AMERICAN Journal of Science on Integration and Human Development
Volume 4, Issue 6, 2026 ISSN (E): 2993-2750

Prevalence, Determinants, and Socioeconomic Inequality of Anaemia Among Women of Reproductive Age in Odisha and Jharkhand, India: A Survey-Weighted Analysis with Machine Learning Insights Using NFHS-5 Data

Deepti Rani Pattanaik[1], Monalisha Pattnaik[2]
[1&2] Department of Statistics, Sambalpur University, Sambalpur, Odisha, India
Email- 97deeptirani@gmail.com

Abstract: Background: Anaemia among females of reproductive age (15-49 years) is a major community health crisis in India. Odisha and Jharkhand, two states with large tribal and economically disadvantaged populations, report anaemia burdens substantially above the national average.
Objective: This study sought to define the dominance of anaemia in Odisha and Jharkhand, investigate its socioeconomic and nutritional causes using survey-weighted multivariable logistic regression, measure wealth-related inequality via the Concentration Index, and pinpoint key predictors using Random Forest classification.
Methods: The analysis performed on pooled data from the National Family Health Survey-5 (NFHS-5, 201921), comprising 27,971 women from Odisha and 26,495 women from Jharkhand, for a entire sample of 54,466 women age bearing 15 to 49 years. Survey-weighted logistic regression models (svyglm with quasibinomial family), Random Forest classification (ntree = 500), Variance Inflation Factor (VIF) analysis, Concentration Index (CI), and Erreygers' Normalized Concentration Index (ECI) were used. Jharkhand, the middle wealth category, and higher education were used as reference groups.
Results: The survey-weighted prevalence of anaemia was 67.1% in Jharkhand and 65.5% in Odisha. There was no statistically significant difference between states (AOR=1.00; 95% CI: 0.951.07; p=0.879). Key factors included: higher BMI (Protective effect, AOR=1.00 per unit; p < 0.001), no education (AOR=1.00 1.20; p < 0.001), primary education (AOR=1.00; p=0.003), secondary education (AOR=1.00; p < 0.001), poorest wealth status (AOR=1.00; p < 0.001), and poorer wealth status (AOR=1.00; p=0.010). The VIF analysis indicated the absence of multicollinearity. According to the Mean Decrease Gini metric, Random Forest identified BMI and Age as the most significant predictors. The Concentration Index (CI = -0.065; Erreygers' ECI = -0.177) indicated that anaemia was significantly more prevalent among poorer populations.
Conclusion: Anaemia in Odisha and Jharkhand is driven primarily by low BMI, lower education, and poverty rather than state of residence. Targeted nutrition, education, and poverty-alleviation interventions are urgently needed in both states.

Keywords: anaemia; NFHS-5; Odisha; Jharkhand; logistic regression; random forest; concentration index; women of reproductive age; socioeconomic determinants.

1. Background

Globally, anaemia is a major public health crisis that impacts around 1.62 billion individuals, with the highest burden felt in low- and middle-income nations. The World Health Organization (WHO) establishes the diagnostic threshold for anaemia in non-pregnant women of childbearing age as a haemoglobin level under 12.0 g/dL. The condition is associated with impaired physical and cognitive performance, reduced work productivity, weakened immunity, adverse pregnancy outcomes, and increased maternal morbidity. Women of reproductive age are especially vulnerable

due to menstrual blood loss, increased nutritional requirements, repeated pregnancies, and inadequate dietary intake, often compounded by socioeconomic disadvantage and poor access to healthcare.

India continues to bear a substantial share of the global anaemia burden. Data from the National Family Health Survey-5 (NFHS-5, 2019–21) indicate that 57.0% of women aged 15–49 years are anaemic, representing an increase from 53.1% reported in NFHS-4. Despite sustained policy efforts, including the National Iron Plus Initiative and Anaemia Mukt Bharat programme, the prevalence of anaemia vestiges alarmingly high. Previous studies have identified nutritional status, educational attainment, household wealth, and social disadvantage as major determinants of anaemia among Indian women. For instance, Ghosal[1] reported persistent socioeconomic inequalities in anaemia among tribal and non-tribal women across India, while Let[2] demonstrated that women from poorer and less educated households face significantly greater risks of anaemia. Similarly, Gnanasekaran[3], using NFHS-5 data, emphasized the important role of nutritional and socioeconomic factors in shaping anaemia prevalence females.

Odisha and Jharkhand are among the most socioeconomically vulnerable states in eastern India and continue to experience anaemia prevalence levels substantially above the national average. Both states have large Scheduled Tribe populations, widespread poverty, limited healthcare accessibility, and persistent nutritional deprivation. These structural disadvantages contribute to elevated risks of anaemia and other nutrition-related disorders. Although several studies have examined anaemia at the national level, comparative evidence focusing specifically on Odisha and Jharkhand remains limited. Furthermore, relatively few studies have simultaneously examined the determinants, predictive factors, and socioeconomic inequality associated with anaemia within these states.

