

Machine Learning-Based Prediction of Acute Morbidity and Counterfactual Policy Simulation among Children 6–23 Months of Age in Punjab, Pakistan: A Secondary Analysis of a Multi-Indicator Cluster Survey 2017–18

Bakhtawar Majeed

University of Veterinary and Animal Sciences

Furqan Awan

University of Veterinary and Animal Sciences

Muniba Khaliq

`muniba.khaliq@pgr.reading.ac.uk`

University of Reading

Muhammad Hassan Mushtaq

University of Veterinary and Animal Sciences

Muhammad Asif Ali

University of Veterinary and Animal Sciences

Research Article

Keywords: Random forest, Machine learning, Childhood morbidity, Counterfactual simulation, Punjab, Pakistan, MICS, Vitamin supplementation

Posted Date: September 4th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10413711/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Background

Acute morbidity from fever, diarrhea, and respiratory infections remains the leading cause of death among children under five years of age in low- and middle-income countries. Existing epidemiological analyses did not quantify the potential population-level impact of targeted interventions. This study aimed to identify the most important socioeconomic and nutritional determinants of acute childhood morbidity and to simulate the effects of predicted factors through counterfactual analysis.

Methods

This study used secondary cross-sectional data from the Punjab Multi-Indicator Cluster Survey (MICS) 2017–18). The analytical sample comprised 10,128 children aged 6–23 months. The outcome variable was acute disease status (fever, diarrhea, or respiratory symptoms in the preceding two weeks), classified as diseased or healthy. Nine policy-relevant predictor variables were included: vitamin and mineral supplement access, wealth quintile index, maternal education, bottle feeding, height-for-age, weight-for-height, minimum dietary diversity, breastfeeding status, and health insurance. A random forest model was applied after hyperparameter optimization via grid search. Counterfactual simulation was conducted by systematically replacing each key predictor with its optimal value across the full dataset to estimate the predicted change in health rates under different scenarios.

Results

Of the 10,128 children included, 49% ($n = 4,982$) were classified as diseased. The tuned random forest model achieved an overall accuracy of 60.5% (95% CI: 58.5–62.2%), significantly exceeding the no-information rate of 50.7% ($p < 0.001$), with an AUC of 0.632 (95% CI: 0.612–0.651) and a kappa coefficient of 0.204. Vitamin and mineral supplement access emerged as the dominant predictor of mean decrease in Gini importance, followed by the wealth quintile index and maternal education. Counterfactual simulation demonstrated that optimizing access to vitamin and mineral supplements for all children was associated with the largest predicted improvement in population-level health rates (+ 12.79 percentage points), followed by achieving higher maternal education (+ 3.61 percentage points).

Conclusions

This study demonstrates the utility of random forest modeling combined with counterfactual simulation as a policy-analytical tool for assessing child health in data-rich but resource-constrained settings. Vitamin and mineral supplementation programmes, maternal education, and dietary diversity

interventions represent the highest-yield policy priorities for reducing acute childhood morbidity in Punjab's most vulnerable communities.

1. Introduction

Morbidity and mortality during childhood are major health challenges worldwide, with enormous consequences for the well-being of individuals and the development of society (1). Although there has been significant progress in decreasing mortality rates among children under 5 years of age in the global context since 1990, with 93 deaths per every 1,000 live births to 37 deaths per 1,000 live births in 2023, a gap between various geographical areas and socioeconomic classes still remains (2). The major causes of death in children under the age of 5, especially in low- and middle-income countries, are infectious diseases, preterm birth complications, and congenital anomalies (3).

Malnutrition also plays a crucial role in exacerbating childhood morbidity and mortality, contributing to approximately 50% of deaths in children under five years of age (4, 5). Approximately 22% of children under 5 years of age worldwide were stunted, and 6.8% suffered from wasting in 2022 (6). In addition to malnutrition, multiple factors contribute to increased childhood morbidity and mortality, including socioeconomic status, maternal education, maternal age, birth intervals, breastfeeding practices, and cultural practices (7). A low Wealth Quintile Index (WQI) has a profound effect on a child's nutritional status and well-being. Approximately twenty-five percent of children from middle-income and low-income countries who are younger than 2 years lack threshold levels of food diversity according to the World Health Organization (WHO) recommendations (8, 9).

In Pakistan, the situation is particularly concerning, with an underfive mortality rate of 58.5 deaths per 1,000 live births and the highest infant mortality rate among South Asian countries and 22nd in the whole world (10–12). The United Nations has set the Sustainable Development Goal 3.2 to reduce the five mortality rates to 25 deaths per 1000 live births by 2030, which demands multisectoral interventions to address this issue (13). Childhood morbidity in Pakistan includes diverse health problems, such as infectious diseases, developmental disorders (14, 15), and most importantly, malnutrition (16). Approximately 40% of children under 5 years of age in Pakistan were stunted, and 17.7% of children experienced wasting in 2018 (17).

The first 1,000 days of life, spanning from conception to a child's second birthday, represent a critical window for improving child health and survival. During this period, rapid fetal growth, brain development, immune system maturation, and organ formation occur, making children highly responsive to nutritional and environmental influences (18). Adequate maternal nutrition, exclusive breastfeeding for the first six months, timely complementary feeding, immunization, infection prevention, and access to quality healthcare during this window substantially reduce the risk of childhood morbidity from infectious diseases such as diarrhea and pneumonia, as well as mortality (19). Therefore, evidence consistently identifies the first 1,000 days as the most effective period for interventions aimed at reducing childhood morbidity and mortality while promoting optimal growth, neurodevelopment, and long-term health

outcomes (20). The previous 1000 days provide a foundation for the next 1,000 days of life (2–5 years of age), which constitute a second critical window for early childhood development, during which nurturing care and supportive environments are essential for sustaining and enhancing the health, growth, and developmental gains achieved during the first 1,000 days. However, most children in low- and middle-income countries do not receive adequate nurturing care, placing them at increased risk of poor developmental outcomes (21).

Punjab, the most populous province of Pakistan with over 110 million residents, contributes disproportionately to national figures given its size. (22, 23). Socioeconomic disparities are evident, with 45% of child morbidity cases occurring in poor families compared with 35% in wealthier households (12). It is critical to address malnutrition and childhood morbidity in Punjab from both a national and regional standpoint as a part of larger initiatives to improve child health and sustainable development. Furthermore, the Punjab government has made enhancing children's health a top priority in its health strategy, highlighting the necessity of evidence-based interventions customized for the region (24).

The Punjab Multi-Indicator Cluster Survey (MICS) 2017–18, conducted by UNICEF in collaboration with the Bureau of Statistics, Government of Punjab, provides the most comprehensive population-representative dataset on child nutrition, health, and socioeconomic status in the province (23). Previous analyses of MICS data have highlighted alarming gaps in healthcare utilization, dietary diversity, and vaccination coverage (25, 26).

Machine learning (ML) algorithms, particularly ensemble tree-based methods such as random forest (RF), offer significant advantages for population health research. The random forest handles high-dimensional datasets, captures complex nonlinear interactions between predictors, is robust to multicollinearity, and generates variable importance metrics that can directly inform priority-setting (27). Several studies have demonstrated the superiority of random forest over logistic regression for identifying key predictors of child health outcomes in low–middle-income countries (LMICs) (28, 29). In addition to prediction, a growing body of literature has applied counterfactual simulations, systematically replacing observed predictor values with hypothetical optimal values and observing model-predicted outcomes to estimate the population-level impact of policy interventions (30). This approach moves ML from a purely predictive framework to a policy-generative framework, offering decision-makers quantified estimates of intervention effects within specific populations.

Despite these methodological advances, no published study to date has applied random forest modeling with integrated counterfactual policy simulation to Punjab MICS data for childhood acute morbidity, leaving the complex interplay of nutritional, dietary, and socioeconomic determinants underexplored and the potential effect of policy interventions unquantified. This gap represents a missed opportunity to translate routinely collected survey data into actionable public health intelligence.

This study therefore has three objectives: (i) to describe the distribution of nutritional and socioeconomic determinants of acute morbidity among children aged 6–23 months in Punjab; (ii) to build and evaluate a random forest model for predicting acute childhood morbidity via policy-relevant

predictors; and (iii) to apply counterfactual simulations to estimate the impact of key population-level policy-relevant predictors on the health status of children. The novel contribution of this study lies in the integration of machine learning-based prediction with counterfactual policy simulation using a nationally representative dataset, providing a replicable analytical framework for child health policy in Punjab and similar LMIC settings.

2. Methods

2.1 Study Design and Data Source

This study used a cross-sectional secondary data analysis design. Data were obtained from the Punjab Multi-Indicator Cluster Survey (MICS) 2017–18, a household-based, population-representative survey conducted by the Bureau of Statistics, Government of Punjab, in collaboration with UNICEF (31). The MICS uses a stratified multistage cluster sampling design and has been conducted periodically since 2003 to provide data on key health, education, and welfare indicators for children and women. The survey employed structured, interviewer-administered questionnaires at the household level. The MICS 2017–18 represents the most recent comprehensive data on a variety of indicators for children up to 23 months of age in Punjab, with data collected from all nine administrative divisions: Bahawalpur, Dera Ghazi Khan (D.G. Khan), Faisalabad, Gujranwala, Lahore, Multan, Rawalpindi, Sahiwal, and Sargodha.

