Kamal VK, Srivastav S, Kumari D, Ranjan M. Identification of distinct risk subsets for under five mortality in India using CART model: an evidence from NFHS-4. *Journal of Global Health Reports*. 2020;4:e2020055. doi:10.29392/001c.13169

Research Articles

Identification of distinct risk subsets for under five mortality in India using CART model: an evidence from NFHS-4

Vineet K Kamal[1], Sharad Srivastav[1], Dolly Kumari[2], Mukesh Ranjan[3]

[1] National Institute of Epidemiology (Indian Council Of Medical Research), Chennai, India, [2] Asian Development Research Institute, Patna, India, [3] Pachhunga University College, Mizoram University, Mizoram, India

Keywords: under five mortality, risk subsets, risk subgroups, recursive partitioning, decision tree, cart

https://doi.org/10.29392/001c.13169

Journal of Global Health Reports

Vol. 4, 2020

Background

The objective of this study was to find the distinct risk subsets or clusters identified by the combination of factors and important factors to classify under five mortality (U5M) in high focused Indian states.

Methods

Using population-based cross-sectional data from the National Family Health Survey (NFHS, 2015-2016) on 1, 40, 427 live births of five years preceding the survey occurred to 99,205 women of high focused Indian states with U5M rate above the national level, a recursive partitioning approach based two classification tree models, one without considering missing values and other with missing together approach, were fitted using binary outcome of U5M and independent factors comprising of socioeconomic, demographic, maternal and biological, nutritional and environmental factors.

Results

There were nine and sixteen sub-groups in model-1 and model-2, respectively. In model-1, breastfeeding = no & birth in past 5 years = (2, 3+ births) and in model-2, breastfeeding = no & birth weight = (<2.5kg, not known) & birth in past 5 years = (2, 3 or more births) were found to be maximum mortality risk sub-groups. In terms of variable importance to predict U5M, model-1 identified birth in past 5 years, breastfeeding, birth order, wealth index, mother‘s age at birth. Model-2 additionally identified delivery complications, birth weight, state, sanitation facility, birth interval, caste, education. Overall correct classification rate was higher for model-1 (66%) than model-2 (64%).

Conclusions

The main observed risk cluster was combination of two factors like breastfeeding and number of births in past 5 years, which for most people are easily modifiable with appropriate strategies and policies. Finally, to combat U5M in high focused states, identifying risk subsets or clusters is important for targeting and intervening purposes, as the intensity and type of policies and programs may differ according to clusters. This method is suitable to identify complex natural interactions between predictors, important variables and hypothesis generation to inform policy maker on intervention strategies, which may be difficult or impossible to uncover using traditional multivariable techniques.

Despite a significant reduction in under-five mortality (U5M) globally over the years, it is still a major public health problem in developing countries. India has a very significant role to play in global efforts to end the avoidable death of children under the age of five, as it contributes the highest share of deaths among the under-fives globally.[1] The global U5M rate in the year 2016 was 41 deaths per 1000 live births, with 56% reduction since 1990.[2] The global community addresses the critical need to end preventable child deaths, making it an essential part of the global strategy for women's, children's and adolescent health (2016-2030) and the third Sustainable Development Goal (SDG) to ensure healthy lives and promote well-being for all people of all ages. The SDG 3.2 objective calls for an end to preventable deaths of newborns and children under the age of five and specifies that all countries should aim at reducing neonatal mortality to at least 12 deaths per 1,000 live births and under-five deaths to at least 25 deaths per 1,000 live births by 2030.[3]

In India, the U5M rate (U5MR) is unevenly distributed across regions and socio-economic groups and the national-level estimate of U5MR masks the huge sub-national geographic variation across Indian states. There are seven states, designated as high focused states, where U5MR is higher than national average (50 per 1,000 live births). The U5MR in these high focused states, namely Uttar Pradesh (UP) (78 per 1,000 live births), Madhya Pradesh (MP) (65 per 1,000 live births), Chhattisgarh (64 per 1,000 live births), Bihar (58 per 1,000 live births), Assam (56 per 1000 live births), Jharkhand (54 per 1,000 live births), Rajasthan (51 per 1000 live births) are almost similar to the U5MR in some of the poor performing African countries.[4]

