Trends in Child Stunting and Contraception Awareness in Women: Analysis of NFHS-4 Data

1st Akshat Jain
BITS Pilani, Hyderabad Campus

2nd Hasnain Zeenwala
BITS Pilani, Hyderabad Campus

3rd Prathma Chowksey
BITS Pilani, Hyderabad Campus

4th Shivangi Sharma
BITS Pilani, Hyderabad Campus

***Abstract*—This project explores two problems: discovering a set of attributes that predicts rate of Stunting, and analysing the influence of Contraception Usage/Awareness on attributes related to health of women and children. The NFHS-4 data has been mined using K-Means Clustering and Association Rule Mining algorithms. The results obtained related to stunting stress the importance of sanitation and health of women (during pregnancy, and overall well-being and education).**

Contraceptive usage/family planning has been shown to have remarkable consequences on the behaviour of households with respect to children and mothers, children are more likely to be immunised, pregnant women are more likely to get proper health care checkups, and women are more likely to be punctilious about their health (for example, getting checkups for Breast and Cervix Cancer).

I. Introduction

In India, 35 per cent of children younger than five years of age are stunted, a manifestation of chronic under-nutrition. While India's economy has been growing at an impressive rate, the country still has the highest number of stunted children in the world, representing one-third of the global total of stunted children under the age of five years.

Stunting is defined as the percentage of children, aged 0 to 59 months, whose height for age is below minus two standard deviations (moderate and severe stunting) and minus three standard deviations (severe stunting) from the median of the WHO Child Growth Standards. It is associated with an under-developed brain, with long-lasting harmful consequences, including diminished mental ability and learning capacity, poor school performance in childhood, reduced earnings and increased risks of nutrition-related chronic diseases, such as diabetes, hypertension, and obesity in the future.[6]

In the Indian context, therefore, it is important to identify the attributes which predict stunting. A literature review revealed that the association of stunting with various factors has been studied in countries like Rwanda, China, and Pakistan[4][2][5]. This analysis has sought to confirm those relationships in the Indian context and explore if other relationships exist.

NFHS(National Family Health Survey-IV 2015-16) data has been analysed using exploratory data analysis techniques like Association Rule Mining and K-Means Clustering and has yielded interesting relationships: *stunting* has been shown to be related to *breastfeeding practices, sanitation in households, literacy, BMI of women, and scrupulous practices during pregnancy*.

Further, since the health of mothers and correct health-care decisions during pregnancy were found to heavily influence the children's health, and patterns were mined to predict them. And it was discovered that *contraceptive awareness* influenced *antenatal care visits, immunisation of infants, institutional births*, and a variety of other factors related to women and children health (the results will be detailed in the report).

II. Problem Definition

- To explore and discover a set of attributes in the NFHS-4 data that predicts Stunting.
- To explore the influence of Contraception Awareness on attributes related to *health of women and children*.

III. Dataset Description

The National Family Health Survey (NFHS) is one of the large-scale surveys carried out in India. It is conducted in multiple rounds. The NFHS provides data on fertility, infant and child mortality, the practice of family planning, maternal and child health, reproductive health, nutrition, anaemia, utilization and quality of health and family planning services, both at the state and national-level in India. It provides essential data on health and family welfare needed by the Ministry of Health and Family Welfare and other agencies for policy and programme purposes. Technical assistance for the NFHS was provided mainly by ORC Macro (USA) and other organizations for the specific issues that cropped up. The funding for different rounds of NFHS has been provided by USAID, DFID, the Bill and Melinda Gates Foundation, UNICEF, UNFPA, and MOHFW, as well as the GOI.[3]

A. Content Of The Dataset

- 93 continuous valued (ratio level) attributes (including but not limited to):
  - Women Literacy percentages
  - Contraception Method percentages
  - Child Immunisation rates
  - Woman Health Indicators
  - Child Health Indicators
- a total of 30 files, one for each state (including Delhi). Each file contained district wise data for the 93 attributes.

Separate data was available for rural and urban households within the districts.

IV. Data Preprocessing

This section describes the preprocessing techniques used on the dataset, along with the reasons for selecting those techniques.

A. Merging data

The original data set contained a total of 30 files. The data from these files was merged together and a single dataset was formed, containing 1549 rows and 97 columns (93 attributes along with 4 meta attributes) and a total of 14889 NA values.