2.2 Study population and analytical samples

The original MICS 2017–18 dataset contains records for 42,408 children under five years of age. The target population for this research was children aged 6–23 months of age, as the WHO Minimum Dietary Diversity (MDD) indicator — a core predictor variable in this study — is exclusively validated and applicable to children aged 6–23 months, as defined by the WHO/UNICEF Infant and Young Child Feeding (IYCF) guidelines (32). After the initial filter, a sample size of 11,364 children aged 6–23 months was obtained. A stepwise data cleaning protocol was applied prior to analysis. First, records with standard MICS missing or “don’t know” response codes (values 8 and 9 for categorical variables) were excluded. Second, implausible z scores for the anthropometric indicators (coded as 99.97, 99.98, or 99.99 in the dataset, denoting measurement failure) were excluded. Third, records with missing values for any of the nine analytic variables were excluded via complete-case analysis. After applying these exclusion criteria, the final analytical sample comprised 10,128 children. The geographic distribution of the analytical sample across Punjab’s nine divisions is reported in Table 1.

2.3 Outcome Variable

The primary outcome variable was acute disease status, which was defined as a dichotomous indicator (yes/no). A child was classified as 'diseased' if the caregiver reported the occurrence of any of the following within the two weeks preceding the survey interview: fever, diarrhea, cough, or difficulty breathing. A child was classified as “healthy” if none of these symptoms were reported. This composite case definition follows the standard MICS childhood morbidity ascertainment methodology and is

consistent with the WHO Integrated Management of Childhood Illness (IMCI) framework, which identifies fever, diarrhea, and cough or difficult breathing as the four primary presenting symptoms of childhood illness (33). This definition has been applied in previous population-based analyses of under-five morbidity using DHS data from comparable settings (34).

2.4 Predictor Variables

Nine policy-relevant predictor variables were selected a priori on the basis of theoretical frameworks of child health determinants, data availability, and evidence from the literature (8, 9). These were:

(1) Vitamin and mineral supplement access: Whether the child had access to vitamin or mineral supplements (yes/no). (2) Wealth quintile index: Household economic status is classified into five quintiles (Poorest, Poor, Middle, Rich, Richest) on the basis of an asset index computed by MICS. (3) Maternal education: The highest educational level attained by the child's mother (none, primary, middle, secondary, or higher). (4) Bottle feeding: the use of a feeding bottle in the preceding 24 hours (yes/no). (5) Height-for-age z score (HAZ) category: Nutritional status classified as severely stunted (< -3 SD), moderately stunted (-3 to < -2 SD), or normal (≥ 2 SD). (6) Weight-for-height z score (WHZ) category: Nutritional status classified as severe wasting (< -3 SD), moderate wasting (-3 to < -2 SD), normal, or overweight ($> +2$ SD). (7) Minimum dietary diversity (MDD): Whether the child achieved the WHO MDD score of ≥ 5 food groups (adequate/inadequate). The dietary groups considered were breastmilk, grains/cereals, legumes, vitamin A-rich fruits and vegetables, other fruits and vegetables, dairy products, flesh foods, and eggs (32). (8) Breastfeeding status—whether the child was still being breastfed at the time of the survey (yes/no). (9) Health insurance—whether the household had any form of health insurance coverage (yes/no).

2.5 Statistical analysis—random forest model

All the analyses were conducted in R version 4.3.1. The RF package was used for model fitting. Prior to modeling, the dataset was split into a training set (70%) and a testing set (30%) via stratified random sampling (set.seed = 1234) to ensure reproducibility and preserve the outcome class distribution across sets.

RF is an ensemble learning method that constructs multiple decision trees from bootstrap samples of the training data, with a random subset of predictors considered at each node split. The final class prediction is determined by the majority vote across all the trees. This procedure reduces the variance associated with individual decision trees, improves generalization to unseen data, and handles complex nonlinear interactions between predictors without requiring pre-specification. The algorithm is also inherently robust to multicollinearity, making it well suited to the interrelated predictor set used in this study (35).

Hyperparameter optimization was conducted via a systematic grid search over the training set, varying three key parameters: the number of variables randomly sampled at each node split (mtry: 2, 3, 4, 5, and 6), the minimum node size (nodesize: 1, 5, and 10), and the maximum number of terminal nodes

(maxnodes: 10, 20, 30, and 40). Model accuracy on the testing set was used as the selection criterion (36). The optimal parameter combination was mtry = 6, nodesize = 10, and maxnodes = 40, with ntree = 1,000 trees used to ensure convergence of the error estimates. The model was fitted with importance=TRUE to enable computation of both the mean decrease Gini coefficient and the mean decrease accuracy importance metric.

2.6 Model evaluation

Model performance on the held-out testing set was evaluated in terms of (i) overall classification accuracy with 95% confidence interval; (ii) sensitivity (true positive rate for the healthy/no class); (iii) specificity (true positive rate for the diseased/yes class); (iv) positive predictive value (PPV); (v) negative predictive value (NPV); (vi) F1 score; (vii) balanced accuracy; (viii) Cohen's kappa coefficient; and (ix) area under the receiver operating characteristic curve (AUC-ROC) with 95% CI computed via the DeLong method via the pROC package. The statistical significance of model accuracy over the no-information rate (NIR) was also assessed (37). Variable importance was assessed via the mean decrease Gini (MDG), which quantifies the total decrease in node impurity averaged over all trees and splits for each predictor. Higher MDG scores indicate greater contributions to class separation (38).

2.7 Counterfactual simulation

To translate the model findings into policy-relevant estimates, the top three predictor variables assessed by the mean decrease Gini coefficient were used for counterfactual simulation. Three counterfactual scenarios were simulated. In each scenario, a single predictor variable was set to its optimal value for all observations in the full dataset, whereas all other variables were held at their observed values. The fitted random forest model was then applied to generate the predicted probabilities of being "healthy" [Disease Status = No] for each observation. The mean predicted health rate across all observations was compared against the baseline predicted health rate (derived from the original observed dataset) to quantify the predicted change attributable to each intervention.

Three scenarios were evaluated: (1) vitamin and mineral supplement access — all children set to vitamin mineral supplements = "yes"; (2) maternal education — all mothers set to mothers' education = "higher"; and (3) minimum dietary diversity — all children set to the MDD = "adequate". This approach mirrors the marginal standardization method used in causal inference and has been applied in comparable machine learning policy simulation studies (39). The results are reported as the absolute change in the predicted percentage of healthy children relative to the baseline.

2.8 Geographic Mapping of Acute Morbidity Prevalence

To visualize the spatial distribution of the acute morbidity burden across Punjab, the division-level two-week period prevalence was computed as the proportion of children classified as diseased within each of Punjab's nine administrative divisions. Administrative boundary shape files for Pakistan at the district level (Level 2) were obtained from the United Nations Office for the Coordination of Humanitarian Affairs (OCHA) Pakistan Common Operational Dataset. District-level polygons were aggregated to the division

level via a district-to-division lookup table constructed according to the official administrative classification of Punjab Province. Spatial data processing and choropleth map generation were performed via the *sf* and *ggplot2* packages in R version 4.3.1. Prevalence estimates were mapped via a continuous yellow-to-red color gradient, with darker shading indicating higher acute illness prevalence.

2.9 Ethical considerations

The MICS 2017–18 was conducted under institutional ethical clearance by the Bureau of Statistics, Government of Punjab, and UNICEF Pakistan. Informed consent was obtained from all the participating households at the time of data collection. This study constitutes a secondary analysis of deidentified, publicly available data; no additional ethical approval was needed.

3. Results

3.1 Sample characteristics

The analytical sample comprised 10,128 children aged 6–23 months, of whom 5,173 (51%) were male and 4,955 (49%) were female. The geographic distribution was broadly proportional to divisional population size, with Gujranwala (16%), Faisalabad (12%), Multan (12%), and Lahore (12%) contributing the largest shares of the sample. All demographic and nutritional characteristics are presented in Tables 1 and 2.

Among the 10,128 children, 4,982 (49%) were classified as diseased, reporting at least one episode of fever, diarrhea, cough, or respiratory difficulty in the two weeks preceding the survey, whereas 5,146 (51%) were classified as healthy. Figure 1 presents the geographic distribution of the two-week period prevalence of acute illness among children aged 6–23 months across the nine administrative divisions of Punjab. Substantial divisional variation was observed, with D.G. Khan recording the highest acute illness prevalence (66%), followed by Sargodha (56.8%), Rawalpindi (52.3%), Multan (54.3%), and Faisalabad (49.2%). The lowest prevalence was observed in Lahore (41%) and Gujranwala (38%).

Nutritional status indicators revealed that 29% of the children were stunted (11% severely stunted, 18% moderately stunted) and that 8.9% were wasted (2.3% severely wasted, 6.6% moderately wasted). The vast majority of children (81%) did not achieve the WHO Minimum Dietary Diversity score. Only 15% of the children had access to vitamin or mineral supplements, whereas only 2.4% of the households reported health insurance coverage. Breastfeeding was reported in 63% of the children, whereas bottle feeding was practiced in 59% of the households.

Socioeconomic characteristics revealed a predominantly poor population: 23% of children belonged to the poorest wealth quintile, and 21% belonged to the poor quintile, with only 15% in the richest quintile. With respect to maternal education, 41% of mothers had no formal education, whereas only 14% had attained higher education. These findings indicate that a sample is concentrated at the lower end of the socioeconomic spectrum.