The hypotheses framing in medical research and development of public health interventions frequently involve the identification of high-risk groups and the effects of individual and other factors on the concerned outcome.[5,6] Although identifying relevant risk subgroups can be difficult with standard statistical methods, the character of these subgroups can provide insight into effect mechanisms and recommend targets for modified interventions. Identifying factors strongly associated with U5M rates is a topic of increased research interest for most developing countries, including India, to formulate appropriate health programs and policies to consider when searching for options to combat child mortality in highly focused states with a view to achieving a higher level of socio-economic development. Many statistical methods have been used so far, including logistic regression, survival analysis, etc., to identify factors that are strongly associated with U5M rates.[7–11] However, many of these studies did not consider important factors such as birth in the past 5 years (birth spacing), complications during delivery, the number of antenatal visits in their research that could have a confounding effect, and the combination (joint effects) of these multiple risk factors that could affect U5M in highly focused states. Additionally, it has been observed that individual, household, and community level factors may have synergistic effects on outcome, and interaction terms in regression models are typically used to test hypotheses regarding synergistic effects (i.e., effect modification). However, for modelling more complicated nonlinear associations, this approach is not perfect.[5] Classification and Regression tree (CART) analysis can be used as an alternative to identify important variables and complex interactions between predictors which may be difficult or impossible to uncover using traditional regression or other multivariate techniques. CART analysis, a means of multivariate data exploration, is a non-parametric technique which uses binary recursive partitioning algorithm to assign the subjects into mutually exclusive homogeneous subsets or clusters according to a set of independent variables and this will produce a classification tree following a series of binary splits dividing children into higher- and lower-risk subgroups for a given outcome based on given predictor variables.[12] CART based recursive partitioning techniques have been used in past to identify high-risk subgroups, mostly to examine clinical or genetic risk factors, for cardiovascular disease, diabetes and population-based studies.[13–15] It would be very important to make a classification rule based on combination of factors with the help of decision tree considering local effect of each factor on outcome to predict U5M. In this regard, none of the studies from India or other settings have been carried out to find the distinct risk subsets or cluster based on decision tree (recursive partitioning). This information can then be used to suggest areas for policymakers to provide a better understanding of the socio-economic, demographic, cultural and environmental factors affecting U5M, keeping in view to achieve SDG 3.2 in high priority states of India, as these states are unlikely to achieve the SDG 3 target by 2030.[1] Therefore, our aim was to find the distinct risk subsets or clusters (identified by the combination of factors) and important factors to classify U5M in high focused states using CART model.

METHODS

DATA AND STUDY POPULATIONS

We used data from the Kids Recode (KR) file of the recently conducted fourth round of the Demographic and Health Survey (DHS) for India, popularly known as the National Family Health Survey-4 (NFHS-4, 2015-16) which is a large-scale, multi-round survey conducted in a representative sample of households throughout India. NFHS-4, based on 1,315,617 children born of 699,686 women in 601,509 households with a response rate of 98%, was conducted under the stewardship of the Ministry of Health and Family Welfare, Government of India (GOI) which provides information on population, health and nutrition for India and each state/union territory.

The survey incorporated 425,563 households from rural areas and 175,946 households from urban areas. The sample size for NFHS-4 was decided based on the need to produce indicators at district and state/union territory (UT). The sample was selected through a two-stage sample design: for the first stage, the Primary Sampling Units (PSU) were villages in rural areas (selected with probability proportional to size) and Census Enumeration Blocks (CEB) for urban areas and in second stage; a random selection of 22 households in each PSU and each CEB were done for rural and urban area, respectively.[16] KR Recode file contains the full birth history of five years preceding the survey of all women interviewed, including information on pregnancy and postnatal care as well as immunization, health and nutrition. The study population considered were high focused states which constitute seven states (UP, MP, Chhattisgarh, Bihar, Assam, Jharkhand, and Rajasthan) having under-five and infant mortality higher than national level (50 deaths per 1,000 live births). There were a total of 2, 59,627 births born between 2010 and 2016, but for the analysis, we considered a total of 1, 40,427 births who were born between 2010 and 2016 in the above high focused seven states.

VARIABLES CONSIDERED

In this cross-sectional study, the binary dependent (outcome) variable was considered as U5M (i.e. death before reaching the fifth birthday). Births which took place preceding five years from the date of survey have been considered for the analysis.

We selected predictors for U5M based on review of literatures and additionally, other important factors seemed otherwise necessary from policy making point of view. There were 24 independent/explanatory variables available at household-level (States, Type of place of residence, Wealth index, Type of cooking fuel, Type of toilet facility, Source of drinking water, Religion, Caste), maternal-level (Highest educational level, Mothers age at birth, Covered by health insurance, Breastfeeding, Birth in past five years), child-level (Preceding Birth Interval, Birth order number, Birth weight in kilograms, Sex, Child is twin), and maternal and child-care program-level (Number of antenatal visits during pregnancy, Delivery by caesarean section, Assistance at deliver, Delivery complications, Place of delivery, Time before postnatal check-up). The definition of these variables, based on,[17] are given in Appendix S1 in the Online Supplementary Document For analysis, these variables were further classified as Demographic factors, Socioeconomic factors, Nutritional factor, Environmental factors, and Maternal and Biological factors as given in Table 1a and 1b. These explanatory variables were recoded and categorized, given in **Table 1a and 1b**, using the literature and NFHS-4 Report.[16]

ANALYTIC APPROACH

Descriptive statistics were used to summarize overall sample characteristics. The variables considered for model fitting are reported with their missing values, if any, in **Table 1a and 1b**. The association between explanatory or predictor variables and outcome was seen using Chi-squared test. The p-values less than 0.05 were taken as statistically significant.

For identification of distinct risk subsets of U5M, we fitted two classification tree models, where model-1 was fitted without considering missing values and model-2 was fitted using missing together (MT) approach.[18] The purpose of these classification trees were to reveal the structure of the dataset with respect to distinct combinations of risk variables that jointly influence the child mortality risk. For the binary outcome variable U5M, software used mainly three steps: (1) constructing an initial large classification tree using recursive partitioning to choose the predictor variables:(2) pruning this tree upward, thereby creating a nested sequence of smaller trees: and (3) selecting an optimum-sized tree from this nested sequence. For deriving classification tree, the data were randomly splitted into two parts, the training set (50%) and testing set (50%). The tree was grown using only the training set, and the testing set was used to estimate the error of all possible subtrees that can be built, and the subtree with the lowest error on the testing set was chosen as the classification tree. The steps of tree building procedures has been given in more detail in Appendix S6 in the Online Supplementary Document.