In recent years, machine learning approaches have gained attention in public health research for identifying influential predictors and exploring complex relationships among health outcomes. Random Forest algorithms, in particular, have been widely used for variable importance assessment and predictive modelling. When combined with traditional epidemiological methods and inequality measures, such approaches can provide a more comprehensive understanding of disease burden and its underlying drivers.

Against this background, the present study investigates anaemia among females of generative age in Odisha and Jharkhand using NFHS-5 microdata. Specifically, the study aims to: (i) estimate and compare survey-weighted anaemia prevalence across the two states; (ii) identify socioeconomic and nutritional determinants of anaemia using multivariable logistic regression; (iii) assess the relative importance of predictors through Random Forest analysis; and (iv) evaluate wealth-related inequality in anaemia using the Concentration Index and Concentration Curve. By integrating conventional statistical methods with machine learning and inequality analysis, the study seeks to generate evidence that can inform targeted and equity-oriented interventions for reducing the burden of anaemia among vulnerable women in eastern India.

2. Methodology

2.1 Data Acquisition and Sample Pooling

This study adopts a cross-sectional framework to evaluate secondary data from fifth iteration of India's National Family Health Survey (NFHS-5), which was conducted between 2019 and 2021. an initiative executed by the IIPS and ICF International for the Ministry of Health and Family Welfare. We retrieved individual state-level data for Odisha (n = 27,971) and Jharkhand (n = 26,495) via the official DHS Program portal. By integrating a state-specific identifier, we successfully aggregated these distinct datasets into a single, pooled cohort. The final statistical analysis comprises (N = 54,466) reproductive-age women between 15 and 49 years old.

2.2 Outcome Variable

The binary outcome indicated the presence or absence of anaemia (yes/ no), based on the variable

AS: women categorized as ' not anemic' were assigned a value of 0 (No); those classified as mild, moderate, or severe were assigned a value of 1 (Yes). A binary numeric variable (anaemia binary: 0/1) was also created for use in regression modeling.

2.3 Exposure Variables

The statistical models adjusted for a comprehensive set of sociodemographic and health indicators. Geographic variation was captured via a state variable, using Jharkhand as the reference category. Maternal age (years) and Body Mass Index (derived by scaling raw values by 100) were treated as continuous variables. Categorical covariates included educational attainment, stratified into higher education (reference group), secondary, primary, and no formal education. The household wealth index was segmented into five quintiles, with the middle quintile serving as the reference, compared against the poorer, poorest, richer, and richest strata. Finally, demographic adjustments included residential locality (rural vs. urban), social category or caste (Scheduled Caste, Scheduled Tribe, Other Backward Class, and General/Other), and religious affiliation (categorized as Hindu or others).

2.4 Survey-weighted Analysis

A complex survey design was defined using the R survey package's svydesign function, incorporating PSU-level clustering (ids = ~psu), stratification (strata = ~strata), and probability weights (weights = ~wt, nest = TRUE). Design-adjusted estimates and standard errors for anaemia prevalence by state were calculated using survey-weighted proportions (svymean). To account for the complex, two-stage stratified sampling design of the (NFHS-5), survey-weighted analytical adjustments must be integrated into the multivariable logistic regression framework to ensure accurate and nationally representative population estimates[4].

2.5 Multivariable Logistic Regression

A survey-weighted multivariable logistic regression was performed using svyglm with a quasibinomial family, where anaemia served as the outcome variable and all covariates were included as predictors, with Jharkhand set as the reference state. Adjusted Odds Ratios (AOR) were obtained by exponentiating the model coefficients; 95% CI were calculated as exp ($\beta \pm 1.96 \times SE$). The model summary included the P-values. Multicollinearity was calculated by means of the Generalised Variance Inflation Factor (GVIF) from the car package; a GVIF value above 10 indicates possible multicollinearity concerns. Socioeconomic and demographic risk factors are evaluated using multivariable logistic regression, an established approach for isolating independent determinants and calculating adjusted odds ratios from complex survey data[5].

2.6 Random Forest Classification

Separate Random Forest models tailored to each stateusing the randomForest package with 500 trees and mtry = 2were trained for Odisha and Jharkhand, following stratification of the cleaned dataset (with na.omit applied). Feature importance was measured using Mean Decrease Gini, which calculates the overall decrease in node impurity across all trees caused by each variable. An additional pooled model was trained using a 70:30 traintest split (set.seed(123)) to evaluate overall model accuracy and AUC (using the pROC package). To complement traditional econometrics, a Random Forest machine learning framework is utilized to rank the relative importance of predictors and capture complex interactions among health determinants[6].