Table 1
Sociodemographic characteristics of the study sample (N = 10,128)

Characteristic	Category	N	%
Gender	Male	5,173	51%
	Female	4,955	49%
Age (months)	6–11 months	3,156	31%
	12–13 months	6,972	69%
Division	Bahawalpur	887	8.8%
	Dera Ghazi Khan	1,122	11%
	Faisalabad	1,196	12%
	Gujranwala	1,643	16%
	Lahore	1,251	12%
	Multan	1,227	12%
	Rawalpindi	1,076	11%
	Sahiwal	676	6.7%
	Sargodha	1,050	10%
Wealth Quintile Index	Poorest	2,341	23%
	Poor	2,145	21%
	Middle	2,173	21%
	Rich	1,922	19%
	Richest	1,547	15%
Maternal Education	None	4,189	41%
	Primary	2,065	20%
	Middle	1,051	10%
	Secondary	1,425	14%
	Higher	1,398	14%
Health Insurance	No	9,880	97.6%
	Yes	248	2.4%

Table 2
Nutritional and health-related characteristics of the study sample (N = 10,128)

Characteristic	Category	N	%
Disease Status	Healthy (No)	5,146	51%
	Diseased (Yes)	4,982	49%
Height-for-Age (HAZ)	Severely Stunted	1,102	11%
	Moderately Stunted	1,807	18%
	Normal	7,219	71%
Weight-for-Height (WHZ)	Severe Wasting	232	2.3%
	Moderate Wasting	664	6.6%
	Normal	9,044	89%
	Overweight	188	1.9%
Minimum Dietary Diversity	Inadequate	8,245	81%
	Adequate	1,883	19%
Breastfeeding	No	3,751	37%
	Yes	6,377	63%
Vitamin/Mineral Supplements	No	8,569	85%
	Yes	1,559	15%
Bottle Feeding	No	4,108	41%
	Yes	6,020	59%
Vitamin A-rich Fruits/Veg	No	5,848	58%
	Yes	4,280	42%

3.2 Random forest model performance

The tuned random forest model (ntree = 1,000; mtry = 6; nodesize = 10; maxnodes = 40) achieved an overall accuracy of 60.5% (95% CI: 58.7–62.3%) on the held-out test set, statistically significantly exceeding the no-information rate of 50.65% ($p < 2.2 \times 10^{-16}$). The AUC-ROC was 0.6322 (95% CI: 0.6125–0.6518), indicating a modest but statistically significant discriminatory ability above chance. Cohen's kappa was 0.2044, indicating fair agreement beyond chance.

The confusion matrix revealed that the model correctly classified 1,342 healthy children (true negatives for disease) and 508 diseased children (true positives for disease). The confusion matrix for the testing

dataset is given in Table 3. With the healthy class (No) set as the positive reference, the model demonstrated high sensitivity for detecting healthy children (86.64%) but lower specificity for detecting diseased children (33.66%). The positive predictive value was 57.28%, and the negative predictive value was 71.05%. The F1 score was 68.96%, and the balanced accuracy was 60.15%. The full performance metrics are presented in Table 4, and the ROC curve is depicted in Fig. 2.

Table 3
Confusion matrix for the testing dataset (N = 3,058;
30% of the sample data).

Disease Status	No (Healthy)	Yes (Diseased)
No (Healthy)	1,342(TN)	1,001 (FN)
Yes (Diseased)	207 (FP)	508 (TP)

Table 4
Performance metrics of the tuned random forest model on the
held-out test set (N = 4,178)

Metric	Value
Overall Accuracy	0.605 (95% CI: 0.5847–0.6224)
No-Information Rate	0.5065
P value (Accuracy > NIR)	$< 2.2 \times 10^{-16}$
AUC-ROC	0.6322 (95% CI: 0.6125–0.6518)
Cohen's Kappa	0.2044
Sensitivity (Healthy class)	0.8664
Specificity (Disease class)	0.3366
Positive Predictive Value	0.5728
Negative Predictive Value	0.5065
F1-Score	0.6896
Balanced Accuracy	0.6015

Note: Positive class = 'No' (Healthy). Sensitivity refers to the model's ability to correctly identify healthy children; specificity refers to the model's ability to correctly identify diseased children. NIR = no-information rate (accuracy of classifying all cases as the majority class).

3.3 Variable Importance

The variable importance assessed by the mean decrease Gini coefficient is illustrated in Fig. 3. The vitamin and mineral supplement had the highest importance score (MDG $\approx$ 253), substantially exceeding the prespecified importance threshold of 200 and far surpassing all other predictors. The wealth quintile

index (MDG ≈ 36) and maternal education (MDG ≈ 28) ranked second and third, respectively. Weight-for-height (MDG ≈ 18), height-for-age (MDG ≈ 13), and minimum dietary diversity (MDG ≈ 08) had intermediate importance scores. The dominance of supplements in the Gini importance ranking is noteworthy. As a binary variable, 85% of the cases were in the “No” category. Accordingly, counterfactual simulation was employed to provide a more direct, population-level estimate of the policy impact of supplement access.

3.4 Counterfactual Policy Simulation

The baseline predicted health rate across the full analytical dataset under the observed conditions was 71.82%. Table 5 and Fig. 4 present the predicted health rates and effects vs. baseline under the three counterfactual policy scenarios.

Table 5
Counterfactual simulation results: predicted health rates under different policy scenarios

Scenario	Predicted Healthy Rate (%)	Absolute Change vs. Baseline (%)
Baseline (observed conditions)	71.82	—
Vitamin/Mineral Supplement Access (all children)	84.61	+ 12.79
Higher Maternal Education (all mothers)	75.43	+ 3.61
Adequate Minimum Dietary Diversity (all children)	72.29	+ 0.47

Vitamin and mineral supplement access was associated with the largest predicted improvement: a 12.79 percentage point increase in the health rate (from 71.82% to 84.61%). This represents the single highest-impact modifiable intervention among all the scenarios tested. Higher maternal education for all mothers in the sample was associated with a 3.61 percentage point increase in predicted health rates (71.82% to 75.43%). Achieving adequate minimum dietary diversity for all children was associated with a modest 0.47 percentage point increase (71.82% to 72.29%). The distribution of individual-level predicted probabilities of being healthy under each scenario, visualized via density and bar plots (Figs. 4 and 5), confirmed that the supplement scenario produced the most pronounced rightward shift in the predicted probability distribution, indicating the broadest population-level benefit.

4. Discussion

This study applied a random forest machine learning algorithm to data from the Punjab MICS 2017–18 to identify key determinants of acute childhood morbidity and to quantify the predicted population-level impact of targeted policy interventions through counterfactual simulation. The principal findings are as follows: (i) acute morbidity prevalence was 49% in the analytical sample, indicating a high and persistent burden of childhood illness; (ii) vitamin and mineral supplement availability, the wealth quintile index, and

maternal education were the most important predictors of disease status; and (iii) vitamin and mineral supplement access was associated with the largest predicted improvement in health rates (+ 12.79 percentage points), underscoring supplementation as the highest-yield policy priority.

4.1 Disease prevalence

The acute morbidity prevalence of 49% observed in this study, defined as any episode of fever, diarrhea, cough, or respiratory difficulty in the preceding two weeks, reflects the composite burden of these conditions, which individually has been documented at 14% for diarrhea (40), 8.4% for acute respiratory symptoms, and 27% for fever in Punjab (41). The prevalence of acute morbidity observed in this study is also consistent with findings from comparable demographic health survey-based analyses from Pakistan, the broader South Asian region (42, 43), and WHO and UNICEF global estimates, which indicate that children in South Asia experience an average of three to five episodes of acute illness per year, with fever and diarrhea predominating (44). The finding that nearly half of the study sample experienced acute morbidity in a two-week recall window points to the urgent need for strengthened preventive interventions. The geographic distribution of acute illness incidence revealed a striking 28 percentage point disparity between D.G. Khan (66%) and Gujranwala (38%), highlighting substantial subprovincial inequity in the childhood morbidity burden within Punjab. This pattern is consistent with evidence from comparable LMIC settings, demonstrating that geographic clustering of child morbidity reflects the uneven distribution of poverty, sanitation infrastructure, and healthcare access across administrative units (25). The D.G. Khan division is among the most economically disadvantaged divisions in Punjab, with the highest proportion of households in the poorest wealth quintile and the most limited access to basic health services (45).

4.2 Model Performance and Methodological Considerations

The random forest model achieved an accuracy of 60.5% and an AUC of 0.63, which, while modest in absolute terms, represents a statistically significant improvement over chance classification (NIR = 50.6%, $p < 0.001$). This level of performance is consistent with comparable machine learning studies of childhood health outcomes using population survey data. A study comparing machine learning and conventional epidemiological approaches for predicting childhood diarrhea using Nigerian DHS data reported a random forest AUC of 0.684, which was closely comparable to that of logistic regression and consistent with the modest discriminatory performance expected from population survey predictors (46). A concurrent meta-analysis of 11 studies applying machine learning to DHS data for malnutrition prediction reported a pooled random forest accuracy of 0.71, attributing the wide variability to differences in setting, outcome definitions, and the inherent limitations of survey-based proxies as individual-level classifiers (47).

The model demonstrated high sensitivity for correctly identifying healthy children (86.6%) but lower specificity for detecting diseased children (33.6%), indicating a bias toward predicting the majority class.

This pattern is common in ensemble tree models when the class distribution is moderately imbalanced and the predictors are predominantly population-level social variables rather than individual clinical markers (48). Importantly, the primary analytical objective of this study was not the clinical diagnosis of individual children, a purpose for which sensitivity for the minority class (diseased children) would be the paramount criterion, but rather the identification of population-level determinants and simulation of policy effects. For this purpose, the AUC and variable importance provide the most meaningful evaluation criteria, as they assess the model's ability to rank children by predicted risk and identify the relative contribution of each predictor to outcome variability (49).

4.3 Key Predictors of Childhood Morbidity

The degree of vitamin and mineral supplement access emerged as the most important predictor of the mean decrease in Gini importance. This finding is corroborated by extensive literature linking micronutrient deficiency, particularly vitamin A, zinc, and iron deficiency, to increased susceptibility to infectious disease in young children (50). Zinc supplementation in children under five years of age has been shown to reduce diarrheal incidence by 13% and pneumonia morbidity by 19%, with estimated reductions in diarrhea and pneumonia mortality of 13% and 15%, respectively (51). Several meta-analyses have confirmed that zinc and vitamin A supplementation significantly reduce the incidence and severity of diarrhea and pneumonia in children under five years of age in LMICs (52, 53). The extremely low supplement coverage in this sample (15%) points to a critical gap in preventive micronutrient programming in Punjab, which is consistent with findings from comparable settings in Pakistan (54, 55).