For details about CART method, readers may consult Breiman et al.[19] The performances of both the CART models were summarized using sensitivity, specificity, area under curve (ROC), correct classification in learning and testing sample data, and overall percentage correct classification. IBM SPSS statistics 23 was used for data preparation and descriptive analysis and Minitab's Salford Predictive Modeler (SPM) 8.3.0 software was used for model fitting. For the analysis, appropriate sampling weight was used. The sampling weights were already specified in the NFHS datasets.

RESULTS

CLASSIFICATION TREE MODELS

Based on two classification trees (model-1 and model-2), we constructed distinct risk subsets, represented in Table 2 and Table 3, respectively. Details about resultant classification tress and other technical details are given in Appendix S2, Appendix S3, Appendix S4, Appendix S5, Appendix S6 and Appendix S7 in the Online Supplementary Document. Terminal subsets comprised of more than 5.6% mortality cases were considered as mortality groups for classification purpose, as the percentage of under-five mortality in the total sample was 5.6%. In the classification tree model-1, there were a total of nine terminal nodes, in which terminal subsets 3, 4, 7, 8 and 9 were considered as the mortality risk subsets. Similarly, in the classification tree model-2, out of 16 terminal nodes, terminal subsets 3, 4, 5, 8, 9, 10, 13, 14, 15 and 16 were considered as the mortality risk subsets.

From Table 2, as expected, the child, who is on breastfed and single child born in last five year tend to have classified as survival (terminal node 1). Similarly, if a child is not on breastfed, and his/her mother is having one birth in last five years and she belongs to poor or middle class and the birth order of the child is first then that child will be classified in mortality subset (terminal node 8). Similar interpretation can be made for other terminal nodes also. Terminal node 4 comprised of combination of factors "Breastfeeding= (Yes)& Birth in past 5 years = (2,3, or more births) & Birth order = (1)" had mortality of 8.3%, which was least of all 5 mortality subsets and terminal node 9 comprised of the combination of factors "Breastfeeding= (No) & Birth in past 5 years = (2, 3, or more births)" had mortality of 16.5%, which was maximum of all 5 subsets.

For model-2, interpretation of Table 3 for classification rule is same as in case of above model-1. In Table 3, the terminal node 8 comprised of combination of factors "Breastfeeding = (No) & Birth weight= (2.5 kg or more) & Birth in past 5 years = (1) & Wealth Index = (Poor, Middle) & Birth order= (>=2) & Mothers age = (30-39 years, >=40 years)" had mortality of 6.1%, which was least of all 10 mortality subsets and terminal node 16 comprised of the combination of factors "Breastfeeding = (No) & Birth weight= (Less than 2.5 kg, Not known) & Birth in past 5 years = (2, 3 or more births)" had mortality of 21.2%, which was maximum of all 10 subsets.

Table 4 shows relative variable importance with normalized score for both the classification tree models, which ranks the variables, as a summary of a variable's contribution to currently selected the overall tree when all nodes are examined and taking into account how good a splitter it is. In our example, the variable 'Birth in past 5 years' and 'Delivery complication' were ranked as most important followed by 'Breastfeeding' in model-1, and model-2, respectively. Explanatory variables missing in the Table 4 received a zero score, indicating that these variables contributed no role in the analysis, either as a primary splitter or as a surrogate. Variable importance has been discussed in more detail in Appendix S8 in the Online Supplementary Document.

Table 5 shows the performances of both the CART models in learning and testing sample data. In model-1 and model-2, the area under curve was 0.789, and 0.827, respec-

Table 1a. Distribution of deaths by socioeconomic, demographic and environmental factors.

Socioeconomic factors	Under-five deaths (yes)	Total	*P-value* *
	n (row %)	N (column %)	
Highest educational level			<0.001
Illiterate	3753(6.5)	57932(41.3)	
Primary	1348(6.1)	22046(15.7)	
Secondary and above	2797(4.6)	60449(43.0)	
Type of place of residence			<0.001
Urban	1214(4.5)	26724(19.0)	
Rural	6684(5.9)	113703(81.0)	
Covered by health insurance			0.243
No	7181(5.6)	127154(90.5)	
Yes	717(5.4)	13273(9.5)	
Wealth index			<0.001
Poor	5515(6.3)	88061(62.7)	
Average	1203(5.3)	22567(16.1)	
Rich	1180(4)	29799(21.2)	
States			<0.001
Assam	518(5)	10309(7.3)	
Bihar	1322(5.2)	25437(18.1)	
Chhattisgarh	539(5.8)	9283(6.6)	
Jharkhand	570(4.7)	12204(8.7)	
Madhya Pradesh	1361(5.5)	24611(17.5)	
Rajasthan	758(4.5)	16832(12.0)	
Uttar Pradesh	2830(6.8)	41751(29.7)	
Demographic factors			
Mothers age at birth(years)			<0.001
<20	1312(7.5)	17546(12.5)	
20-29	5184(5.1)	101380(72.2)	
30-39	1246(6.3)	19931(14.2)	
>40	156(9.9)	1570(1.1)	
Religion			0.044
Hindu	6481(5.7)	114195(81.3)	
Muslim	1278(5.5)	23264(16.6)	
Other	139(4.7)	2968(2.1)	
Caste			<0.001
Scheduled caste	1760(6.4)	27511(19.6)	
Scheduled tribe	1168(5.8)	20095(14.3)	
Other Backward Classes	3758(5.5)	68271(48.6)	
Other	1065(5)	21430(15.3)	
Missing	147(4.7)	3120(2.2)	
Environmental factors			
Type of cooking fuel			<0.001
Safe	1166(4.3)	26835(19.1)	
Unsafe	6238(6)	104728(74.6)	
Not a de jure resident	494(5.6)	8864(6.3)	
Type of toilet facility			<0.001