2.7 Concentration Index and Concentration Curve

The inequality in anaemia related to wealth was measured using the Concentration Index (CI), calculated through the standard covariance formula. The wealth index quintile, ranked from poorest (1) to richest (5), was used as the socioeconomic ranking variable. Erreygers' Normalized CI (ECI = 4 × mu × CI) was calculated to adjust for the bounded binary outcome. The Concentration Curve displays the cumulative share of anaemia against the cumulative share of the population ranked by wealth from poorest to richest; when the curve lies above the line of equality, it signals pro-poor concentration (CI < 0). The standard concentration index is widely used to compute the degree of

socioeconomic-related health inequity by plotting a concentration curve against a diagonal line of perfect equality[7].

3. Result

3.1 Sample Characteristic

The pooled analytical sample comprised 54,466 women aged 15-49 years, including 27,971 respondents from Odisha and 26,495 from Jharkhand. Sociodemographic features of the study population are presented in Table 1.

The mean age of respondents was 30.25 years, ranging from 15 to 49 years. Body Mass Index (BMI) values ranged from 12.17 to 59.97 kg/m², with a mean of 21.39 kg/m² and a median of 20.75 kg/m². The interquartile range (IQR) was 18.59-23.37 kg/m². A total of 1,343 BMI observations (2.5%) were missing and excluded from regression and machine learning analyses using complete-case (na.omit) procedures.

The socioeconomic profile indicated substantial economic deprivation among respondents. About 42.6% of women belonged to the poorest wealth quintile, while 24.6% were from the poorer category, together accounting for 67.2% of the pooled sample. The study population included women from diverse educational, residential, caste, and religious backgrounds, enabling comprehensive assessment of socioeconomic and nutritional determinants of anaemia.

Table 1. Sociodemographic Characteristics of the Study Sample (N = 54,466)

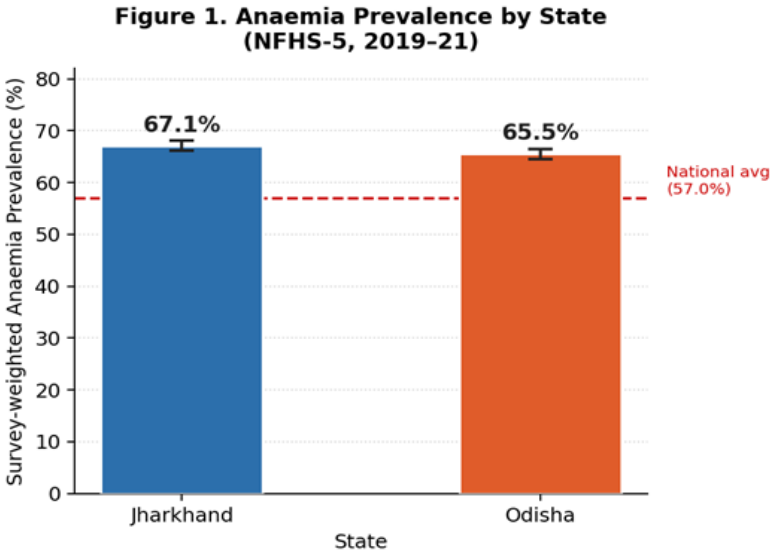
Variable	Category	N	% or Mean (SD) / Median (IQR)
State	Odisha	27,971	51.4%
	Jharkhand	26,495	48.6%
Age (years)	Mean (SD)	54,466	30.25 (9.8)
BMI (kg/m²)	Mean (SD)	53,123	21.39 (4.8)
	Median (IQR)	53,123	20.75 (18.59–23.37)
Wealth Index	Poorest	23,191	42.6%
	Poorer	13,405	24.6%
	Middle	8,650	15.9%
	Richer	5,654	10.4%
	Richest	3,566	6.5%
Anaemia	Yes	36,199	66.5%
	No	18,267	33.5%

3.2 Anaemia Prevalence by State

Table 2 and Figure 1 show the anaemia prevalence across states. The survey-weighted prevalence of anaemia was 67.1% (SE: 0.51%) in Jharkhand and 65.5% (SE: 0.46%) in Odisha. Initial estimates indicated 17,907 anaemic women in Jharkhand (67.6%) and 18,292 in Odisha (65.4%). Univariable survey-weighted logistic regression (with Jharkhand as reference) produced an odds ratio of 0.93 (95% CI: 0.880.99; p = 0.023), suggesting Odisha had slightly lower odds of anaemia. Both estimates are significantly higher than the national NFHS-5 average of 57.0%.

