DATAWISE
Intelligence. Insights. Infinity.

An Application of Demographic Data as a Surrogate for Epidemiological Profiling in India

+91 40 40204837

info@mydatawise.com

Datawise
Hyderabad, India

The absence of nation-wide profiling of the health condition of people in India has lead to the Government of India tending to take a broad brush approach with very little reference to the specific health condition of people in a region. This causes inefficient health management mechanism with the government erring on the side of caution in order to ensure coverage. The key challenges arising out of inadequate and poor quality of data are uneven health care spending, and mismatch between disease profiles and provision of care.

As a result, a significant gap exists in the ability of policy makers to understand the implications for public health in those geographies which are not extensively covered. A resultant approach is a broad- brush health intervention which is inefficient while achieving its objective of universal health coverage.

Many of the data collected through the Census have a strong correlation to the health status of the individual. A question therefore arises as to whether it is possible to use demographic data available through the Census, for the purpose of deriving healthcare indicators, and thus using some of the data available through census as a means of deriving epidemiological data.

The results of the mapping of each of the demographic data collected from Census, initially within the NFHS data set, and subsequently, mapped to the Census dataset provides an insight into the potential disease that individuals in various geographies are likely to suffer. This data, once extrapolated to the ward level in a city provides a level of granularity which helps in a better understanding of the health parameters that are actionable.

[1]

INTRODUCTION

One of the key challenges of public health policy in India has been the absence of nation-wide profiling of the health condition of people. In a country as vast as India, in the absence of specific epidemiological profiling, most health interventions by the Government of India tend to take a broad brush approach with very little reference to the specific health condition of people in a region. This leads to an inefficient health management mechanism with the government erring on the side of caution in order to ensure coverage. The key challenges arising out of inadequate and poor quality of data are uneven health care spending, and mismatch between disease profiles and provision of care.

Multiple studies are conducted in India on a regular basis to determine the health status of individuals. The most important ones that have been conducted on a large scale are the Annual Health Survey

[1] This Paper is a summarized version of the Paper presented by Datawise at IIM Ahmedabad by K Vinay Kumar and K Vijay Kumar.

(AHS) which is an exercise sponsored by the Census of India, and the National Family Health Survey (NFHS Series III, conducted in 2006). While these studies have helped in understanding the general health condition, they suffer from severe limitations. The AHS survey provides information on vital public health measures such as CBR, CDR, MMR, etc. but this is limited to the nine states (Chhattisgarh, Uttar Pradesh, Assam, Odisha, Madhya Pradesh, Bihar, Jharkhand, Rajasthan and Uttarakhand) identified as being most critical to measure the health parameters. On the other hand, the NFHS survey is extensive in its geographical coverage, but is limited in the number of households it covers in each geography (the total coverage, for example in the Birth Surveys is only about 260 thousand respondents). Further, in order to mask HIV case identification, the district level data has not been provided in the NFHS data making comparability difficult.

"As a result, a significant gap exists in the ability of policy makers to understand the implications for public health in those geographies which are not extensively covered. A resultant approach is a broad-brush health intervention which is inefficient while achieving its objective of universal health coverage."

As a result, a significant gap exists in the ability of policy makers to understand the implications for public health in those geographies which are not extensively covered. A resultant approach is a broad-brush health intervention which is inefficient while achieving its objective of universal health coverage. In future surveys, the NFHS and AHS are proposed to be merged into a single comprehensive survey which would provide data at a district level granularity. However, the results of these survey are expected to be available only late in 2015 or in 2016. Even when compiled, they would have limited utility since no comparable baseline data would be available.

LITERATURE REVIEW

Many of the data collected through the Census have a strong correlation to the health status of the individual. A question therefore arises as to whether it is possible to use demographic data available through the Census, for the purpose of deriving healthcare indicators, and thus using some of the data available through census as a means of deriving epidemiological data.

It is well researched that improved education results in improved healthcare (Ross and Wu, 1995). Other research similarly indicates that socio-economic class has an impact on the state of health (Adler, and Ostrove, 2006). Other research has shown that there is an inverse relation between literacy levels and obesity (Ferrer, McMunn, Dommarco and Brunner, 2014).