Improved	1694(4.4)	38139(27.2)	
Unimproved	5710(6.1)	93424(66.5)	
Not a de jure resident	494(5.6)	8864(6.3)	
Source of drinking water			0.238
Improved	6662(5.7)	117610(83.8)	
Unimproved	742(5.3)	13953(9.9)	
Not a de jure resident	494(5.6)	8864(6.3)	
Total	**7898(5.6)**	**140427(100)**	

* Chi-square (χ^2) test

tively for the testing samples. The misclassification rates for training and testing samples for both the CART models were 5.9 % and 5.6 %, respectively. The overall correct classification, i.e., the percent of correctly classified cases out of all cases in the dataset, for model-1 and model-2 were about 66% and 64%, respectively.

DISCUSSION

By applying CART model based recursive partitioning technique to NFHS-4 data, we identified the distinct risk subsets (identified by the combination of factors) and important factors to classify U5M in high focused states. This method may be appropriate for identifying situations where the etiologic role of one factor depends on the presence or absence of one or more other factor. Also, this method can be used to identify important variables and complex interactions between predictors which may be difficult or impossible to uncover using traditional multivariate techniques. The findings of the present study are found to be consistent with the literature with additional information which has not been revealed by any of other studies, particularly in India. In addition to finding out the important factors for U5M, the classification tree models were able to depict the combination or natural interaction of the maternal and biological factors within themselves or with other environmental as well as socioeconomic and demographic factors. In other words, we cannot say that a single factor alone matters, actually it is a combination of factors which matter in terms of prediction of U5M. For instance, in model-1, Breastfeeding = (No) & Birth in past 5 years = (2, 3+ births) and in model-2, Breastfeeding= (No) & Birth weight= (<2.5kg, Not known) & Birth in past 5 years = (2, 3 or more births) were found to be the natural interaction determined by the model itself gives the maximum mortality risk subset.

One of the rationales for this study is to find modifiable factors that can help the country in reaching the SDG 3 by 2030. A study[1] used the NFHS-4 and reported that 177 districts of India are unlikely to achieve the SDG 3 target on the U5M rate by 2030 where the majority of high risk districts are located in high focused states. In order to achieve the SDG 3, where U5M rate is taken as one of the milestones to be reduced to 25 deaths per 1000 live births, these seven high focused states (Assam, Bihar, Chhattisgarh, Jharkhand, MP, Rajasthan, and UP) have prevalence of 66.5% U5M. Effective policies and programs can be put in place to enhance accelerated reduction of U5M in these high focused states if we have information based on child level, household/mother's level, community level, and child-care program-level factors and the way and manner they influence the child mortality. In such situation, a classification rule based on combination of factors with the help of decision tree based on recursive partitioning algorithm considering local effect of each factor on outcome can be very useful to predict under five mortality. The identification of distinct risk subgroups by recursive partitioning of data could lead to the aiming of specific groups for primary or secondary intervention to avoid U5M. A cluster or subset having "no breastfeeding and more number of children in last five years" (Model-1's terminal node 9) or "no breastfeeding, low birth weight of a baby and more number of children in last five years" (Model-2's terminal node 16) put together highest contribution in U5M in high focused states of India. These all are modifiable factors which can be improved with appropriate policy implementation and intervention. Breastfeeding is one of the most important interventions of child survival, and it is reflected in present study also, despite this, only 65% of children are exclusively breastfed for the first six months and 45% of children receive breastfeeding within one hour of birth in India.[20] According to NFHS-4's report, timely initiation of breastfeeding is particularly less for women with no schooling, for home deliveries, and for births delivered by unskilled personnel. The percentage of children with early initiation of breastfeeding (i.e., breastfed within one hour of birth) is very low in UP (25%), one of high focused states with highest percentage of U5M (78 deaths per 1,000 live births). Initial breastfeeding and exclusive breastfeeding for the first six months of life avoids around 20% deaths in newborn and 13% deaths in under-five,[21] and can also decrease mortality due to neonatal infections (sepsis, pneumonia, diarrhea and tetanus).[22] The number of births in last 5 years is related to family planning. The use of contraceptive methods could act as guard against a greater number of children and help in adequate birth spacing and could influence against undesirable pregnancies which in turn will cut the risk associated with early weaning of the child and health complications such as maternal depletion syndrome. Also, number of births in last 5 years and increase in total children ever born could consequence in low birthweight, lack of care, and premature births on the limited household resources, and children have to compete for the small resources available for their survival.

Table 1b. Distribution of deaths by nutritional, maternal and biological factors.