Table 2. State-wise Anaemia Prevalence (NFHS-5, 2019–21)

State	Total n	Anaemic n	Crude %	Survey-wtd %	SE
Jharkhand	26,495	17,907	67.6%	67.1%	0.51%
Odisha	27,971	18,292	65.4%	65.5%	0.46%
Total	54,466	36,199	66.5%	66.5%	—



Error bars represent 95% confidence intervals (±1.96 × SE). Dashed red line = national average (57.0%, NFHS-5). Source: Authors' analysis of NFHS-5 microdata.

Figure 1. Survey-Weighted Anaemia Prevalence by State (NFHS-5, 2019-21)

3.3 Assessment of Multicollinearity

Before conducting multivariable modeling, multicollinearity was evaluated using the Generalised Variance Inflation Factor (GVIF) within a linear model. All GVIF values were well under the 10 threshold (Table 3), confirming that multicollinearity among the predictors was not an issue.

Table 3. Variance Inflation Factor (VIF) Results

Variable	GVIF	df	GVIF^(1/2Df)
Age	1.433	1	1.197
BMI	1.224	1	1.106
Education Level	1.787	3	1.102
Wealth Index	1.905	4	1.084
Residence	1.395	1	1.166

3.4 Logistic Regression with Multiple Variables

Table 4 and Figure 2 display the outcomes of the survey-weighted multivariable logistic regression, revealing that initial, raw differences in anaemia odds between Odisha and Jharkhand disappear after controlling for socioeconomic and nutritional covariates (AOR=1.00; 95% CI: 0.95-1.07; p=0.879). This indicates that geographic variation is entirely driven by underlying regional disparities rather than local baseline differences. Physiological markers show that Body Mass Index (BMI) acts as a crucial protective factor, with each unit increase corresponding to a 3% decrease in anaemia threat (AOR=0.97; 95% CI: 0.96-0.98; $p < 0.001$), reinforcing the premise that overall nutritional sufficiency mitigates vulnerability. Socioeconomic factors exhibit a strong, inverse gradient; compared to highly educated women, those with no education (AOR=1.20; $p < 0.001$), primary schooling (AOR=1.16; p=0.003), and secondary schooling (AOR=1.15; $p < 0.001$) experience significantly elevated risk, potentially due to differences in dietary health literacy. Similarly, household wealth follows a stark pro-poor pattern, where women in the poorest quintile face 30% greater odds (AOR=1.30; 95% CI: 1.21-1.40; $p < 0.001$) and the poorer quintile faces 10% greater odds (AOR=1.10; 95% CI: 1.02-1.18; p=0.010) relative to the middle quintile, though this risk plateaus without significant variance in the richer and richest groups. Conversely, age is not a

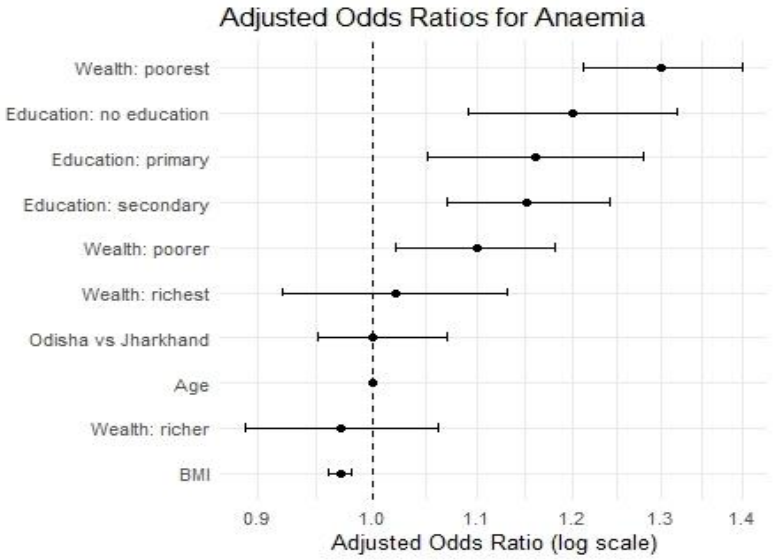
significant predictor in this adjusted model (AOR=1.00; p=0.385), demonstrating that the impact of age on anaemia is fully mediated by the proximate socioeconomic and physiological determinants captured in the analysis. Ultimately, these findings underscore that anaemia in these states is not a random geographic burden, but rather a structural issue deeply rooted in systemic deprivation. The strong, intersecting patterns of low wealth and limited education emphasize that clinical interventions alone will be insufficient without addressing broader social determinants of health. Consequently, public health strategies must move beyond uniform iron supplementation toward targeted, multi-sectoral policies that improve female literacy, enhance economic security, and promote overall nutritional welfare.