B. Feature Creation

The following three features were added to the data for ease of data visualisation and analysis:

- State
- Zone (Northern/Southern/Eastern/Western/Central/North Eastern)
- Region (Urban/Rural)

The total number of attributes in the data after the above features were created was 100.

C. Data Cleaning

Data cleaning is the process of preparing data for analysis by removing or modifying data that is incorrect, incomplete, irrelevant, duplicated, or improperly formatted. This data may hinder the process of analysis or provide inaccurate results.

The Dataset contained district wise information about a certain number of households surveyed. Those rows which had 0 households surveyed were removed. This resulted in the removal of 9 rows. The features were divided into various categories, according to the relevance in our study. Data was cleaned by removing those attributes which contained a disproportionate number of 'NA' values. This resulted in the removal of 10 attributes. Finally, rows that contained missing values in at least one category were removed. 248 such rows were removed from the dataset. 393 data cells which still contained missing values were replaced by the means of the features they fell under. The dataset then contained 1292 rows, 90 columns and 0 NA values.

D. Attribute Subset Selection

Attribute Subset Selection is the process to reduce the dimensionality of data by using only a subset of all its features. It involves the removal of irrelevant and redundant features.

Domain knowledge was used to remove features that were irrelevant, after which 53 attributes remained. Some features were removed because they were redundant, having high correlation with other features—correlation threshold was set to 0.7. After this step, 34 attributes remained in the dataset.

E. Unsupervised Discretisation

Discretisation is the process of transferring continuous functions, models, variables, and equations into discrete counterparts. The dataset was discretised into 'low' , 'medium' and 'high' categories using Equal-Frequency Discretisation. This was done to facilitate the use of Association Rule Mining as one of the data analysis techniques.

F. Normalisation

Normalisation or standardisation refers to the process of re-scaling the values of the variables in a data set so that they share a common scale. Due to the sensitivity of clustering algorithms to the scale of features, Z-score normalisation was performed on the data.

$$z_i = \frac{x_i - \bar{x}}{s} \tag{1}$$

V. Data Visualisation

All the major visualisation techniques (histogram, boxplot, pie-chart, scatter plot, etc.) were used to discover relationships between gender, literacy, household facilities, women and child health, etc. We obtained some interesting avenues to explore through these visualisations. It made apparent how the interplay of factors is different in urban and rural households.

We discovered a few interesting relationships like that of household sanitation facilities and clean cooking fuel use, literacy/education rates, BMI of mother, breastfeeding practise with stunting. Awareness about the use of contraceptives was found to be correlated with increased awareness of child immunisation, antenatal care and breast and cervix cancer in women. (Fig 1)(Fig 2)

VI. Methods of Data Analysis (Stunting)

In the initial stages, few interesting relationships were discovered (shown in the scatter plots below) like that of household sanitation facilities, literacy/education rates, BMI of mother, breastfeeding practise with stunting. Awareness about the use of contraceptives was found to be correlated with increased awareness of child immunisation, antenatal care and breast and cervix cancer in women.

To determine risk factors for stunting, Clustering Analysis, Association Rule Mining and Regression Analysis techniques were employed.

A. Regression Analysis

19 variables relevant to stunting (on the basis of correlation analysis and literature review) were selected. To determine risk factors for stunting, first, a full regression model with all 19 variables was conducted and the p-value was observed for each attribute. Then, a new regression model with all variables with p-values less than 0.05 was developed. The remaining variables were discarded. This was performed recursively until all final variables were significant (p-value < 0.05). The stat models library in Python was used to develop the regression model and find p-values.

B. Cluster Analysis

To explore how they affected stunting, attributes (dimensions) were individually picked, and *K-Means Clustering* algorithm was implemented and run along the picked dimension(s) to cluster the data points. The Z-Score normalised data was fed to the clustering algorithm.

The data points (districts) were grouped into three clusters for each of the following attributes:

- Percentage of households with improved sanitation facilities.
- Percentage of households that use clean fuel for cooking.
- Percentage of households with access to improved water sources.
- Percentage of women with below normal BMI.
- Percentage of children under age 3 years breastfed within one hour of birth.
- Percentage of mothers who consumed iron folic acid for 100 days or more when they were pregnant.
- Percentage of women who are literate.
- Percentage of women who are anaemic.