Nutritional factor	Under-five deaths	Total	*P-value* *
	n (row %)	N (column %)	
Breastfeeding			<0.001
No	4832(10.7)	44999(32.0)	
Yes	3066(3.2)	95428(68.0)	
Maternal and Biological factors			
Preceding Birth Interval(months)			<0.001
<=24	5164(6.7)	77565(55.2)	
>24	2734(4.3)	62862(44.8)	
Birth in past five years			<0.001
1	2102(3.3)	63254(45.0)	
2	3701(6)	62036(44.2)	
3+	2095(13.8)	15137(10.8)	
Birth order number			<0.001
1	2813(6)	47237(33.6)	
>=2	5085(5.5)	93190(66.4)	
Birth weight in kilograms			<0.001
Less than 2.5 kg	1295(7.4)	17556(12.5)	
2.5 kg or more	2439(3.2)	77270(55.0)	
Missing	4164(9.1)	45601(32.5)	
Sex of child			<0.001
Male	4277(5.8)	73394(52.3)	
Female	3621(5.4)	67033(47.7)	
Child is twin			<0.001
Single	7278(5.3)	138053(98.3)	
Multiple	620(26.1)	2374(1.7)	
Number of antenatal visits during pregnancy			<0.001
No antenatal visits	1134(5.1)	22259(15.9)	
More than 4visits	1679(3.7)	44897(32.0)	
At least 4 visits	888(2.8)	31402(22.4)	
Missing	4197(10)	41869(29.8)	
Delivery by caesarean section			<0.001
No	7313(5.7)	128068(91.2)	
Yes	585(4.7)	12359(8.8)	
Assistance at delivery			<0.001
Unskilled	2521(7)	35975(25.6)	
Skilled	5329(5.1)	104452(74.4)	
Missing	48(100)	48(0.0)	
Delivery complications			<0.001
No	1736(3.4)	51681(36.8)	
Yes	1958(4.1)	47477(33.8)	
Missing	4204(10.2)	41269(29.4)	
Place of delivery			<0.001
Institutional	5099(5.1)	100111(71.3)	
Non institutional	2751(6.8)	40316(28.7)	
Missing	48(100)	48(0.0)	

Time before postnatal check up			<0.001
<4 hours	611(3.1)	19862(14.1)	
4-23 hours	21(1.7)	1241(0.9)	
1-2 days	85(2.4)	3521(2.5)	
3+ days	124(1.7)	7438(5.3)	
No check-up	2877(4.3)	66881(47.6)	
Missing	4180(10.1)	41484(29.5)	
Total	**7898(5.6)**	**140427(100)**	

* Chi-square (χ^2) test

Table 2. Classification tree model-1.

Terminal node	Risk subsets or classification rules	Class (%)
1	Breastfeeding= (Yes)& Birth in past 5 years = (1 birth)	Survival(100.0)
2	Breastfeeding= (Yes)& Birth in past 5 years = (2,3, or more births) & Birth order = (>=2) & Birth in past 5 years = (2 births)	Survival(97.8)
3	Breastfeeding= (Yes)& Birth in past 5 years = (2,3, or more births) & Birth order = (>=2) & Birth in past 5 years = (3 or more births)	Mortality(9.3)
4	Breastfeeding= (Yes)& Birth in past 5 years = (2,3, or more births) & Birth order = (1)	Mortality(8.3)
5	Breastfeeding= (No)& Birth in past 5 years = (1 birth) & Wealth Index = (Rich)	Survival(96.7)
6	Breastfeeding= (No)& Birth in past 5 years = (1 birth) & Wealth Index = (Poor, Middle) & Birth order = (>= 2) & Mothers age = (<20 years, 20-29 years)	Survival(95.1)
7	Breastfeeding= (No)& Birth in past 5 years = (1 birth) & Wealth Index = (Poor, Middle) & Birth order = (>= 2) & Mothers age = (30-39 years, >=40 years)	Mortality(9.7)
8	Breastfeeding= (No)& Birth in past 5 years = (1 birth) & Wealth Index = (Poor, Middle) & Birth order = (1)	Mortality(15.5)
9	Breastfeeding= (No)& Birth in past 5 years = (2, 3, or more births)	Mortality(16.5)

% indicates percentage of class cases in terminal nodes

Methodologically, CART is quite different from the more commonly used statistical methods like logistic regression, survival analysis and other conventional multivariate methods, with the primary benefit of illustrating the natural interaction and important variable selection related to outcome. The other benefits of CART are that recursive partitioning does not make any distributional assumptions about the modeled variables, and among variables, it accounts for multilevel interactions. Also, nonlinear relationships between parameters do not affect tree performance. The outcome from CART analysis is easy to interpret and explain to policy makers.

To the best of our knowledge, this study is the first attempt of its kind to examine combination of factors in hierarchical manner associated with U5M accounting for demographic, socioeconomic, health, and environmental variables in high focused states of India. Several studies have established the relationship that exists between the socioeconomic, demographic factors and U5M in India and other developing countries,[7,8,23,24] on the other hand researchers in the public health and medical sciences had carried out researches to explain U5M by examining the effects of some proximate determinants.[9,11,25] In the present study it is found that breastfeeding, birth interval, birth order, type of birth, wealth index and mother's age at birth are the factors that mostly influence U5M. Other studies from Indian subcontinent also have reported similar factors.[11,26] Apart from Indian setting, studies[25] done in other countries also showed that birth spacing is a very significant proximate variable through which socioeconomic variables such as mother's education, age, place of residence, and wealth status influence child mortality in Nigeria. Similar to this study, factors like birth in past 5 years and preceding birth interval have come out as the most important risk factors in present study. Present study also shows that unimproved sanitation facilities could influence U5M like other similar study from India.[27] A study from South Africa used Random survival forests demonstrated that covariates that were originally excluded from the survival analysis due to violation of the PH assumption were important in explaining under-five child mortality rates.[28] In terms of outcomes other than child mortality, a study from India used this method to show the negative association of breastfeeding practice, economic status, and antenatal care of mother with malnutrition among tribal children.[29]