Another study also found a strong correlation between being educated and lower Body Mass Index (BMI) (von Hippel, Lynch and Jamie, 2014). Further, although it was found that as education categories fall, the probability of being overweight increase, the size of the household did not appear to have any impact on obesity (Eidsdóttir, Sigridur, Kristjánsson, Álfgeir, Sigfúsdóttir, Inga D, Garber, Carol E, Allegrante, John P, 2013). This has also been corroborated in a study in China, which has similar demographic patterns to India (Xiao, Zhao, and Wang, 2013).

Studies on Asthma have suggest that younger males and those within less educated families may be more vulnerable to aeroallergens as reflected by hospitalization for asthma (Cakmak, Dales, Judek, and Coates, 2005).

Similar studies on diabetes have shown strong positive linkages with employment (Ruston, Smith, and Fenrnando, 2013). Epidemiological surveys on tuberculosis (TB) in China showed that pulmonary TB was impacted by socio-economic and geographical factors (Ngui, Lim, Ai ian, Chuen, Sek, 2010). Univariate analysis demonstrated that low level of mother's education, non-working parents low household income were significantly associated with the high prevalence of Anemia ((Ngui, Lim, Ai ian, Chuen, Sek, 2012). Socioeconomic factors such as wealth, education, employment, occupation of the partner, presence of toilet facility, and preventive health measures were strongly related to the Hemoglobin levels of women Haverkate, Smits, Meijerink, and Hinta, 2014).

DATA FOR ANALYSIS

The NFHS data was obtained from the third survey conducted in 2005-06 and made public in 2010, and available for research from Demographic and Health Surveys (recode structure DHS V). The data for Census for India was available from the 2011 data which is available to the public from the Census website.

Health data was selected from NFHS at an individual respondent level for the key respondent in a household. A total of about 250 thousand records were obtained, which included multiple records for the same household.

For comparison purposes, Census data was sought. However, in order to ensure that there is a level of granularity that would make it comparable, we had to consider only ward level data. This was available on a comparable basis with NFHS for the following eight cities: Delhi, Mumbai, Kolkata, Chennai, Hyderabad, Indore, Meerut, and Nagpur. The data was available at an aggregated level for a total of 1,311 wards. Of these, data for 25 wards was not considered for the purpose of the analysis where data was outliers in respect of average household size (e.g., Hyderabad, Ward 1 has an average household size of more than 10), or sex ratios were skewed (e.g., Mumbai Ward 313 had a sex ratio of 3.04), or and only the remaining 1,286 wards were considered, distributed in Table 1 as follows:

Table 1: Description of cities, the number of wards, and population

	Number of Wards	Population
Chennai	155	4,646,732
Hyderabad	107	3,936,561
Indore	213	2,424,348
Kolkata	138	4,447,138

Meerut	263	1,725,923
Mumbai	38	3,067,558
Nagpur	362	3,168,234
New Delhi	10	140,315

Clearly, the entire ward level information was not available for the two mega-cities of Mumbai and New Delhi, nor was the population indicators complete. However, for the purpose of completeness of analysis, these cities and their populations were included at a ward level.

A comparable dataset from NFHS yielded a total of 14,217 records for the same cities. For comparison purposes, only urban population was considered in both the cities.

Parameters were identified from the NFHS database on the basis of a broad classification as follows:

1. Parameters which were available as indicators of the demographic markers of the respondent included respondent age, State, Highest educational level, Source of drinking water, Type of toilet facility, Literacy, Number of living children, Partner's educational attainment, Partners age, Child's age in months, Child's weight in kilograms, City\Town\Countryside, Acres of agricultural land, Household has cows/bulls/buffalo, Household has a BPL card, Household structure, and Marital status.
   However, on a comparable basis only the indicators for household size, literacy levels, and employment levels were available across both the census data and NFHS data. Each of these indicators is also available in the Census and can be link the NFHS data file to the census file to arrive at similar demographic profiles. Each of the above indicators also provides an indirect indication of the potential health status of the individual.
2. Parameters which indicate the actual health status of the individual. We have included the following eight health parameters that could be identified as direct evidence of health of an individual. These include 4 indicators of health, and 4 indicators of occurrence of specific disease:
   a. Body-Mass Index
   b. Rohrer's Index
   c. Anemia level
   d. Hemoglobin Levels
   e. Incidence of Diabetes
   f. Incidence of Asthma
   g. Incidence of Thyroid or Goiter
   h. Incidence of Tuberculosis

Each of these indicators provides a means to correlate the indicators in the incidents which impact health with the actual status of health.