According to the value of their centroids (prototypes), each cluster was labelled 'high', 'medium', or 'low', the cluster with the largest value of the centroid being labelled 'high'. The Stunting values for each district, after discretisation, had already been categorised into 'high', 'medium', and 'low'. *So the ratio(%) of data points (districts) belonging to each stunting category for each cluster was then calculated.*

C. Association Rule Mining

Apriori algorithm was implemented and run to mine association rules related to Stunting. The discretised data, with the values of each attribute having been categorised into 'high', 'medium', and 'low', was available. In order to convert it into a suitable input format for the association rule mining algorithm, categorical attributes were converted into assymetric binary attributes. *The support and the confidence thresholds were 0.15 and 0.6, respectively.*

VII. Methods of Data Analysis (Contraception Awareness)

A. Clustering

The districts were clustered on the basis of the attribute, "Any Modern Method". The attribute indicated what percent of households in a district used a modern method of contraception.

After the clusters were obtained, in each cluster, the ratio of data points falling in the 'high', 'medium', and 'low' categories for the following attributes was calculated:

- Children aged 12-23 months who were fully immunized(%).
- Children aged 9-59 months who received a vitamin A dose in last 6 months(%).
- Mothers who had at least 4 antenatal care visits(%).
- Institutional births(%).
- Women who went for Cervix cancer checkups(%).
- Women who went for Breast cancer checkups(%).

B. Association Rule Mining

Apriori algorithm was implemented to mine association rules about how *contraception awareness* predicts other factors related to health of women and children. The thresholds for support and confidence were 0.15 and 0.50 respectively.

The code for the methods discussed above can be found in the referenced link. [1]

VIII. Results (Stunting)

A. Regression

A multivariate regression analysis of 19 attributes was conducted and p-values were studied for each attribute. 6 attributes were removed from the regression model as they had p-values greater than the decided threshold (0.05). A new regression model was developed for the remaining attributes. All of these 13 attributes turned out to be significant in the new model (p-value < 0.05), hence they were retained. The coefficients of the regression model were studied to determine their effect on Stunting. The following attributes were found to be positively related with increased child stunting:

- Households with an improved drinking-water source1 (%)
- Women whose Body Mass Index (BMI) is below normal (BMI <18.5 kg/m2) (%)
- All women age 15-49 years who are anaemic (%)

The following attributes were found to be negatively related with increased child stunting:

- Sex ratio of the total population (females per 1000 males)
- Households using improved sanitation facility (%)
- Households using clean fuel for cooking (%)
- Women who are literate (%)
- Total children age 6-23 months receiving an adequate diet (%)

However, it was suspected that these results could be biased due to high degree of multi-collinearity in the independent variables,which is a common problem in multivariate regression. Hence, the effect of these variables was further studied using clustering analysis and association rule mining.

B. Clustering

1) Clustering on the basis of "Households using improved sanitation facility" yielded the following cluster centres, $C_0 = -1.0056$ (low sanitation), $C_1 = 1.28156$ (high sanitation), $C_2 = 0.12155$ (medium sanitation). In cluster C_0 (corresponding to a low level of sanitation), 62.7% districts have 'high' stunting, and 7.5% have 'low' stunting. On the other hand, in cluster C_1 (corresponding to a high level of sanitation), 6.2% have 'high' stunting, and 65.3% districts have 'low' stunting. (Fig 3)
2) Clustering on the basis of "Children under age 3 years breastfed within one hour of birth" yielded the following cluster centres, $C_0 = 0.071931$ (medium), $C_1 = -1.017726$ (low), $C_2 = 1.343879$(high). In cluster C_1 (low breastfeeding), 47% have districts 'high' stunting and 19% have 'low' stunting. On the other hand, in

cluster C_2 23% have 'high' stunting and 44% have 'low' stunting.

3) Clustering on the basis of "Women having below normal BMI (BMI <18.5 kg/m2)" yielded the following cluster centres, $C_0 = 1.13491$ (high), $C_1 = -1.14934$ (low), $C_2 = 0.0131351$ (medium). In cluster C_0 (high percentage of women having below normal BMI), 64% districts have 'high' stunting rates, and 3.8% have 'low' stunting rates. On the other hand, in cluster C_1 (low percentage of women having below normal BMI), 5.7% districts have 'high' stunting, and 67.8% have 'low' stunting.
4) Clustering on the basis of "Mothers who consumed iron folic acid for 100 days or more when they were pregnant" yielded the follow cluster centres, $C_0 = -0.87373$ (low), $C_1 = 1.53358$ (high), $C_2 = 0.35245$ (medium). In cluster C_1 (high percentage of mothers who consumed folic acid), 8.6% districts have 'high' stunting, and 68.8% have 'low' stunting. On the other hand, in cluster C_0 (low percentage of mothers who consumed folic acid), 49.9% districts have 'high' stunting, and 16% have low stunting.