The wealth quintile index and maternal education ranked as the second and third most important predictors, respectively, which is consistent with the established literature on social determinants of child health. Socioeconomic gradients in child morbidity are well documented in South Asia: children from the poorest households face a two- to threefold higher risk of acute illness than do those from wealthier families, which is mediated through pathways including food insecurity, poor water and sanitation, overcrowding, and limited healthcare access (43, 56). Maternal education is recognized as one of the most powerful determinants of child health globally, operating through improved care-seeking behavior, better hygiene practices, superior dietary practices, and greater uptake of preventive services (57). The finding that 41% of mothers in this sample had no formal education highlights a structural vulnerability that substantially limits protective caregiving behaviors.

Bottle feeding was identified as an important predictor of disease status, which is consistent with established evidence that bottle-fed children have a significantly greater risk of diarrheal disease and respiratory infections than exclusively breastfed children do (58). The high prevalence of bottle feeding (59%) in this sample is concerning, given that 63% of the children were also still breastfeeding, suggesting that mixed feeding with bottle use may drive elevated morbidity risk. This finding aligns with a rapid review and meta-analysis reporting that exclusive breastfeeding has been shown to reduce diarrheal disease incidence by 43–50% in infants under six months compared with nonexclusive or formula feeding (59).

Anthropometric indicators (height-for-age and weight-for-height categories) have intermediate importance, confirming the bidirectional relationship between malnutrition and infection that has been extensively documented in the literature: malnourished children are more susceptible to infection, and recurrent infection further impairs nutritional status (60). The stunting prevalence of 29% and wasting prevalence of 8.9% observed in this sample are below the WHO emergency thresholds but remain substantially above global averages, indicating a chronic nutritional crisis that sustains elevated morbidity risk.

4.4 Counterfactual Simulation Findings

The counterfactual simulation identified vitamin and mineral supplement access as the single largest modifiable determinant of population-level health rates, with a predicted improvement of + 12.79 percentage points. This finding has direct programmatic implications. Punjab's existing nutrition programmes — including the National Nutrition Programme and the Punjab Food Authority micronutrient fortification initiatives — have focused predominantly on food fortification rather than targeted supplementation for underfive children. The simulation results suggest that targeted supplementation programs, reaching 85% of children currently without access, could yield substantial reductions in the acute morbidity burden at the population level. These findings are consistent with those of Nugroho and Handini (61), who demonstrated that micronutrient supplementation significantly improved nutritional outcomes in vulnerable populations, with iron reducing anemia rates, vitamin A decreasing infection-related morbidity and mortality, and zinc alleviating stunting while reducing diarrheal and respiratory infections in children (62).

The predicted effect of higher maternal education (+ 3.61 percentage points) reinforces the importance of investing in girls' education as a long-term child health intervention. This effect is comparable to a global systematic review and meta-analysis of 300 studies from 92 countries reporting that each additional year of maternal schooling is associated with a 3.04% reduction in underfive mortality (95% CI: 2.82–3.23%), which is mediated through improved health literacy, care-seeking behavior, and household decision-making capacity (63). The relatively modest absolute effect in this simulation reflects the multistep pathway through which education influences child health and the inherent limitations of single-variable counterfactual analysis in capturing these indirect mechanisms. The effect of achieving MDD (+ 0.47 percentage points) was the smallest among the three scenarios. This finding appears to underestimate the known nutritional importance of dietary diversity and could be interpreted in the context of the model's complexity: the MDD variable's relatively low Gini importance suggests that other predictors, particularly supplements and wealth, absorbed a large proportion of the dietary diversity signal, likely due to the high correlation between access to diverse foods and household economic status in this sample. The substantially larger counterfactual effect of supplement access (+ 12.79 percentage points) compared with MDD (+ 0.47 percentage points) is not simply a reflection of differential variable importance but reflects a fundamental difference in the wealth-correlation structure of these two interventions. Dietary diversity in low-income South Asian settings is strongly wealth-constrained, and affordable sources of key micronutrients from food are largely inaccessible to the

poorest households in Pakistan (64), meaning that the training data rarely contain the combination of achieving MDD alongside low wealth. Counterfactual simulation of this combination therefore extrapolates outside the model's training distribution, producing a conservative estimate. In contrast, supplement access is distributed through government and international programme channels, including Lady Health Worker outreach and UNICEF campaigns, which specifically target poor households regardless of wealth quintile (65, 66), meaning that the model encounters supplement access across all wealth levels during training. The counterfactual simulation of supplement access therefore remains within the model's learned distribution, producing a more reliable and generalizable prediction effect.

The marked geographic heterogeneity in acute illness prevalence across Punjab's nine divisions – spanning 38% in Gujranwala to 66% in D.G. Khan – underscores the importance of geographically targeted implementation of supplementation and maternal education programmes, with D.G. Khan, Sargodha, and Rawalpindi representing the highest-priority divisions for immediate public health investment. This would maximize the population-level impact of the policy interventions identified in the counterfactual simulation

4.5 Strengths and Limitations

This study has several important strengths. The use of a large, population-representative household survey provides generalisability to the population up to 23 months of age in Punjab. The integration of machine learning with counterfactual simulation represents a methodological advance over conventional regression analyses of MICS data, enabling a transition from association description to policy-relevant effect estimation. The systematic hyperparameter tuning procedure ensures that the model is optimized prior to evaluation. The reporting of model performance against the no-information rate benchmark provides an appropriate context for interpreting the predictive results.

Several limitations must be acknowledged. First, complete-case analysis reduced the analytical sample from 11,364 to 10,128 children, reflecting the cumulative effect of requiring non-missing values across nine predictor variables simultaneously, a trade-off inherent to multivariate survey-based modeling that has been documented in comparable MICS analyses (67). Multiple imputation was initially applied as a sensitivity analysis; however, given that missing data in household surveys such as MICS are concentrated in the poorest and most rural households, a pattern consistent with Missing Not At Random (MNAR) rather than Missing At Random (MAR) mechanisms, complete case analysis was retained as the primary analytic approach. Research has shown that under MNAR, multiple imputation assuming MAR can itself produce biased estimates, whereas complete case analysis remains valid (68). The analytical sample of 10,128 children with complete records provided adequate statistical power for the random forest model and the counterfactual simulation, and the substantive findings were consistent across both approaches. Second, acute morbidity was ascertained by two-week caregiver recall, which is subject to recall bias and may underestimate or overestimate true illness episodes. Future research should integrate digital health-based real-time symptom recording to replace recall-based ascertainment, enabling more precise outcome measurement. A scoping review of data science and AI applications in maternal, newborn and child health identified real-time digital data collection as

the single most impactful methodological advance for improving ML model accuracy in LMIC child health research, underscoring the importance of this transition for studies such as ours (69). Third, the low sensitivity of the model for detecting diseased children (30.0%) reflects the inherent limitations of population-level social variables as individual-level diagnostic classifiers; the model's utility lies in population-level risk stratification rather than individual prediction.

5. Conclusions

This study demonstrates that random forest machine learning combined with counterfactual policy simulation provides a powerful analytical framework for translating population survey data into actionable child health policy guidance. Among the policy-relevant determinants of acute childhood morbidity examined in Punjab, vitamin and mineral supplement access emerged as the highest-impact modifiable factor, with universal supplement access predicted to increase population-level health rates by 12.79 percentage points. Maternal education and dietary diversity represent secondary priorities with meaningful but modest predictive effects. The findings support urgent investment in targeted micronutrient supplementation programmes, education, and dietary diversity interventions as high-yield strategies for reducing the acute morbidity burden among children in Punjab, especially in divisions with the highest burden of illness, such as D.G. Khan, who reported the highest prevalence of acute morbidity (66%), Sargodha (56.8%), and Rawalpindi (52.3%) as identified via geographical mapping. This analytical framework is replicable via MICS data from other provinces and countries, offering a scalable tool for evidence-based child health policymaking in similar resource-constrained settings.

Abbreviations

AUC-ROC

Area under the curve

DHS

Demographic and Health Surveys

D.G. Khan

Dera Ghazi Khan

HAZ

Height-for-age z-score

IYCF

Infant and Young Child Feeding

IMCI

Integrated Management of Childhood Illness

LMICs

Low- and middle-income countries

MAR

Missing at random

MDG

Mean decrease Gini
MDD
Minimum Dietary Diversity
MICS
Multiple Indicator Cluster Surveys
ML
Machine learning
MNAR
Missing not at random
NIR
No-information rate
OCHA
United Nations Office for the Coordination of Humanitarian Affairs
RF
Random Forest
ROC
Receiver operating characteristic
SD
Standard deviation
UNICEF
United Nations Children's Fund
WHO
World Health Organization
WHZ
Weight-for-height z-score
WQI
Wealth Quintile Index

Declarations

Ethics Approval and Consent to Participate:

This study constitutes a secondary analysis of deidentified, publicly available data from the Punjab MICS 2017–18 (31). The original survey was conducted under ethical clearance by the Bureau of Statistics, Government of Punjab, and UNICEF Pakistan. Informed consent was obtained from all participating households during the original survey. No additional ethical approval was required for this secondary analysis.

Clinical trial number

Not applicable.

Consent for Publication:

Not applicable.

Competing Interests:

The authors declare that they have no competing interests.

Funding:

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Author Contribution

BM: conceptualization, data curation, data analysis, writing – original draft. FA: supervision, data analysis, writing – review and editing. MHM: methodology, validation, writing–review and editing. MK: supervision, methodology, writing–review and editing. MAA: methodology, writing – review and editing. All the authors read and approved the final manuscript.