Table 4. Survey Weighted Multivariate Regression Model

	AOR	95% CI Lower	95% CI Upper	p-value	Sig.
State (ref: Jharkhand)					
Odisha	1.00	0.95	1.07	0.879	ns
Continuous					
Age (per year)	1.00	1.00	1.00	0.385	ns
BMI (per unit, kg/m2)	0.97	0.96	0.98	<0.001	***
Education (ref: Higher)					
No education	1.20	1.09	1.32	<0.001	***
Primary	1.16	1.05	1.28	0.003	**
Secondary	1.15	1.07	1.24	<0.001	***
Wealth Index (ref: Middle)					
Poorest	1.30	1.21	1.40	<0.001	***
Poorer	1.10	1.02	1.18	0.010	*
Richer	0.97	0.89	1.06	0.498	ns
Richest	1.02	0.92	1.13	0.724	ns

AOR=Adjusted Odds Ratio; CI=Confidence Interval; analysis uses a survey-weighted model (svyglm, quasibinomial, R survey package) with reference groups set to Jharkhand (State), Higher

*(Education), and Middle (Wealth Index), where significance levels are denoted by *** p < 0.001, p < 0.01, * p < 0.05, and ns = not significant.*

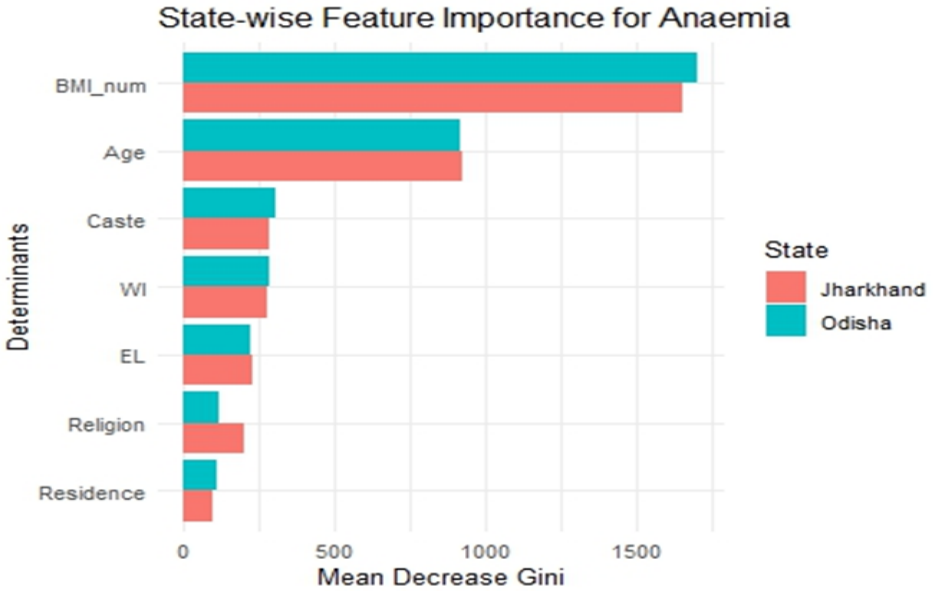


Horizontal lines = 95% CI. Dashed vertical line = OR of 1.0 (null). Log scale on x-axis. Reference: Jharkhand; higher education; middle wealth.

Figure 2. Forest Plot: Adjusted Odds Ratios for Anaemia (Survey-weighted Multivariable Logistic Regression)

3.5 Feature Importance in Random Forest

State-specific Random Forest models were developed for Odisha and Jharkhand, identifying BMI and age as the primary drivers of anemia, with caste, wealth index, and education as intermediate factors. Religion and residence had minimal influence, while the state variable showed the least importance, correlating with the non-significant adjusted odds ratio from logistic regression. Validated through a 70:30 train-test splitting, the model accomplished a testing accuracy of 66.3% but displayed significant class imbalance, particularly with a 98.9% error rate in the 'No' category, reflecting the high baseline prevalence of anemia at 66.5%. Although the classification performance was poor, the analysis fulfilled its methodological aim of robust feature extraction and variable importance ranking, successfully highlighting the sociodemographic and health determinants of anemia in both states.