Though this study offers unique findings and policy sug-

Table 3. Classification tree model-2

Terminal node	Risk subsets or classification rules	Class (%)*
1	Breastfeeding= (Yes)& Delivery Complication=(No, Yes)	Survival(100.0)
2	Breastfeeding= (Yes)& Delivery Complication= (Not known) & Birth weight= (2.5kg or more) & State = (Bihar, Chhattisgarh, Jharkhand, MP, Rajasthan) & Birth in past 5 years = (1,2 births)	Survival(96.0)
3	Breastfeeding= (Yes)& Delivery Complication= (Not known) & Birth weight= (2.5kg or more) & State = (Bihar, Chhattisgarh, Jharkhand, MP, Rajasthan) & Birth in past 5 years = (3 or more births)	Mortality(7.9)
4	Breastfeeding= (Yes)& Delivery Complication= (Not known) & Birth weight= (2.5kg or more) & State = (Assam, UP)	Mortality(6.3)
5	Breastfeeding= (Yes)& Delivery Complication= (Not known) & Birth weight= (Less than 2.5 kg, Not known)	Mortality(12.5)
6	Breastfeeding= (No)& Birth weight= (2.5kg or more) & Birth in past 5 years = (1 birth) & Wealth Index = (Rich)	Survival(97.7)
7	Breastfeeding= (No)& Birth weight= (2.5kg or more) & Birth in past 5 years = (1 birth) & Wealth Index = (Poor, Middle) & Birth order= (>=2) & Mothers age = (<20 years,20-29 years)	Survival(96.8)
8	Breastfeeding= (No)& Birth weight= (2.5kg or more) & Birth in past 5 years = (1) & Wealth Index = (Poor, Middle) & Birth order= (>=2) & Mothers age = (30-39 years,>=40 years)	Mortality(6.1)
9	Breastfeeding= (No)& Birth weight= (2.5kg or more) & Birth in past 5 years = (1) & Wealth Index = (Poor, Middle) & Birth order= (1)	Mortality(10.3)
10	Breastfeeding= (No)& Birth weight= (2.5kg or more) & Birth in past 5 years = (2, 3, or more births)	Mortality(10.7)
11	Breastfeeding= (No)& Birth weight= (Less than 2.5 kg, Not known) & Birth in past 5 years = (1 birth) & Birth Order = (>=2) & Mothers age = (<20 years,20-29 years) & Sanitation facility = (Improved)	Survival(95.5)
12	Breastfeeding= (No)& Birth weight= (Less than 2.5 kg, Not known) & Birth in past 5 years = (1 birth) & Birth Order = (>=2) & Mothers age = (<20 years,20-29 years) & Sanitation facility = (Unimproved) & Birth Interval = (<=24 months)	Survival(94.6)
13	Breastfeeding= (No)& Birth weight= (Less than 2.5 kg, Not known) & Birth in past 5 years = (1 birth) & Birth Order = (>=2) & Mothers age = (<20 years,20-29 years) & Sanitation facility = (Unimproved) & Birth Interval = (>24 months)	Mortality(6.4)
14	Breastfeeding= (No)& Birth weight= (Less than 2.5 kg, Not known) & Birth in past 5 years = (1 birth) &Birth Order = (>=2) & Mothers age = (30-39 years, >=40 years)	Mortality(11.7)
15	Breastfeeding= (No)& Birth weight= (Less than 2.5 kg, Not known) & Birth in past 5 years = (1 birth) & Birth Order = (1)	Mortality(18.2)
16	Breastfeeding= (No)& Birth weight= (Less than 2.5 kg, Not known) & Birth in past 5 years = (2, 3 or more births)	Mortality(21.2)

% indicates percentage of class cases in terminal nodes

gestions, some methodological issues need to be taken care in consideration. We built two models, one (model-1) without considering missing values and other (model-2) treating missing values as a level of that variable. We built model-2 because of availability of missing observation in substantial manner and this is the limitations of secondary data. Mainly, this is because the information on antenatal-care interventions were collected only for most recent births or last child. This restriction may introduce considerable bias in the results. In model-2, delivery complication was found to be most important, but when this variable formed risk subset or cluster in combination with other variables, the interpretation was not much meaningful, for example: Breastfeeding = (Yes) & Delivery Complication= (Not known) & Birth weight= (Less than 2.5 kg, Not known) was classified as mortality group. Similar interpretation might hold true for other clusters of model-2 also. Although the performance in terms of discriminative ability i.e., area under curve was greater, both in learn and test sample, for model-2 than model-1, overall correct classification rate was higher for model-1 as compared to model-2. Therefore, in terms of interpretation and correct classification, model-1 seems more meaningful and could be recommended for policy suggestion.

LIMITATIONS

There could be potential data quality problem due to recall bias, as the quality of mortality estimates calculated from birth histories depends on the mother's ability to recall all of the children she has given birth to, and their birth dates and ages at death as well. Underestimation of childhood mortality might have taken place due to the selective omission from the birth histories of those births that did not sur-

Table 4. Variable importance for classification tree models.