Acknowledgments:

The authors acknowledge the Bureau of Statistics, the Government of Punjab, and UNICEF Pakistan for making the MICS 2017–18 data publicly available.

Data Availability

The Punjab MICS 2017–18 dataset is publicly available upon request from the UNICEF MICS program (<https://mics.unicef.org/surveys>). The analyzed dataset and R analysis code are available in the Zenodo repository (<https://doi.org/10.5281/zenodo.21328577>).

References

1. Reiner RC, Olsen HE, Ikeda CT, Echko MM, Ballestreros KE, Manguerra H, et al. Diseases, injuries, and risk factors in child and adolescent health, 1990 to 2017: findings from the global burden of diseases, injuries, and risk factors 2017 study. *JAMA Pediatr.* 2019;173(6):e190337–e. <https://doi.org/10.1001/jamapediatrics.2019.0337>.
2. Fact sheets-malnutrition: World Health Organization. 2021 [[Accessed 25 March 2025]. Available from: <https://www.who.int/news-room/fact-sheets/detail/child-mortality-under-5-years>

3. Levels & Trends in Child Mortality: Report 2022, WHO/UNICEF/World Bank Group. [Accessed 01 Jun 2025]. Available from: <https://childmortality.org/wp-content/uploads/2023/01/UN-IGME-Child-Mortality-Report-2022.pdf>
4. Hailu M, Kassahun W, Yesuf HA, Lingerew A. Serum electrolyte disturbances and treatment outcomes in hospitalized children with Severe Acute Malnutrition in Ethiopia. *Sci Rep.* 2026;16. <https://doi.org/10.1038/s41598-026-44123-1>.
5. Facts Sheets: Malnutrition: World Health Organization. 2024 [[Accessed 01 Jun 2026]. Available from: <https://www.who.int/news-room/fact-sheets/detail/malnutrition>
6. Levels and trends in child malnutrition: UNICEF/WHO/World Bank Group Joint Child Malnutrition Estimates: Key findings of the 2023 edition. World Health Organization. 2023. Report No.: 9240073795 [Accessed 26 March 2025]. Available from: <https://www.who.int/publications/i/item/9789240073791>
7. Tharwani ZH, Bilal W, Khan HA, Kumar P, Butt MS, Hamdana AH, et al. Infant & child mortality in Pakistan and its determinants: a review. *Inquiry.* 2023;60:00469580231167024. <https://doi.org/10.1177/004695802311670>.
8. Kumar R, Mahmood T, Naeem N, Khan SA, Hanif M, Pongpanich S. Minimum dietary diversity and associated determinants among children aged 6–23 months in Pakistan. *Sci Rep.* 2024;14(1):2639. <https://doi.org/10.1038/s41598-024-51705-4>.
9. Shaukat S, Ullah A, Chani MI. Child Mortality and its Socio-Economic, Maternal and Environmental Determinants in Pakistan: An Empirical Evidence from Micro Level Data. *Adv J Econom Finance.* 2026;4(1):170–80. <https://doi.org/10.63075/wxarg680>.
10. Trends in under-five mortality rate in Pakistan: UNICEF. 2023 [[Accessed 10 Jun 2025]. Available from: <https://data.unicef.org/country/pak/>
11. Saleem J, Zakar R, Butt MS, Aadil RM, Ali Z, Bukhari GMJ, et al. Application of the Boruta algorithm to assess the multidimensional determinants of malnutrition among children under five years living in southern Punjab, Pakistan. *BMC Public Health.* 2024;24(1):167. <https://doi.org/10.1186/s12889-024-17701-z>.
12. Zainab U, Abbas M, Urooj A, Rabia M, Sun H, Ahmed Shehzad M. Undesired nexus poor health status of child under-five: A case study of Pakistan. *PLoS ONE.* 2025;20(5):e0323845. <https://doi.org/10.1371/journal.pone.0323845>.
13. Wang F, Gillani S, Balsalobre-Lorente D, Shafiq MN, Khan KD. Environmental degradation in South Asia: Implications for child health and the role of institutional quality and globalization. *Sustain Dev.* 2025;33(1):399–415. <https://doi.org/10.1002/sd.3124>.
14. Hayat B, Pervaiz S, Khan Z, Zaidi SHA, Masood A, Sherazi SH, et al. The Impact of Integrated Community-Based Management of Respiratory Infections in Reducing Child Mortality: Integrated Community-Based Management of Respiratory Infections in Reducing Mortality. *Pak J Health Sci.* 2024;05(10):233–9. <https://doi.org/10.54393/pjhs.v5i10.2365>.

15. Zakar R, Zakar NM, Shahzad R, Tekian A, Fischer F. Community health needs assessment: findings from a community-university partnership strengthening program on maternal and child health in Pakistan. *BMC Public Health*. 2025;25(1):654. <https://doi.org/10.1186/s12889-025-21947-6>.
16. Fatima S, Haider M, Hameed A, Saleem SG, Karim S. Comorbidities and their outcomes in children with severe acute malnutrition visiting pediatric emergency department at a tertiary care hospital in urban slums of Karachi, Pakistan. *Ann Pediatr*. 2021;4(1):1068.
17. Rehan KZ, Sania Z. Child Malnutrition in Pakistan: Evidence from National Nutritional Survey 2018. *CEERAT J Soc Dev*. 2025;2(1):9–22.
18. Zhou Y, Xu Y. Nutrition and Metabolism in the First 1000 Days of Life. *Nutrients*. 2023;15(11):2554. <https://doi.org/10.3390/nu15112554>.
19. Juan J, Yang H. Early life 1000 days: Opportunities for preventing adult diseases. *Chin Med J*. 2022;135(05):516–8. <https://doi.org/10.1097/CM9.0000000000001920>.
20. Schwarzenberg SJ, Georgieff MK, Committee on N, Daniels S, Corkins M, Golden NH, et al. Advocacy for improving nutrition in the first 1000 days to support childhood development and adult health. *Pediatrics*. 2018;141(2):e20173716. <https://doi.org/10.1542/peds.2017-3716>.
21. Draper CE, Yousafzai AK, McCoy DC, Cuartas J, Obradović J, Bhopal S, et al. The next 1000 days: building on early investments for the health and development of young children. *Lancet (London England)*. 2024;404(10467):2094–116. [https://doi.org/10.1016/s0140-6736\(24\)01389-8](https://doi.org/10.1016/s0140-6736(24)01389-8).
22. Malik MI, Nadeem M, Haq W. Factors Contributing to Disparities in Early Child Development in Punjab, Pakistan: An In-depth Analysis Through National Survey. *Child Indic Res*. 2025;18(5):2223–43. <https://doi.org/10.1007/s12187-025-10242-w>.
23. Wazir MA, Goujon A. Assessing the 2017 census of Pakistan using demographic analysis: A sub-national perspective. *Vienna Institute of Demography Working Papers*; 2019. [Accessed 22 Jun 2025]. Available from: <https://www.econstor.eu/bitstream/10419/207062/1/1667013416.pdf>
24. Gardezi NuZ. Public health insurance and birth outcomes: evidence from Punjab, Pakistan. *Health Policy Plan*. 2020;36(1):1–13. <https://doi.org/10.1093/heapol/czaa115>.
25. Zhang Y, Akram S, Hassan MU, Shaheen M, Shahzad MF. Inequality of opportunity in early childhood development and health among children under five in Pakistan: a multidimensional analysis using MICS data. *Crit Public Health*. 2026;36(1):2600716. <https://doi.org/10.1080/09581596.2025.2600716>.
26. Faisal S, Zahid FM, Kamal S, Shahzad K, Seidu A-A, Ahinkorah BO. Modeling the factors associated with incomplete immunization among children. *Math Probl Eng*. 2022;2022(1):8460837.
27. Khalilia M, Chakraborty S, Popescu M. Predicting disease risks from highly imbalanced data using random forest. *BMC Med Inf Decis Mak*. 2011;11(1):51. <https://doi.org/10.1186/1472-6947-11-51>.
28. Fenta HM, Zewotir T, Muluneh EK. A machine learning classifier approach for identifying the determinants of under-five child undernutrition in Ethiopian administrative zones. *BMC Med Inf Decis Mak*. 2021;21(1):291. <https://doi.org/10.1186/s12911-021-01652-1>.