C. Association Rule Mining

The association rules were obtained in the format: (antecedent)—>(consequent) (confidence). The following rules contained *High Stunting* as the consequent.

1) (Households using improved sanitation facility (%) (low))—>(Children under 5 years who are stunted (%) (high)) (0.684454)
2) (Women who are literate (%) (low))—>(Children under 5 years who are stunted (%) (high)) (0.64733)
3) (Women whose Body Mass Index (BMI) is below normal (BMI <18.5 kg/m2) (%) (high))—>(Children under 5 years who are stunted (%) (high)) (0.6325)
4) (Women who are literate (%) (low), Women whose Body Mass Index (BMI) is below normal (BMI <18.5 kg/m2) (%) (high)')—>(Children under 5 years who are stunted11 (%) (high)) (0.76288)
5) (Households using improved sanitation facility (%) (low), Women who are literate (%) (low)) —> (Children under 5 years who are stunted (%) (high)) (0.77815)
6) (Households using clean fuel for cooking (%) (low), Women who are literate (%) (low)) —> (Children under 5 years who are stunted (%) (high)) (0.74339)
7) (Households using improved sanitation facility (%) (low), Men who are literate (%) (low)) —> (Children under 5 years who are stunted (%) (high)) (0.76264)
8) (Households using improved sanitation facility (%) (low), Women age 20-24 years married before age 18 years (%) (high)) —> (Children under 5 years who are stunted (%) (high)) (0.78629)
9) (Women who are literate (%) (low), Men who are literate (%) (low)) —> (Children under 5 years who are stunted (%) (high)) (0.70662)
10) (Women who are literate (%) (low), Women age 20-24 years married before age 18 years (%) (high)) —> (Children under 5 years who are stunted (%) (high)) (0.6877)
11) (Women who are literate (%) (low), Mothers who consumed iron folic acid for 100 days or more when they were pregnant (%) (low)) —> (Children under 5 years who are stunted (%) (high)) (0.7238)
12) (Households using improved sanitation facility2 (%) (low), Mothers who consumed iron folic acid for 100 days or more when they were pregnant (%) (low)) —> (Children under 5 years who are stunted (%) (high)) (0.8132)
13) (Households using improved sanitation facility (%) (low), Women whose Body Mass Index (BMI) is below normal (BMI <18.5 kg/m2) (%) (high)) —> (Children under 5 years who are stunted (%) (high)) (0.7448)
14) (Households using improved sanitation facility (%) (low), Households using clean fuel for cooking(%) (low)) —> (Children under 5 years who are stunted (%) (high)) (0.73178)

IX. Results (Contraception Awareness)

A. Clustering

The data was clustered on the basis of the attribute, "Any Modern Method of Contraception Used(%)". The cluster centres obtained were the following, $C_0 = -1.21763$ (low usage of modern contraception), $C_1 = 1.133296$ (high usage of modern contraception), $C_2 = 0.038585$ (medium usage of modern contraception). The following results were obtained:

1) In cluster C_0, 51.2% districts have 'low' child immunisation rates, and 15.8% have 'high' immunisation rates. On the other hand, in cluster C_1, 14% have 'low' immunisation rates, and 52.4% have 'high' immunisation rates. (Fig 4)
2) In cluster C_0, 50.5% have 'low' % of children who received vitamin A dose, and 14.8 % have 'high' percentage for the same. On the other hand, in cluster C_1, 52.4% have 'high' % of children who received vitamin A dose and 14% have 'low' % of children who received vitamin A dose.
3) In cluster C_0, 67.4% have 'low' % of women who had at least 4 antenatal care visits, and only 11 % have 'high' percentage of the same. On the other hand, in cluster C_1, 54.7% have 'high' % of women who had at least 4 antenatal care visits and only 10% have 'low' % of women who had at least 4 antenatal care visits.
4) In cluster C_0, 62.3% have 'low' percentage of institutional births, and only 9.4% have 'high' percentage of institutional births. On the other hand, in cluster C_1, 54.9% have 'high' percentage of institutional births and 10.9% have 'low' percentage of institutional births.
5) In cluster C_0, 47.6% and 51.2% of districts have a 'low' % of women went for breast cancer and cervix cancer checkups respectively. On the other hand, in cluster C_1, 55.2% and 57.4% of districts have a 'high' % of women went for breast cancer and cervix cancer checkups respectively.