Mean Decrease Gini quantifies each variable's total contribution to node purity improvement across all 500 trees. Higher values

indicate greater importance. Models trained separately for Odisha and Jharkhand (randomForest package, ntree = 500)

Figure 3. State-wise Random Forest Feature Importance (Mean Decrease Gini)

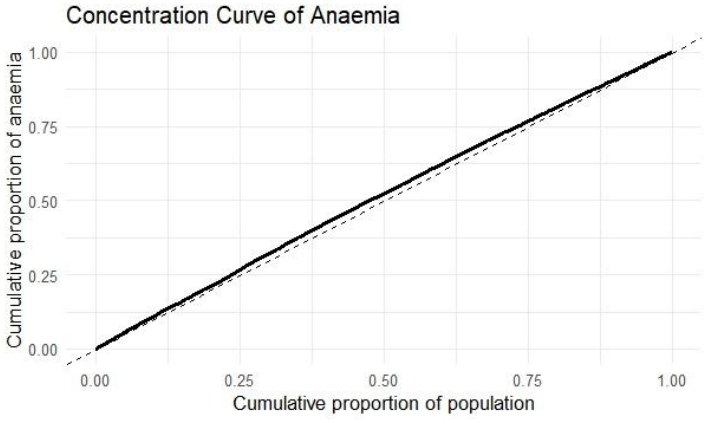
3.6 Concentration Index and Concentration Curve

Table 5 presents the Concentration Index results. The CI was -0.065 (Erreygers' Normalised CI= -0.177), confirming statistically significant pro-poor concentration of anaemia. The negative CI indicates that anaemia is disproportionately borne by women in lower wealth quintiles. The mean anaemia prevalence used in ECI computation was 67.8% (weighted sample). Figure 4 presents the Concentration Curve. The curve lies entirely above the line of equality throughout its range, confirming that anaemia is concentrated among the poor. The poorest 42.6% of women account for approximately 48% of all anaemia cases well above their proportional share.

Table 5. Concentration Index Results for Anaemia by Wealth Quintile

Wealth Quintile	Total n	Anaemia n*	Prevalence	Cumulative Pop. Share
Poorest	23,191	17,393	75.0%	0–42.6%
Poorer	13,405	9,316	69.5%	42.6–67.2%
Middle	8,650	5,493	63.5%	67.2–83.1%
Richer	5,654	3,099	54.8%	83.1–93.5%
Richest	3,566	1,648	46.2%	93.5–100%
Concentration Index	−0.065			Pro-poor concentration
Erreygers' ECI	−0.177			Mean prevalence: 67.8%

Anaemia n estimated as quintile N × quintile-specific prevalence. CI computed via standard covariance formula using survey weights. ECI = Erreygers' Normalised Concentration Index = 4 × μ × CI



The concentration curve lies above the line of equality, indicating that anaemia is disproportionately concentrated among poorer

women (CI = −0.065; ECI = −0.177). The x-axis shows the cumulative population share from poorest to richest, while the y-axis represents the cumulative share of anaemia cases. Source: Authors' analysis of NFHS-5 microdata.

Figure 4. NFHS-5 Anaemia Concentration Curve for Reproductive-Age Women in Odisha and Jharkhand

4. Discussion

This study provides a comprehensive assessment of anaemia among females of procreative age in Odisha and Jharkhand by integrating survey-weighted logistic regression, Random Forest machine learning, and socioeconomic inequality analysis. The findings revealed that anaemia prevalence remains alarmingly high in both Jharkhand (67.1%) and Odisha (65.5%), substantially exceeding the national NFHS-5 estimate of 57.0%. Furthermore, BMI, educational attainment, and household wealth emerged as significant determinants of anaemia, while socioeconomic inequality analyses demonstrated that the burden of anaemia is disproportionately concentrated among poorer women. A notable finding was that the initially observed difference in anaemia prevalence between Odisha and Jharkhand became statistically non-significant after adjusting for socioeconomic and nutritional factors. This suggests that variations in anaemia prevalence between the two states are largely explained by differences in population characteristics rather than state-specific contextual factors. The finding highlights the importance of addressing underlying social and nutritional vulnerabilities instead of relying solely on geographically targeted interventions.

BMI emerged as a significant protective factor against anaemia, with apiece unit rise in BMI linked with a 3% dropping in the odds of anaemia. In addition, BMI was identified as the most influential predictor in the Random Forest models for both states. These results are consistent with the findings reported by Ghosal and Gnanasekaran who also identified poor nutritional status as a major contributor to anaemia among Indian women. Women with lower BMI are more likely to experience inadequate dietary intake, micronutrient deficiencies, and impaired haemoglobin synthesis, thereby increasing their susceptibility to anaemia. The strong influence of BMI underscores the need for integrated nutrition interventions targeting undernourished women, particularly among tribal and economically disadvantaged populations.

Educational attainment demonstrated a significant inverse association with anaemia. Women with no proper schooling, prime education, and subordinate education had significantly higher odds of anaemia compared with women who attained higher education. Parallel outcomes have been reported by Let and Gnanasekaran, who observed that educational advancement is associated with improved nutritional awareness, healthcare utilization, dietary practices, and compliance with iron-folic acid supplementation programmes. These findings suggest that investments in female education may yield long-term benefits in falling the encumbrance of anaemia and improving overall maternal health outcomes.