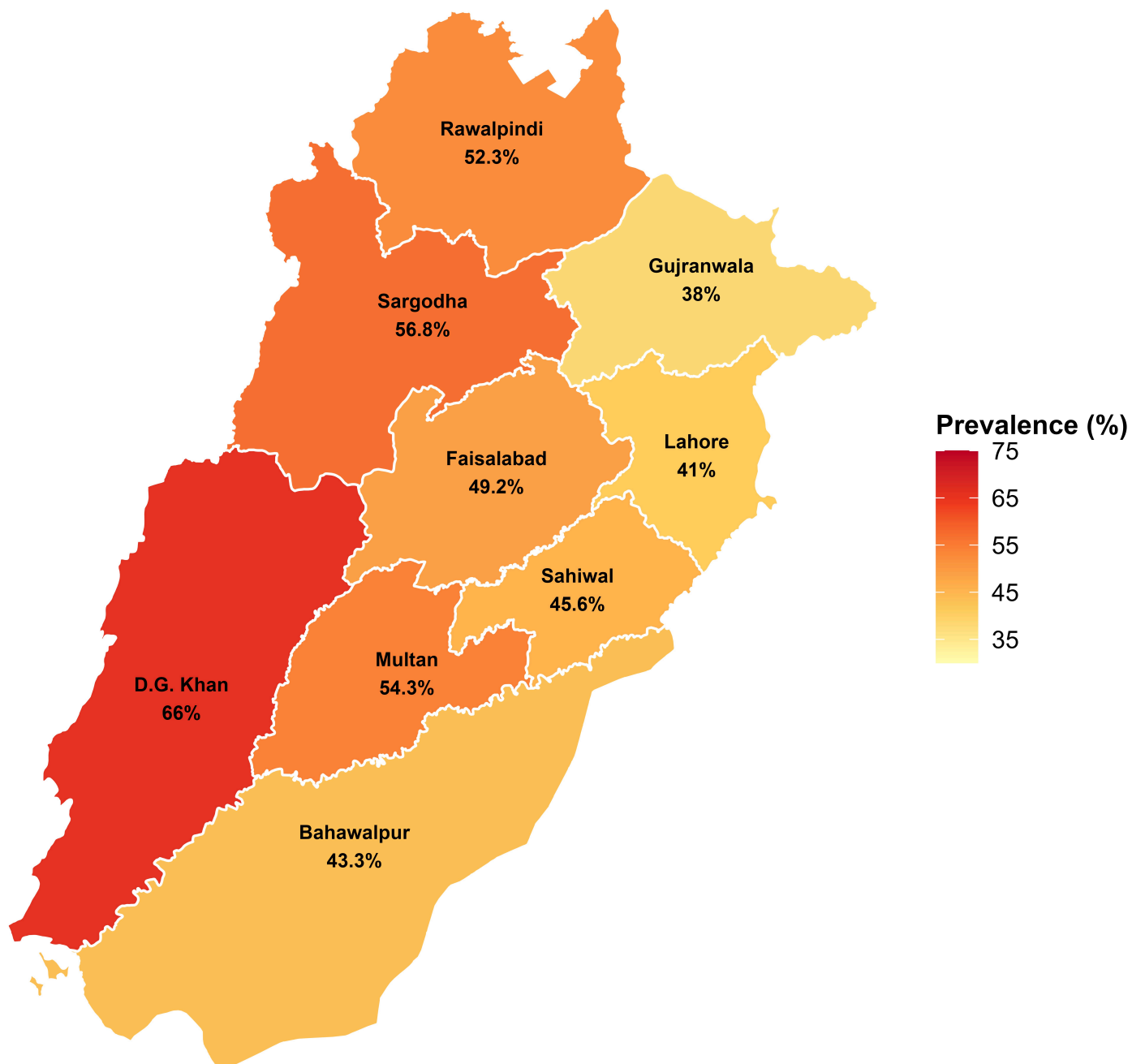
29. Talukder A, Ahammed B. Machine learning algorithms for predicting malnutrition among under-five children in Bangladesh. *Nutrition*. 2020;78:110861. <https://doi.org/10.1016/j.nut.2020.110861>.
30. Wang X, Komura T, Arakawa Y, Chen R, Nakagomi A, Steptoe A, et al. Estimating effects of hypothetical public health interventions on health disparities: a standardization-based simulation approach. *Am J Epidemiol*. 2026;195(6):1585–95. <https://doi.org/10.1093/aje/kwaf279>.
31. Multi Indicator Cluster Surveys Pakistan (MICS 2017-18) [Internet]. UNICEF. 2017. Available from: [https://mics.unicef.org/surveys?display=card&f\[0\]=region:2521&f\[1\]=year:2017](https://mics.unicef.org/surveys?display=card&f[0]=region:2521&f[1]=year:2017)
32. Indicators for assessing infant and young child feeding practices: definitions and measurement methods. World Health Organization. 2021. Report No.: 9240018387 [Accessed 16 March 2025]. Available from: <https://www.who.int/data/gho/data/indicators/indicator-details/GHO/minimum-dietary-diversity-6-23-months>
33. Integrated Management of Childhood Illness Chart Booklet. Geneva: World Health Organization. 2014. [Accessed 26 May 2025]. Available from: <https://cdn.who.int/media/docs/default-source/mca-documents/child/imci-integrated-management-of-childhood-illness/imci-in-service-training/imci-chart-booklet.pdf>
34. Ahinkorah BO. Maternal age at first childbirth and under-five morbidity in sub-Saharan Africa: analysis of cross-sectional data of 32 countries. *Archives of public health = Archives belges de sante publique*. 2021;79(1):151.10.1186/s13690-021-00674-5.
35. Salman HA, Kalakech A, Steiti A. Random forest algorithm overview. *Babylon J Mach Learn*. 2024;2024:69–79. <https://doi.org/10.58496/BJML/2024/007>.
36. Yang L, Shami A. On hyperparameter optimization of machine learning algorithms: Theory and practice. *Neurocomputing*. 2020;415:295–316. <https://doi.org/10.1016/j.neucom.2020.07.061>.
37. Lumumba VW, Wanjuki TM, Njoroge EW. Evaluating the Performance of Ensemble and Single Classifiers with Explainable Artificial Intelligence (XAI) on Hypertension Risk Prediction. *Comput Intell Mach Learn*. 2025;6(1):22–34. <https://doi.org/10.36647/CIML/06.01.A004>.
38. Wang H, Yang F, Luo Z. An experimental study of the intrinsic stability of random forest variable importance measures. *BMC Bioinformatics*. 2016;17(1):60. <https://doi.org/10.1186/s12859-016-0900-5>.
39. Harvey JS, Feng G, Meesala SA, Zhao T, Mehta D. Interpretable Model-Aware Counterfactual Explanations for Random Forest. *arXiv preprint arXiv:251027397*: arXiv preprint; 2025.
40. Jabeen S, Saha UR, van Wesenbeeck CFA, Mushtaq K. An overview of diarrhea among infants and under-five in Punjab-Pakistan. *J Pediatr Nurs*. 2023;71:e28–37. <https://doi.org/10.1016/j.pedn.2023.04.011>.
41. Punjab Survey Finding Report:Multiple Indicator Cluster Survey. Islamabad. 2018. [Accessed 2 Feb 2025]. Available from: <https://www.unicef.org/pakistan/reports/punjab-mics-2017-18>
42. Hossain J, Md. Ripon Rouf AS, Tareq M, Rokunuzzaman M, Sharma SKD. Assessing the impact of early marriage and socioeconomic determinants on under-five morbidity: a cross-country analysis in South Asia. *BMC Pediatr*. 2026;26(1):220. <https://doi.org/10.1186/s12887-026-06596-x>.

43. Mahmud S, Sultana R, Hossain MS, Akhter M, Hossain ME. Multilevel modeling and mediation analysis of acute respiratory infections among under five children in Bangladesh: evidence from multiple indicator cluster survey. *BMC Public Health*. 2026;26. <https://doi.org/10.1186/s12889-026-27018-8>.
44. Pneumonia and Diarrhoea: Tackling the Deadliest Diseases for the World's Poorest Children. New York: WHO/UNICEF. 2012. [Accessed 25 March 2025]. Available from: https://data.unicef.org/wp-content/uploads/2015/12/Pneumonia_Diarrhoea_2012_35.pdf
45. Ilyas M, Bashir MK, Ali A, Zulfikar H. Understanding poverty distribution in Punjab, Pakistan: a disaggregated approach. *J Sustain Econ Develop*. 2025;3(1):31–40. <https://doi.org/10.22194/JSD/25.41>.
46. Ashaolu JO, Do LT, Isaac KR, Akanji TS, Some SYM. Determinants of childhood diarrhea in low- and middle-income countries: a comparative analysis of epidemiological and machine learning approaches. *BMC Public Health*. 2025;25(1):3845. <https://doi.org/10.1186/s12889-025-25076-y>.
47. Rao B, Rashid M, Hasan MG, Thunga G. Machine Learning in Predicting Child Malnutrition: A Meta-Analysis of Demographic and Health Surveys Data. *Int J Environ Res Public Health*. 2025;22(3). <https://doi.org/10.3390/ijerph22030449>.
48. Fletcher RR, Nakeshimana A, Olubeko O. Addressing fairness, bias, and appropriate use of artificial intelligence and machine learning in global health. *Front Artif Intell*. 2021;3:561802. <https://doi.org/10.3389/frai.2020.561802>.
49. Laitinen EK, Gini Coefficient, AUC in Assessing Predictive Model Performance. Effect of Ranks. *Univ Vaasa Open Repository (Osuva)*. 2025;15(5):1251–66. <https://doi.org/10.4236/tel.2025.155070>.
50. Black RE, Victora CG, Walker SP, Bhutta ZA, Christian P, de Onis M, et al. Maternal and child undernutrition and overweight in low-income and middle-income countries. *Lancet (London England)*. 2013;382(9890):427–51. [https://doi.org/10.1016/S0140-6736\(13\)60937-X](https://doi.org/10.1016/S0140-6736(13)60937-X).
51. Yakoob MY, Theodoratou E, Jabeen A, Imdad A, Eisele TP, Ferguson J, et al. Preventive zinc supplementation in developing countries: impact on mortality and morbidity due to diarrhea, pneumonia and malaria. *BMC Public Health*. 2011;11(S3):S23. <https://doi.org/10.1186/1471-2458-11-s3-s23>.
52. Harika R, Faber M, Samuel F, Mulugeta A, Kimiywe J, Eilander A. Are low intakes and deficiencies in iron, vitamin A, zinc, and iodine of public health concern in Ethiopian, Kenyan, Nigerian, and South African children and adolescents? *Food Nutr Bull*. 2017;38(3):405–27. <https://doi.org/10.1177/0379572117715818>.
53. Chitekwe S, Parajuli KR, Paudyal N, Haag KC, Renzaho A, Issaka A, et al. Individual, household and national factors associated with iron, vitamin A and zinc deficiencies among children aged 6–59 months in Nepal. *Matern Child Nutr*. 2022;18(S1):e13305. <https://doi.org/10.1111/mcn.13305>.
54. Key Findings from National Nutrition Survey. Islamabad. 2018. [Accessed 24 Jan 2025]. Available from: <https://www.unicef.org/pakistan/media/1871/file/KeyFindings---NationalNutritionSurvey2018.pdf>