B. Association Rule Mining

The association rules were obtained in the format: (antecedent)—>(consequent) (confidence). The following rules contained *High use of contraception* as the antecedent.

1) (Any modern method of contraception used (%) (high)) —> (Checkup for Breast Cancer (%) (high)) (0.5452)
2) (Any modern method of contraception used (%) (high)) —> (Mothers who consumed iron folic acid for 100 days or more when they were pregnant (%) (high)) (0.5150)
3) (Any modern method of contraception used (%) (high)) —> (Institutional births (%) (high)) (0.5475)
4) (Any modern method of contraception used (%) (high)) —> (Mothers who had at least 4 antenatal care visits (%) (high)) (0.5475)
5) (Any modern method of contraception used (%) (high)) —> (Children age 9-59 months who received a vitamin A dose in last 6 months (%) (high)) (0.5220)
6) (Any modern method of contraception used (%) (high)) —> (Women with 10 or more years of schooling (%) (high)) (0.5000)
7) (Any modern method of contraception used (%) (high)) —> (Checkup for Cervix Cancer (%) (high)) (0.5638)

X. Conclusion

A. Stunting

Running K-Means Clustering and the Apriori algorithms on the data, as described above, yielded some interesting results. These results confirmed already existing results in the literature related to stunting, and produced some novel results as well.

Clustering revealed that proper sanitation in households, correct breastfeeding practices, proper nutrition of women (as revealed by BMI), and scrupulous care during pregnancy (as revealed by the Folic Acid tablet usage) are indicative of reduced *stunting rate*.

Association Rule Mining confirmed the above results, but produced some new results as well. Association rule VIII.C.2 reveals the importance of literacy of women in reducing the rate of stunting. On the other hand, literacy in men wasn't shown to be strongly associated with low stunting rate ((*High Men Literacy, Low Stunting*) appeared in our *frequent item set*, but couldn't satisfy the *confidence threshold*). In all likelihood, better education of mothers causes them to make better choices during pregnancy and in the nutrition of their children, leading to proper growth of their children.

B. Contraception Awareness

Clustering and association rule mining revealed that high contraception awareness is associated with whether *children are completely immunised, whether children receive vitamin A dose, whether mothers go for regular antenatal checkups* (which is important for the health of both mothers and children); and furthermore, it was associated with whether women get checkups for Breast and Cervix Cancer.

Contraception usage and awareness indicates if households plan their families. So, in light of the results mentioned, contraception usage and by extension family planning was shown to have serious consequences for the health of children and women.

XI. Plots

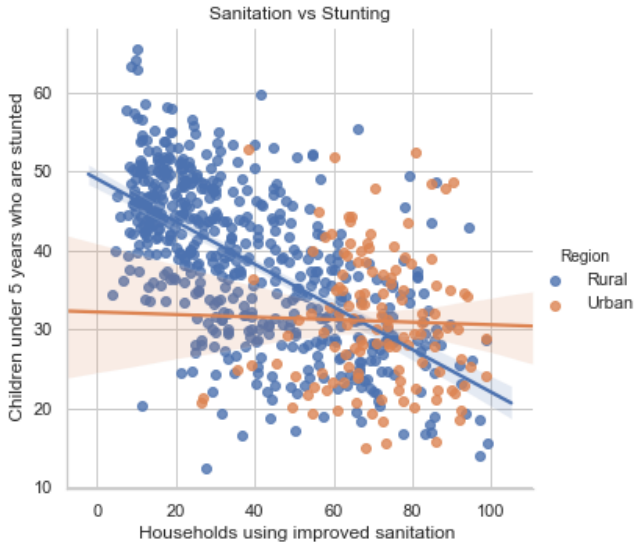

Fig. 1.

