? If so, please provide a description.*
   - The first column of each instance corresponds to the state it represents, which is the label associated with it.

6. *Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (for example, because it was unavailable). This does not include intentionally removed information, but might include, for example, redacted text.*
    - 5 instances contain cases of missing information. In each case, the missing values are due to unavailable data. Namely, Himachal Pradesh, Madhya Pradesh, West Bengal, Tamil Nadu and Andhra Pradesh are all missing `Percent of Children Stunted` and `Percent of Children Wasted`.
7. *Are relationships between individual instances made explicit (for example, users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.*
    - There are no relationships between individual instances.
8. *Are there recommended data splits (for example, training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.*
    - There are no recommended data splits.
9. *Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.*
    - There are no errors, sources of noise, or redundancies in the dataset.
10. *Is the dataset self-contained, or does it link to or otherwise rely on external resources (for example, websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (that is, including the external resources as they existed at the time the dataset was created); c) are there any restrictions (for example, licenses, fees) associated with any of the external resources that might apply to a dataset consumer? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.*
    - The dataset is self-contained.
11. *Does the dataset contain data that might be considered confidential (for example, data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals' non-public communications)? If so, please provide a description.*
    - There is no confidential data, and the dataset is publicly available.
12. *Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.*
    - Columns that might cause anxiety include: `Percent of women age 20-24 married before age 18`, due to systemic norms in the Western society, and `Percent of living children under four years of age wasted` because of the dismal implications of that heading.
13. *Does the dataset identify any sub-populations (for example, by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.*
    - The dataset entirely comprises women and children 4 years and less.
14. *Is it possible to identify individuals (that is, one or more natural persons), either directly or indirectly (that is, in combination with other data) from the dataset? If so, please describe how.*
    - It is not possible to identify individuals in any way.
15. *Does the dataset contain data that might be considered sensitive in any way (for example, data that reveals race or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.*

- Sensitive columns may include but are not limited to: `Percent of women age 20-24 married before age 18`.

16. *Any other comments?*
    - None.

Collection process

1. *How was the data associated with each instance acquired? Was the data directly observable (for example, raw text, movie ratings), reported by subjects (for example, survey responses), or indirectly inferred/derived from other data (for example, part-of-speech tags, model-based guesses for age or language)? If the data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.*
    - The data associated was reported by subjects through interviews from all 25 states.
2. *What mechanisms or procedures were used to collect the data (for example, hardware apparatuses or sensors, manual human curation, software programs, software APIs)? How were these mechanisms or procedures validated?*
    - Manual human curation, i.e. the process of humans taking down the information given during interviews, was used to collect the data.
3. *If the dataset is a sample from a larger set, what was the sampling strategy (for example, deterministic, probabilistic with specific sampling probabilities)?* -Geographic stratification subdivided each state into regions, from which villages were further stratified following: village size, distance from nearest town, proportion of nonagricultural workers, proportion of the population belonging to scheduled casts, and female literacy. Primary Sampling Units were selected systematically, with probability proportional to size of the strata.
4. *Who was involved in the data collection process (for example, students, crowdworkers, contractors) and how were they compensated (for example, how much were crowdworkers paid)?*
    - The operations were supervised by the senior field staff of the concerned CO and PRC in each state. Compensation data is not available.
5. *Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (for example, recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.*
    - The data was collected over April 1992 to September 1993.
6. *Were any ethical review processes conducted (for example, by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.*
    - Ethical review processes were not conducted.
7. *Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (for example, websites)?*
    - We obtained the data via the Demographic and Health Surveys website: dhsprogram.com
8. *Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.*
    - The individuals voluntarily interviewed with data collectors. The notice of data collection is not available.

9. *Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.*
    - The individuals consented to the collection and use of their data. The exact language to which consent was granted is not available.
10. *If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).*
    - A mechanism to revkoe consent was not provided.
11. *Has an analysis of the potential impact of the dataset and its use on data subjects (for example, a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.*
    - An analysis of the potential impact of the dataset and its use on data subjects was not conducted.
12. *Any other comments?*
    - Note that the population primarily consists of younger people due to the population composition of India in 1992.

Preprocessing/cleaning/labeling

1. *Was any preprocessing/cleaning/labeling of the data done (for example, discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remaining questions in this section.*
    - The data was originally obtained in PDF format. The table from the survey PDF was converted to a usable data frame in R using the library pdftools for R.
2. *Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (for example, to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.*
    - The raw data obtained from the PDF is saved in inputs/data/raw_data.csv
3. *Is the software that was used to preprocess/clean/label the data available? If so, please provide a link or other access point.*
    - R Software is avalaible at https://www.R-project.org/
4. *Any other comments?*
    - The library used to help import data from PDF is available at https://docs.ropensci.org/pdftools/

Uses

1. *Has the dataset been used for any tasks already? If so, please provide a description.*
    - The dataset has not been used for other tasks yet.
2. *Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.*
    - https://github.com/haqbilal/India-1992-DHS-State-Findings
3. *What (other) tasks could the dataset be used for?*
    - The dataset can be used for examining the state of women and children in India in 1992-1993.

4. *Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a dataset consumer might need to know to avoid uses that could result in unfair treatment of individuals or groups (for example, stereotyping, quality of service issues) or other risks or harms (for example, legal risks, financial harms)? If so, please provide a description. Is there anything a dataset consumer could do to mitigate these risks or harms?*
    - The cleaning process is very specific to the way this table was formatted in the original PDF and may not work on other tables.
5. *Are there tasks for which the dataset should not be used? If so, please provide a description.*
    - The dataset would not be appropriate for purposes other than examining the plight of women and children in 1992 India.