55. Victora CG, Christian P, Vidaletti LP, Gatica-Domínguez G, Menon P, Black RE. Revisiting maternal and child undernutrition in low-income and middle-income countries: variable progress towards an unfinished agenda. *Lancet (London England)*. 2021;397(10282):1388–99. [https://doi.org/10.1016/S0140-6736\(21\)00394-9](https://doi.org/10.1016/S0140-6736(21)00394-9).
56. Joe W, Prakash A, Yaqoob Azimi SM, Galvin M, Murira Z, Paez Salamanca GN, et al. Levels, severity, and determinants of stunting in children 0–59 months in Afghanistan: Secondary analysis of Multiple Indicator Cluster Survey, 2022-23. *PLOS Glob Public Health*. 2025;5(4):e0004423. <https://doi.org/10.1371/journal.pgph.0004423>.
57. Rezaeizadeh G, Mansournia MA, Keshtkar A, Farahani Z, Zarepour F, Sharafkhah M, et al. Maternal education and its influence on child growth and nutritional status during the first two years of life: a systematic review and meta-analysis. *eClinicalMedicine*. 2024;71. <https://doi.org/10.1016/j.eclinm.2024.102574>.
58. Nguyen P, Binns CW, Ha AVV, Chu TK, Nguyen LC, Duong DV, et al. Prelacteal and early formula feeding increase risk of infant hospitalisation: a prospective cohort study. *Arch Dis Child*. 2020;105(2):122–6. <https://doi.org/10.1136/archdischild-2019-316937>.
59. Ganbold G, Farnaz N, Scutts T, Borg B, Mhrshahi S. The Association Between Exclusive Breastfeeding and Diarrhoea Morbidity in Infants Aged 0–6 Months: A Rapid Review and Meta-Analysis. *Matern Child Nutr*. 2025;21(3):e70042. <https://doi.org/10.1111/mcn.70042>.
60. Morales F, Montserrat-De la Paz S, Leon MJ, Rivero-Pino F. Effects of malnutrition on the immune system and infection and the role of nutritional strategies regarding improvements in children's health status: a literature review. *Nutrients*. 2023;16(1):1. <https://doi.org/10.3390/nu16010001>.
61. Nugroho MI, Handini LS. The effectiveness of micronutrient supplementation in improving nutritional status: a literature review. *J Divers Med Res*. 2025;2(3):97–103. <https://doi.org/10.33005/jdiversemedres.v2i3.143>.
62. Csölle I, Felső R, Szabó É, Metzendorf M-I, Schwingshackl L, Ferenci T, et al. Health outcomes associated with micronutrient-fortified complementary foods in infants and young children aged 6–23 months: a systematic review and meta-analysis. *Lancet Child Adolesc Health*. 2022;6(8):533 – 44. [https://doi.org/10.1016/S2352-4642\(22\)00147-X](https://doi.org/10.1016/S2352-4642(22)00147-X).
63. Balaj M, York HW, Sripada K, Besnier E, Vonen HD, Aravkin A, et al. Parental education and inequalities in child mortality: a global systematic review and meta-analysis. *Lancet (London England)*. 2021;398(10300):608–20. [https://doi.org/10.1016/s0140-6736\(21\)00534-1](https://doi.org/10.1016/s0140-6736(21)00534-1).
64. Ryckman T, Beal T, Nordhagen S, Murira Z, Torlesse H. Affordability of nutritious foods for complementary feeding in South Asia. *Nutr Rev*. 2021;79(Suppl 1):52–68. <https://doi.org/10.1093/nutrit/nuaa139>.
65. Pakistan Pograms, Islamabad. Nutrition International; 2025. [Accessed 22 Mar 2025]. Available from: <https://nutritionintl.org/wp-content/uploads/2025/09/Pakistan-Country-Profile-June-2025-WEB.pdf>

66. Khan GN, Kureishy S, Ariff S, Habib MA, Usmani AA, Mubarik A, et al. Specialized Nutritious Food Combined With Cash Transfers and Social and Behavior Change Communication to Prevent Stunting Among Children Aged 6 to 23 Months in Pakistan: Protocol for a Cluster Randomized Controlled Trial. *JMIR Res protocols*. 2020;9(8):e19001. <https://doi.org/10.2196/19001>.
67. Ahissou NCA, Nonaka D, Takeuchi R, de los Reyes C, Uehara M, Khampheng P, et al. Trend of sociodemographic and economic inequalities in the use of maternal health services in Lao People's Democratic Republic from 2006 to 2017: MICS data analysis. *Trop Med Health*. 2023;51(1):56. <https://doi.org/10.1186/s41182-023-00548-2>.
68. Hughes RA, Heron J, Sterne JAC, Tilling K. Accounting for missing data in statistical analyses: multiple imputation is not always the answer. *Int J Epidemiol*. 2019;48(4):1294–304. <https://doi.org/10.1093/ije/dyz032>.
69. Akuze J, Otieno GP, Amare SY, Ngatia B, Wanduru P, Kirakoya-Samadoulougou F, et al. Data science and artificial intelligence for maternal, newborn and child health: scoping review and thematic analysis. *BMC Public Health*. 2025;26(1):314. <https://doi.org/10.1186/s12889-025-25430-0>.

Figures



Data source: Punjab MICS 2017–18

Figure 1

Geographic distribution of acute illness prevalence among children across nine administrative divisions of Punjab.

Note: The color gradient represents the division-level prevalence ranging from 38% (Gujranwala) to 66% (D. G. Khan).

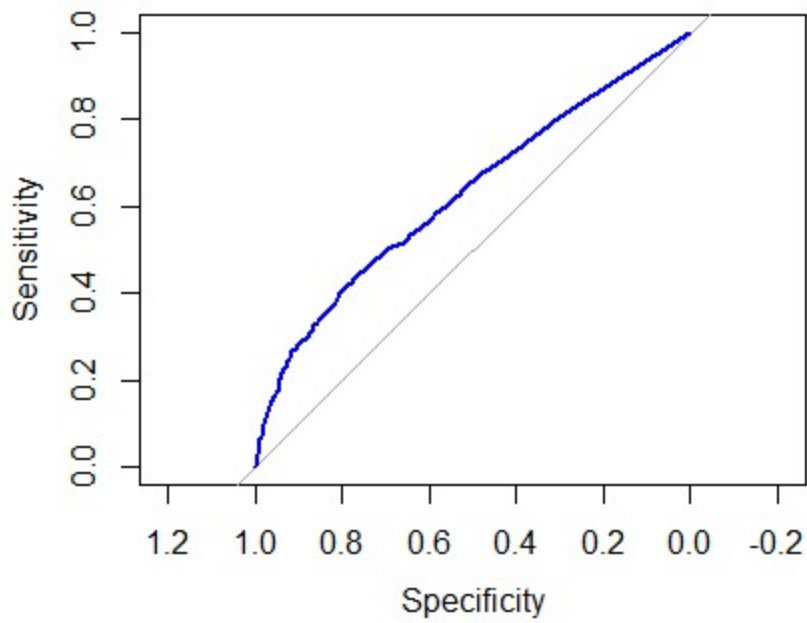


Figure 2

Receiver operating characteristic (ROC) curve for the tuned RF model on the testing dataset.

Note: The curve depicts the trade-off between sensitivity and specificity across classification thresholds.
AUC = 0.6322 (95% CI: 0.6125–0.6518),