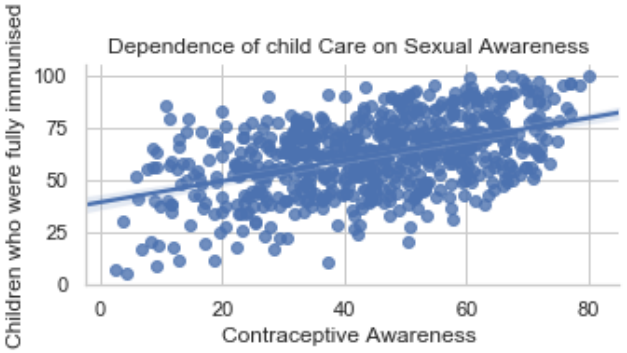

Fig. 2.

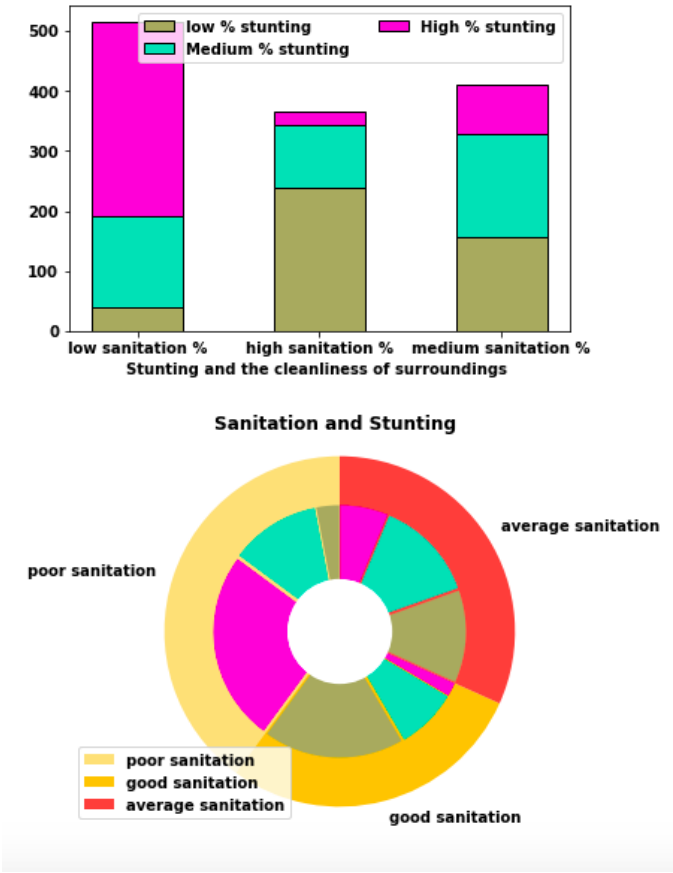

Fig. 3.

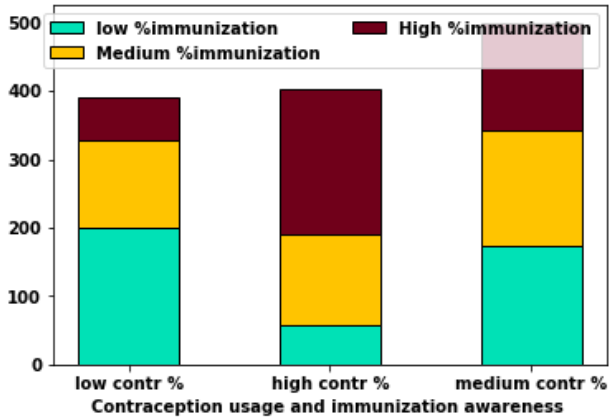

Fig. 4.

References

[1] Prathma Chowksey et al. *Trends in Child Stunting and Contraception Awareness in Women: Analysis of NFHS-4 Analysis*. URL: https://github.com/prathmachowksey/NFHS-4-Analysis/tree/master/Code.

[2] Weigang Liang. *Association of solid fuel use with risk of stunting in children living in China*. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/ina.12627.

[3] *NFHS*. URL: https://data.gov.in/catalog/key-indicators-national-family-health-survey-nfhs.

[4] Alphonse Nshimyiryo. *Risk factors for stunting among children under five years: a cross-sectional population-based study in Rwanda using the 2015 Demographic and Health Survey*. URL: https://bmcpublichealth.biomedcentral.com/articles/10.1186/s12889-019-6504-z.

[5] Ali SS1. *Association of literacy of mothers with malnutrition among children under three years of age in rural area of district Malir, Karachi.* URL: https://www.ncbi.nlm.nih.gov/pubmed/16438277.

[6] UNICEF. *Stop stunting*. URL: unicef.org/india/what-we-do/stop-stunting.