The present study also revealed a pronounced socioeconomic gradient in anaemia prevalence. Women belonging to the underprivileged wealth quintile had 30% higher odds of anaemia compared with those in the middle wealth category. Furthermore, the negative Concentration Index (CI = -0.065) and Erreygers' Concentration Index (ECI = -0.177) confirmed a significant pro-poor concentration of anaemia. These conclusions are consistent with the work of Ghosal, who documented persistent socioeconomic inequalities in anaemia among tribal and non-tribal women across India. Similarly, Chakrabarti [8] reported that improvements in household socioeconomic conditions were important contributors to reductions in anaemia prevalence among Indian women. Economic deprivation often limits access to nutritious foods, healthcare services, sanitation facilities, and other health-promoting resources, thereby increasing vulnerability to anaemia and perpetuating health inequalities.

The Random Forest analysis provided complementary evidence supporting the regression findings. BMI, age, wealth index, and educational status consistently emerged as important predictors of

anaemia across both states. These findings are comparable to those reported by Jha, who identified socioeconomic and nutritional variables as key predictors of anaemia using machine learning techniques. Although the predictive performance of the model was modest (AUC = 0.554), the primary objective of the Random Forest analysis was variable importance ranking rather than clinical prediction. The relatively low classification accuracy for non-anaemic women is likely attributable to the high prevalence of anaemia and resulting class imbalance within the dataset. Future studies may improve predictive performance through the application of advanced class-balancing approaches such as Synthetic Minority Oversampling Technique (SMOTE), cost-sensitive learning, or ensemble boosting algorithms.

4.1 Strengths and Limitations

This study possesses several distinct methodological strengths, characterized by a large, nationally representative pooled sample (N = 54,466) that meticulously adjusts for a complex survey design across all stages of analysis. By triangulating evidence through a robust combination of multivariable logistic regression, machine learning (Random Forest), and formal inequality tools (Concentration Index analysis), the study offers a highly comprehensive evaluation of the data, further strengthened by the use of objective, hemoglobin-based outcomes and a VIF-confirmed absence of multicollinearity. Nevertheless, these insights must be balanced against inherent limitations; the cross-sectional nature of the data precludes any definitive causal inferences, and the application of standardized hemoglobin thresholds did not account for potential confounding variations in altitude or individual pregnancy status. Furthermore, potential biases may have been introduced through missing data, specifically the 4.0% of missing hemoglobin values and the 2.5% of missing BMI records excluded via listwise deletion (na.omit). Finally, the precision of the inequality estimations may have been constrained by utilizing categorical wealth quintiles rather than a continuous economic ranking, while the predictive performance (AUC) of the Random Forest model was inherently limited by structural class imbalance within the dataset.

5. Conclusion

Women of reproductive age in Odisha and Jharkhand face a disproportionately high burden of anaemia; their combined prevalence stands at 66.5%, a figure that tracking well above the national benchmark. The findings indicate that state of residence alone is not a significant determinant of anaemia after adjusting for nutritional and socioeconomic factors. Instead, low BMI, poor educational status, and economic deprivation emerged as the key modifiable risk factors. The negative Concentration Index values (CI = -0.065; ECI = -0.177) and the concentration curve further reveal that the burden of anaemia is disproportionately concentrated among poorer women. Random Forest analysis also highlighted the dominant role of BMI and wealth-related factors in predicting anaemia.

Reducing the burden of anaemia in these states requires a comprehensive and equity-focused strategy. Strengthening universal iron–folic acid supplementation, improving dietary diversity and food fortification among poor households, promoting girls' education, expanding deworming and WASH interventions, and providing nutrition counselling for underweight women are essential measures. In addition, culturally sensitive healthcare interventions for tribal communities are necessary to improve programme effectiveness. Overall, these findings provide important evidence for policymakers and programme planners working towards achieving the goals of Anaemia Mukt Bharat in Odisha and Jharkhand.

a) **Ethics Approval**

NFHS-5 data are publicly available and anonymized. Permission to use the data was obtained from the DHS Program. Ethical approval was not required for secondary analysis of anonymized data.

b) **Data Availability Statement** The data utilizing for this study can be accessed through the DHS Program, subject to reasonable request and official authorization.

c) **Conflict of Interest Statement** The authors state that they have no competing financial or personal interests that could influence this work.

d) **Funding Statement** This study was conducted without any outside financial support or external grants.

References

[1] J. Ghosal, M. Bal, M. Ranjit, and J. S. Kshatri, "To what extent classic socio-economic determinants explain trends of anaemia in tribal and non-tribal women of reproductive age in India? Findings from four National Family Health Surveys (1998–2021)," *BMC Public Health*, vol. 23, Art. no. 856, 2023, doi: 10.1186/s12889-023-15838-x.