Health data from NFHS was further categorized into two broad indicators – healthy, or unhealthy, depending on the health indicators. In the case of Body Mass Index (BMI), a too low BMI, or a too high BMI both were considered to be an indicator of poor health. Rohrer's Index, which is widely considered to be a derivative of BMI was not used for analysis. Anemia levels were broken into anemic or not anemic, hemoglobin levels were available in a range of between

28 to 198, and were into a range of 2 indicators as abnormal or normal. Diabetes, Asthma, Thyroid, and Tuberculosis were considered on a simple existing (0) or not existing (1) scale

ANALYSIS

The results of the mapping of each of the above two sets of parameters, initially within the NFHS data set, and subsequently, mapped to the Census dataset provides an insight into the potential disease that individuals in various geographies are likely to suffer. This data, once extrapolated to the ward level provides a level of granularity which helps in a better understanding of the health parameters that are actionable.

First, the NFHS data was split into two sets, divided along cities. The correlation between each of the health and demographic parameters was first compared across one set, and then validated against the other set of data. Results of only those sets of data which were validated across the first as well as the second set of data were considered as acceptable for further analysis. The division of the data across cities was as shown in Table 2 below:

Table 2: Segregation of NFHS data into the two sets

Set 1	Set 2
Chennai	Hyderabad
Delhi	Mumbai
Indore	Meerut
Kolkata	Nagpur

Population in each of the set of data was seen to be comparable both in terms of size as well as the mix of cities, and therefore, considered good for the purpose of validation. A chi-square test showed significant relationship between the following set of parameters, both for the Set 1 as well as when validated against Set 2. The details of the chis-square tests are shown in Table 3 below:

Table 3: Association between the health indicators and the demographic indicators for the two sets of NFHS data

	Set 1			Set 2		
	Literacy Levels	Household Size	Employment Status	Literacy Levels	Household Size	Employment Status

Body-Mass Index	Significant	Significant	Significant	Significant	Significant	Significant
Rohrer's Index	Significant	-	-	-	-	-
Anemia	-	Significant	Significant	-	Significant	Significant
Hemoglobin	Significant	-	-	-	-	-
Diabetes	-	-	-	-	-	-
Asthma	-	-	Significant	-	-	Significant
Thyroid or Goiter	Significant	Significant	Significant	Significant	-	Significant
Tuberculosis	Significant	Significant	-	-	Significant	-

*p value<0.05

For the data above where there were significant relationships established in both the first as well as the second set of data, Spearman's rank correlation was conducted on each of the demographic parameters with each of the health parameters. Strong correlations were found for each of the following parameters, as explained below.

As can be expected, Employment status and Tuberculosis (TB) are negatively correlated with each other (-0.67). Employment status and BMI are also negatively correlated to each other (-0.59), possibly indicating that working persons tend to take poorer care of their health, possibly due to obesity, than unemployed persons due to malnourishment. Employment status and Anemia are negatively correlated to each other (-0.70), possibly indicating that persons who are employed possibly have access to better nourishment. The probability of an employed person or unemployed person suffering any of the health conditions are as shown in Table 4 below:

Table 4: Probability of TB, poor BMI, or Anemia for employed versus unemployed persons

	TB	BMI	Anemia
Employed	0.13%	14.0%	3.54%
Unemployed	0.27%	35.3%	14.15%

Household size and Tuberculosis are negatively correlated to each other (-0.68) which shows that larger households have a greater probability of suffering from TB. Household size and Anemia are also negatively correlated to each other (-0.73), indicating again that larger households are likely to suffer from poorer nourishment. The household size versus the probable incidence of for TB or Anemia are shown in Table 5 below:

Table 5: Probability of TB, or Anemia for small, medium and large household sizes

	TB	Anemia
Small Household size	0.19%	14.46%

Medium Household size	0.13%	2.55%
Large Household size	0.08%	0.68%

Literacy and Thyroid are positively correlated to each other (0.62). That indicates that people who are literate are more likely to have thyroid than people who are not literate. Similarly, literacy and Body Mass Index are positively correlated (0.65), i.e., literate persons are more likely to have an unhealthy Body Mass Index than illiterate persons. The probability of a literate person or illiterate person suffering any of the health conditions are as shown in Table 6 below:

Table 6: Probability of Thyroid or poor BMI for literate versus illiterate persons

	Thyroid	BMI
Literate	1.41%	37.58%
Illiterate	0.17%	11.74%

All of the above are clear indications of probable health condition and can be applied to a given population that is divided based on a similar categorization.

IMPLICATIONS

From the census data, population was categorized on the same basis as the NFHS data, at a ward level, to arrive at an indication of number of employed versus unemployed, literate versus illiterate, and household size (categorized as small, medium and large). A superimposition of the probable incidence of the disease identified above against each of the demographic indicators would ten yield a ward level health statistic. A range of probable incidence of the following health conditions was then derived using the same measures, which is indicated in Table 7 below, at an overall city level:

Table 7: Distribution of probable incidence of various health conditions in cities in India

	Thyroid	TB	BMI	Anemia
Chennai	2.81%-2.82%	0.25%-0.26%	42.64%	3.06%-3.07%
Hyderabad	2.81%-2.82%	0.15%-0.26%	42.64%	0.81%-3.07%
Indore	0.36%-2.91%	0.00%-0.33%	13.33%-42.77%	0.73%-3.20%
Kolkata	2.81%-2.82%	0.14%-0.36%	42.64%	0.81%-17.35%
Meerut	0.30%-2.87%	0.10%-0.52%	13.33%-42.72%	0.75%-12.26%
Mumbai	2.81%-2.82%	0.25%-0.26%	42.64%	3.06%-3.07%
Nagpur	2.73%-2.93%	0.13%-0.34%	42.53%-42.77%	0.79%-3.22%

New Delhi	2.81%-2.82%	0.24%-0.26%	42.63%-42.65%	3.06%-3.07%

The same data has also been derived at a ward level although for the sake of space, it has not been replicated in this paper.

While we have been limited by the availability of complete census data even at a city level, given that there are clear correlations which have been established between some of the demographic indicators and the health indicators points to the exciting possibility of being able to produce similar results across a much larger geography.

About DATAWISE®

DATAWISE® offers a suite of products and solutions suited to the needs of various situations and industries. Solutions provided for one customer are not necessarily suitable for others, and readers are advised to use their own judgment regarding the suitability of these solutions to their business needs.

DATAWISE®'s business analysis services support the full spectrum of clients' needs with services directed mainly at helping companies discover opportunities for improvement through use of analytical capabilities. We offer analytical services in the following areas:

Strategic Analytics: Alignment of strategic intent with actual work, requiring strategic analytics to answer key decision support questions such as whether to enter into a new segment of business or not, whether to reach new customers or not, and other go, no-go decisions.

Behavioral Analytics: Assistance in determining the 'why' and 'how' of a customer behavior (rather than the 'what') in order to ensure that marketing plans yield the desired results through capturing customer events and actions over time and using these stored interactions to determine typical behavior and deviations from that behavior.

Tactical Analytics: Tactical analytics models that we deploy are typically short-term in nature, and are focused on answering immediate questions rather than aligning to a longer-term goal.

Predictive Analytics: We created complex multi-dimensional models that collate data generated from several interaction points to create models that enable the prediction of future events to help identify of both risks and opportunities.

DATAWISE® has also developed proprietary analytics models OPTLIOX™, CREST™, Infinity™ and DATTAB™, catering to specific customer needs.

• Hyderabad • Delhi • Mumbai • Bangalore • Jaipur • USA

www.mydatawise.com mail at info@mydatawise.com