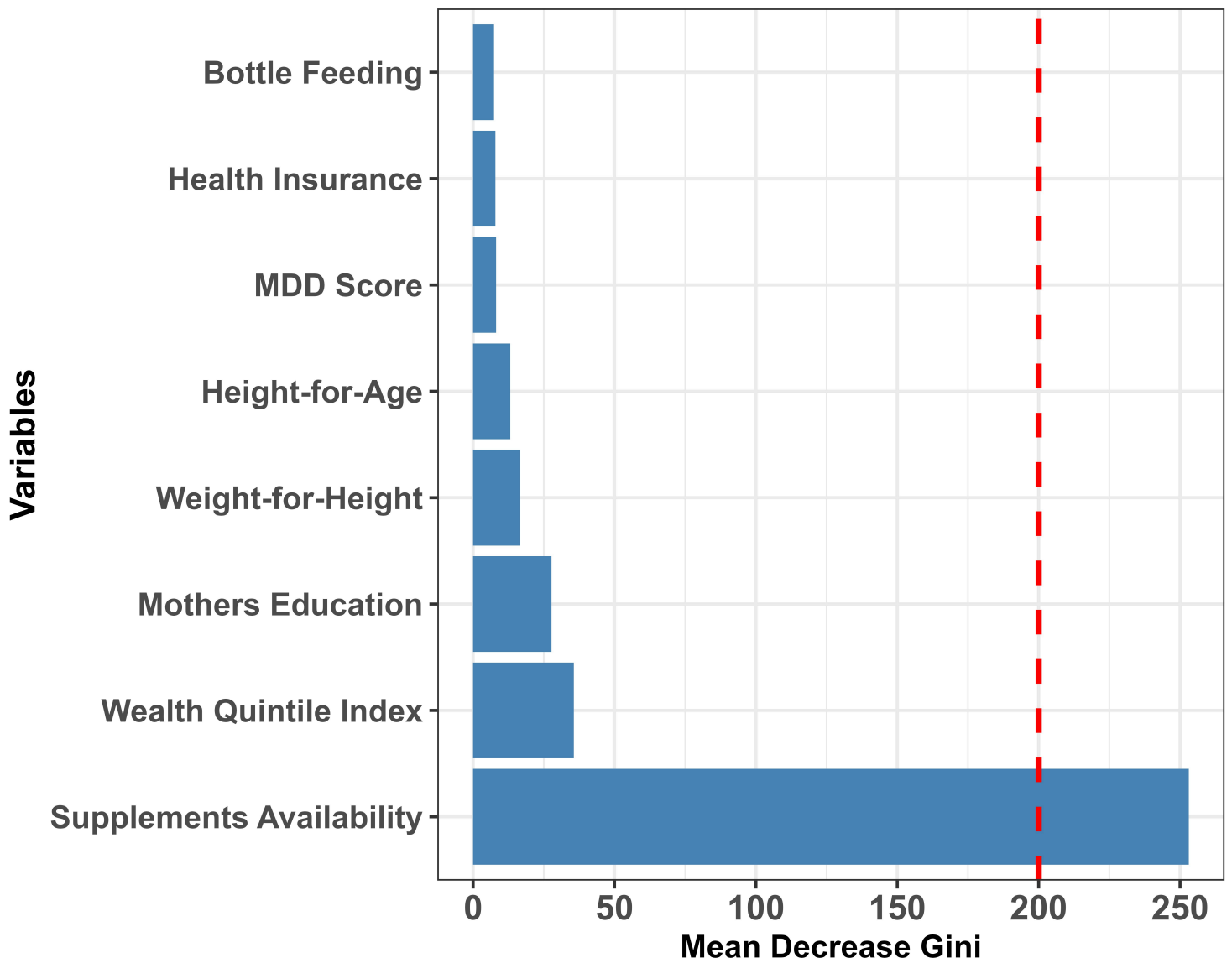


Figure 3

Variable importance plot (mean decrease Gini) for the tuned RF model.

Note: Supplements availability = Vitamin/Mineral supplements access; MDD score = Minimum Dietary Diversity Score

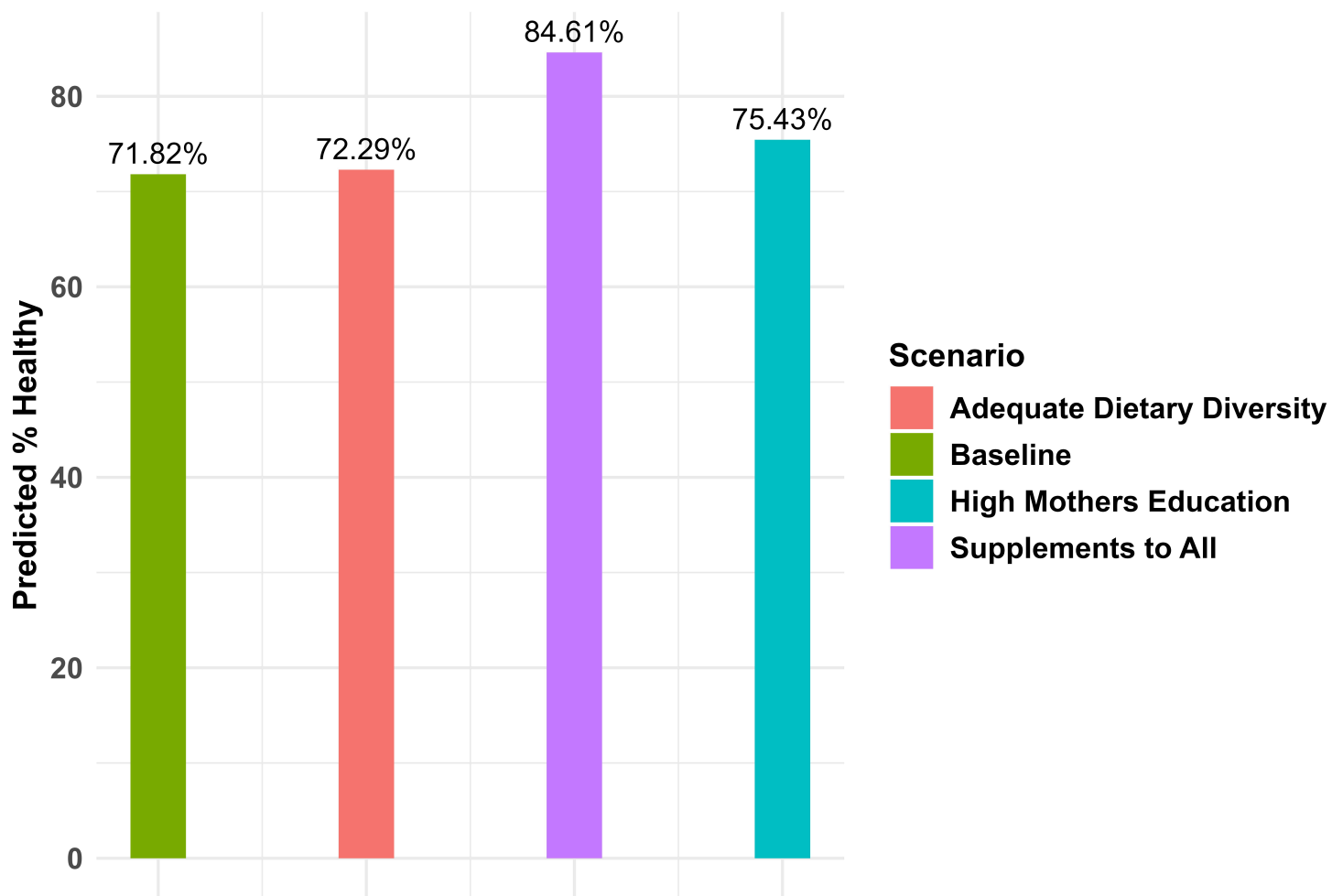


Figure 4

Bar chart depicting the predicted percentage of healthy children under each counterfactual scenario.

Note: The baseline represents the predicted health rate under the observed conditions (71.82%).
Supplements access produced the largest predicted improvement (84.61%, +12.79 percentage points).

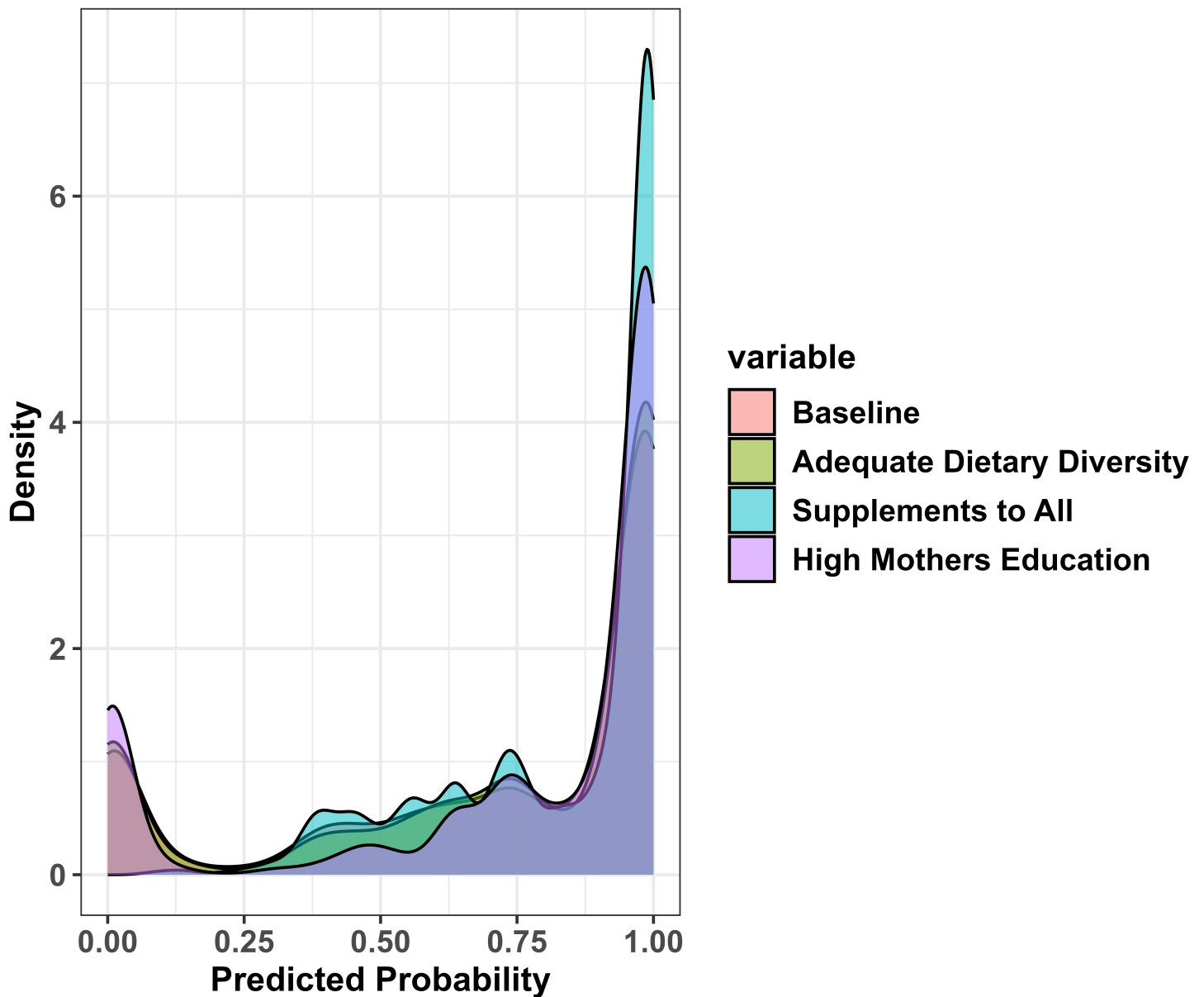


Figure 5

Density plot of individual predicted probabilities of being healthy across all counterfactual scenarios.

Note: The spike at probability = 1.0 reflects the model’s strong tendency to assign a high healthy probability to the majority of observations; the shift of the “Supplements to All” distribution toward higher probabilities confirms the population-level benefit of supplementation.