#	Model-1*		Model-2**	
	Variables Overall Normalized Scores (%)			
1	Birth in past 5 years	100	Delivery complication	100
2	Breastfeeding	70.87	Breastfeeding	44.74
3	Birth order	27.21	Birth weight	14.74
4	Wealth index	7.49	Birth in past 5 years	7.11
5	Mothers age at birth	3.92	Birth order	5.31
6	-	-	Mothers age at birth	2.75
7	-	-	Wealth index	2.72
8	-	-	State	1.32
9	-	-	Sanitation facility	0.51
10	-	-	Birth interval	0.3
11	-	-	Caste	0.3
12	-	-	Education	0.19

*Classification tree model without using missing together approach

** Classification tree model using missing together approach

Table 5. Model error measures

	Model-1*		Model-2*	
Classification tree models	Learn	Test	Learn	Test
N (1,40,427)	70,207	70,220	70,207	70,220
Average Log Likelihood	-0.19	-0.185	-0.177	-0.172
ROC (Area Under Curve)	0.789	0.782	0.83	0.827
Misclass Rate Overall	0.059	0.056	0.059	0.056
Class. Accuracy	0.661	0.66	0.646	0.643
Specificity	72.60%	72.59%	62.92%	62.71%
Sensitivity	70.25%	67.45%	91.57%	90.36%
Overall % Correct Classification	66.03%		64.26%	

*Classification tree model without using missing together approach

** Classification tree model using missing together approach

vive. The shift of birth dates, which may distort mortality trends. This can happen if an interviewer knowingly records a birth as occurring in a different year than the one in which it occurred. This may occur if an interviewer is trying to expurgate her/his overall work load, because live births occurring during the five years before the interview are the subject of a lengthy set of extra questions. Due to recall bias, the quality of reporting of age at death may suffer. Age pattern of mortality could be distorted due to misreporting the child's age at death. Some parts of birth weight records were based on the mother's recall and should be interpreted with caution.

Other limitations related with information of ANC visit (available in 70.2% cases), postnatal check-up (available in 70.5% cases), which were available for recent births or youngest child only. For children other than youngest, we treated observations as missing, although these were not collected. Information for household variables (Source of drinking water, Type of toilet facility, Type of cooking fuel) contained 6.3% responses as “not a jure resident”.

In the face of minor changes in the training sample, the standard Classification trees are often unstable; and sometimes, findings may be difficult to reproduce and should be interpreted with caution.

CONCLUSIONS

The main partitioning variables in our results were breastfeeding and number of births in past 5 years, which for most people are easily modifiable with appropriate strategies and policies. Other variables like birth order, birth weight, wealth index, and mother's age at birth were found to be important predictors of U5M. To combat U5M, identifying risk subsets or clusters is important for targeting and intervening purposes, as the intensity and type of policies and programs may differ according to clusters. In situation where data is available at multilevel, this method can be used to identify homogeneous subsets or clusters defined by combinations of individual characteristics, complex natural interactions between predictors, selection of important variables, hypothesis generation, and data exploration to inform policy maker on intervention strategies, which may be difficult to uncover using traditional multivariate techniques.

Acknowledgements: We thank the staffs of the Division of Epidemiology & Biostatistics, ICMR-National Institute of Epidemiology, Chennai, India for their support for this research. No ethics statement is required for this work as this study is based on publicly available dataset with no identifiable information on the survey participants.

Data Availability: All related data are within the paper and its supporting information files. Data are available at https://dhsprogram.com/data/available-datasets.cfm.

Funding: No funding was available for this study

Authorship contributions: VKK contributed to conception and design of the study, analysis and interpretation of data, and was involved in supervision, drafting and editing the manuscript. SS contributed to conception and study design, analysis of the data, drafting and editing the manuscript. DK contributed to conception, interpretation of the data and editing the manuscript. MR contributed to interpretation of the data and editing the manuscript. All authors approved the final manuscript for publication.

Competing interests: The authors completed the Unified Competing Interest form at www.icmje.org/coi_disclosure.pdf,and declare no conflicts of interest.

Correspondence to:
Vineet Kumar Kamal, Ph.D.
Scientist ‘C’,
Division of Epidemiology & Biostatistics,
National Institute of Epidemiology, Indian Council of Medical Research (ICMR),
R-127, 3rd Avenue, Tamil Nadu Housing Board,
Ayapakkam, Chennai-600077, India.
vineetstats@gmail.com

REFERENCES

1. Bora JK, Saikia N. Neonatal and under-five mortality rate in Indian districts with reference to Sustainable Development Goal 3: An analysis of the National Family Health Survey of India (NFHS), 2015-2016. Moise IK, ed. *PLoS ONE*. 2018;13(7):e0201125. doi:10.1371/journal.pone.0201125

2. Levels and Trends in Child Mortality Report 2017 | UNICEF Publications | UNICEF [Internet]. https://www.unicef.org/publications/index_101071.html.

3. Goal 3: Sustainable Development Knowledge Platform [Internet. https://sustainabledevelopment.un.org/sdg3. Accessed January 9, 2020.

4. Kumar C, Singh PK, Rai RK. Under-five mortality in high focus states in India: A district level geospatial analysis. Baradaran HR, ed. *PLoS ONE*. 2012;7(5):e37515. doi:10.1371/journal.pone.0037515

5. Van Hulst A, Roy-Gagnon M-H, Gauvin L, Kestens Y, Henderson M, Barnett TA. Identifying risk profiles for childhood obesity using recursive partitioning based on individual, familial, and neighborhood environment factors. *Int J Behav Nutr Phys Act*. 2015;12(1):17. doi:10.1186/s12966-015-0175-7

6. Garzotto M, Beer TM, Hudson RG, et al. Improved detection of prostate cancer using classification and regression tree analysis. *Journal of Clinical Oncology*. 2005;23(19):4322-4329. doi:10.1200/jco.2005.11.136

7. Mondal MNI, Hossain MK, Ali MK. Factors Influencing Infant and Child Mortality: A Case Study of Rajshahi District, Bangladesh. *Journal of Human Ecology*. 2009;26(1):31-39. doi:10.1080/09709274.2009.11906162

8. Chowdhury QH, Islam R, Hossain K. Socio-economic determinants of neonatal, post neonatal, infant and child mortality. :8.