Distribution

1. *Will the dataset be distributed to third parties outside of the entity (for example, company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.*
    - No, the dataset is openly available and being used for personal uses only.
2. *How will the dataset be distributed (for example, tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?*
    - The dataset will be distributed using Github.
3. *When will the dataset be distributed?*
    - The dataset will be distributed in April 2022.
4. *Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/ or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.*
    - The dataset will be released under the MIT license
5. *Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.*
    - There are no restrictions
6. *Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.*
    - No such controls or restrictions are applicable.
7. *Any other comments?*

Maintenance

1. *Who will be supporting/hosting/maintaining the dataset?*
    - Bilal Haq and Ritvik Puri
2. *How can the owner/curator/manager of the dataset be contacted (for example, email address)?*
    - Can be contacted via github
3. *Is there an erratum? If so, please provide a link or other access point.*

- There is no erratum available currently.

4. *Will the dataset be updated (for example, to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to dataset consumers (for example, mailing list, GitHub)?*
    - Currently there is no plan of updating the dataset. If there are updates in the future, it will be done through Github.
5. *If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (for example, were the individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.*
    - The dataset was made via survey findings conducted in India. There are no applicable limits as the people took part voluntarily.
6. *Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to dataset consumers.*
    - The older versions would not be hosted. Dataset consumers will be able to check whether the dataset has been updated through Github commit history.
7. *If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to dataset consumers? If so, please provide a description.*
    - There is no mechanism for accepting contribtuions from other users as of now.

B Additional details

Code and data are available at: https://github.com/haqbilal/India-1992-DHS-State-Findings

You can clone the repository and reproduce the paper as needed, and are encouraged to conduct your own analysis. We used 00-simulation.R to create a simulation of the data, and develop a plan to assess. Then we gathered the data using 01-gather_data.R and cleaned it using 02-clean_and_prepare.R. Each script is documented, explaining each step in the process. Please note that obtaining and cleaning the dataset is difficult in an OCR'd PDF because of misrecognition of characters, the table spanning two pages, and the column headings spanning multiple lines.

References

Allaire, JJ, Yihui Xie, Jonathan McPherson, Javier Luraschi, Kevin Ushey, Aron Atkins, Hadley Wickham, Joe Cheng, Winston Chang, and Richard Iannone. 2022. *Rmarkdown: Dynamic Documents for r*. https://github.com/rstudio/rmarkdown.

Bank, The World. 2019. "Individuals Using the Internet (." *Worldbank.org*. International Telecommunications Union. https://data.worldbank.org/indicator/IT.NET.USER.ZS?locations=IN.

Government of Canada, Statistics Canada. 2021. "3.3.1 Data Collection Methods," September. https://www150.statcan.gc.ca/n1/edu/power-pouvoir/ch2/methods-methodes/5214773-eng.htm.

Iannone, Richard. 2022. *DiagrammeR: Graph/Network Visualization*. https://CRAN.R-project.org/package=DiagrammeR.

Iannone, Richard, and Mauricio Vargas. 2022. *Pointblank: Data Validation and Organization of Metadata for Local and Remote Tables*. https://CRAN.R-project.org/package=pointblank.

IIPS. 2009. "National Family Health Survey." *NFHS, India*. IIPS. http://rchiips.org/nfhs/.

IIPS/India, International Institute for Population Sciences-. 1995. "India National Family Health Survey 1992-93," August. https://dhsprogram.com/publications/publication-frind1-dhs-final-reports.cfm.

India, Know. 2020a. "Literacy." *Know India*. india.gov.in. https://knowindia.india.gov.in/profile/literacy.php.

———. 2020b. "Sex Ratio." *Know India: National Portal of India*. india.gov.in. https://knowindia.india.gov.in/profile/sex-ratio.php.

Müller, Kirill. 2020. *Here: A Simpler Way to Find Your Files*. https://CRAN.R-project.org/package=here.

Ooms, Jeroen. 2022. *Pdftools: Text Extraction, Rendering and Converting of PDF Documents*. https://CRAN.R-project.org/package=pdftools.

Pedersen, Thomas Lin. 2020. *Patchwork: The Composer of Plots*. https://CRAN.R-project.org/package=patchwork.

R Core Team. 2020. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

UNICEF. 2017. "Immunization." *Unicef.org*. United Nations Children's Fund. https://www.unicef.org/india/what-we-do/immunization.

water.org. 2020. "Water in India - India's Water Crisis & Sanitation Issues in 2021." *Water.org*. Water.org. https://water.org/our-impact/where-we-work/india/.

Wickham, Hadley. 2007. "Reshaping Data with the reshape Package." *Journal of Statistical Software* 21 (12): 1–20. http://www.jstatsoft.org/v21/i12/.

———. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Xie, Yihui. 2014. "Knitr: A Comprehensive Tool for Reproducible Research in R." In *Implementing Reproducible Computational Research*, edited by Victoria Stodden, Friedrich Leisch, and Roger D. Peng. Chapman; Hall/CRC. http://www.crcpress.com/product/isbn/9781466561595.

———. 2015. *Dynamic Documents with R and Knitr*. 2nd ed. Boca Raton, Florida: Chapman; Hall/CRC. https://yihui.org/knitr/.

———. 2022. *Knitr: A General-Purpose Package for Dynamic Report Generation in r*. https://yihui.org/knitr/.