9. Buwembo P. *Factors Associated with Under-5 Mortality in South Africa: Trends 1997-2002*. (Doctoral dissertation, University of Pretoria)

10. Mohanty SK. Multidimensional poverty and child survival in India. Bhutta ZA, ed. *PLoS ONE*. 2011;6(10):e26857. doi:10.1371/journal.pone.0026857

11. Mani K, Dwivedi SN, Pandey RM. Determinants of Under-Five Mortality in Rural Empowered Action Group States in India: An Application of Cox Frailty Model. *Int J MCH AIDS*. 2012;1(1):60-72. doi:10.21106/ijma.9

12. Lewis RJ. An introduction to classification and regression tree (CART) analysis. In: *Annual Meeting of the Society for Academic Emergency Medicine in San Francisco, California*. Vol 14. ; 2000.

13. Gomez F, Wu YY, Auais M, Vafaei A, Zunzunegui M-V. A Simple Algorithm to Predict Falls in Primary Care Patients Aged 65 to 74 Years: The International Mobility in Aging Study. *Journal of the American Medical Directors Association*. 2017;18(9):774-779. doi:10.1016/j.jamda.2017.03.021

14. Costello T, Swartz M, Sabripour M, Gu X, Sharma R, Etzel C. Use of tree-based models to identify subgroups and increase power to detect linkage to cardiovascular disease traits. *BMC Genet*. 2003;4(Suppl 1):S66. doi:10.1186/1471-2156-4-s1-s66

15. Stern SE, Williams K, Ferrannini E, DeFronzo RA, Bogardus C, Stern MP. Identification of individuals with insulin resistance using routine clinical measurements. *Diabetes*. 2005;54(2):333-339. doi:10.2337/diabetes.54.2.333

16. IIPS/India II for PS-, ICF. India National Family Health Survey NFHS-4 2015-16. 2017. https://dhsprogram.com/publications/publication-fr339-dhs-final-reports.cfm. Accessed April 16, 2019.

17. DHS Recode Manual (English) [Internet]. https://www.dhsprogram.com/publications/publication-dhsg4-dhs-questionnaires-and-manuals.cfm. Accessed January 15, 2020.

18. Zhang H, Holford T, Bracken MB. A tree - based method of analysis for prospective studies. *Statistics in medicine*. 1996;15(1):37-49. doi:10.1002/(sici)1097-0258(19960115)15:1

19. Breiman L, ed. *Classification and Regression Trees. Repr*. Boca Raton: Chapman & Hall [u.a.]; 1998.

20. Home | Ministry of Women & Child Development | GoI [Internet]. https://wcd.nic.in/. Accessed January 10, 2020.

21. Jones G, Steketee RW, Black RE, Bhutta ZA, Morris SS, , Bellagio Child Survival Study Group. How many child deaths can we prevent this year? *The Lancet*. 2003;362(9377):65-71. doi:10.1016/s0140-6736(03)13811-1

22. Mullany LC, Katz J, Li YM, et al. Breast-Feeding Patterns, Time to Initiation, and Mortality Risk among Newborns in Southern Nepal. *The Journal of Nutrition*. 2008;138(3):599-603. doi:10.1093/jn/138.3.599

23. Agarwal S, Srivastava A. Social determinants of children's health in urban areas in India. *J Health Care Poor Underserved*. 2009;20(4 Suppl):68-89. doi:10.1353/hpu.0.0232

24. Worku Z. Factors That Affect Under-Five Mortality among South African Children: Analysis of the South African Demographic and Health Survey Data Set. 2009.

25. Oni G, Samuel G. Effect of Birth Spacing on Under-five Mortality in Nigeria: A Proximate Determinant Approach (Birth Spacing and Under-five Mortality). 2016.

26. Sahu D, Nair S, Singh L, Gulati BK, Pandey A. Levels, trends & predictors of infant & child mortality among Scheduled Tribes in rural India. *INDIAN J MED RES*. 2015;11.

27. Dwivedi LK, Banerjee K, Jain N, Ranjan M, Dixit P. Child health and unhealthy sanitary practices in India: Evidence from Recent Round of National Family Health Survey-IV. *SSM - Population Health*. 2019;7:100313. doi:10.1016/j.ssmph.2018.10.013

28. Nasejje JB, Mwambi H. Application of random survival forests in understanding the determinants of under-five child mortality in Uganda in the presence of covariates that satisfy the proportional and non-proportional hazards assumption. *BMC Res Notes*. 2017;10(1). doi:10.1186/s13104-017-2775-6

29. Debnath A, Bhattacharjee N. Factors Associated with Malnutrition among Tribal Children in India: A Non-Parametric Approach. *Journal of Tropical Pediatrics*. 2014;60(3):211-215. doi:10.1093/tropej/fmt106

SUPPLEMENTARY MATERIALS

Online Supplementary Document

Download: https://www.joghr.org/article/13169-identification-of-distinct-risk-subsets-for-under-five-mortality-in-india-using-cart-model-an-evidence-from-nfhs-4/attachment/36059.pdf