[2] S. Let, S. Tiwari, A. Singh, S. Kumar, and S. Pedgaonkar, "Prevalence and determinants of anaemia among women of reproductive age in Aspirational Districts of India: An analysis of NFHS 4 and NFHS 5 data," *BMC Public Health*, vol. 24, Art. no. 437, 2024, doi: 10.1186/s12889-024-17789-3.

[3] S. Gnanasekaran, M. K. Gupta, V. Jayaraj, G. Mohan, and V. Rajendran, "Distribution and determinants of anemia among reproductive age group women in India: An insight from the fifth round of the National Family Health Survey (2019–21)," *Indian Journal of Community Medicine*, vol. 51, no. 2, pp. 305–313, 2026, doi: 10.4103/ijcm.ijcm_429_24.

[4] M. Das, M. Verma, P. Barman, and D. K. Behera, "Prevalence of anaemia among married women with recent birth history and high-risk fertility behaviour: Secondary data analysis of the National Family Health Survey-India (2019–21)," *BMJ Open*, vol. 14, no. 1, Art. no. e073395, 2024, doi: 10.1136/bmjopen-2023-073395.

[5] P. Bharati, M. Pal, S. Chakraborty, and S. Bhakta, "Socio-economic and demographic risk factors of anemia among Indian children using NFHS data," *AIP Conf. Proc.*, vol. 2163, no. 1, Art. no. 020021, 2019, doi: 10.1063/1.5174115.

[6] R. K. Jha *et al.*, "Socio-demographic factors associated with anaemia among non-pregnant and non-lactating women from low-income families in India: A random forest analysis," *International Journal of Community Medicine and Public Health*, vol. 9, no. 11, pp. 4130–4138, 2022, doi: 10.18203/2394-6040.ijcmph20222934.

[7] O. O'Donnell, S. O'Neill, T. Van Ourti, and B. Walsh, "Conindex: Estimation of concentration indices," *The Stata Journal*, vol. 16, no. 1, pp. 112–138, 2016, doi: 10.1177/1536867X1601600112.

[8] S. Chakrabarti, N. George, M. Majumder, N. Raykar, and S. Scott, "Identifying sociodemographic, programmatic and dietary drivers of anaemia reduction in pregnant Indian women over 10 years," *Public Health Nutrition*, vol. 21, no. 13, pp. 2424–2433, 2018, doi: 10.1017/S1368980018000903.

[9] C. Anjanamma, K. A. Jyotsna, B. Sravani, K. Shilpa, C. V. L. Narayana, and K. S., "Predicting childhood anemia prevalence with machine learning: Evidence from global nutritional data," in *Proc. 6th Int. Conf. Inventive Research in Computing Applications (ICIRCA)*, 2025, pp. 1459–1463, doi: 10.1109/ICIRCA65293.2025.11089631.

[10] S. B. Askandar, *Integration of Artificial Intelligence Algorithms with Automated Hematology Analyzers to Enhance Differential Diagnosis of Anemia Subtypes* [Data set]. Zenodo, 2025. doi: 10.5281/zenodo.17492692.

[11] E. C., V. Sathya, G. S. Priyatharsini, A. K. V., I. Vasudevan, and M. Umapathy, "Machine learning models for predicting anemia: Evaluation and performance insights," in *Proc. First Int. Conf. Innovations in Communications, Electrical and Computer Engineering (ICICEC)*, 2024, pp. 1–7, doi: 10.1109/ICICEC62498.2024.10808776.

[12] S. C., A. M. R., M. D. D., and D. M., "Curability prediction model for anemia using machine learning," in *Proc. 8th Int. Conf. Smart Structures and Systems (ICSSS)*, 2022, pp. 1–7, doi: 10.1109/ICSSS54381.2022.9782233.

[13] P. T. Dalvi and M. A. Gawas, "A comprehensive review of machine learning approaches for detecting anemia and abnormal red blood cells using RBC indices and medical imaging," *Discovery Artificial Intelligence*, vol. 6, Art. no. 66, 2026, doi: 10.1007/s44163-025-00698-8.

[14] S. Gurudatha and R. Majhi, "Robust machine learning methods for prediction of childhood anemia – A case of the Empowered Action Group States of India," in *Proc. 2nd Int. Conf. Recent Advances in Information Technology for Sustainable Development (ICRAIS)*, 2024, pp. 188–

193, doi: 10.1109/ICRAIS62903.2024.10811745.
[15] M. K. Hirok, S. Rahman, and M. Parvin, "Anemia prediction and classification of all classes with and without anemia patients using a machine learning model," in *Proc. IEEE Int. Conf. Computing, Applications and Systems (COMPAS)*, 2024, pp. 1–6, doi: 10.1109/COMPAS60761.2024.10796730.
[16] A. J. Kario and R. Kurniawan, "Prediction of anemia using machine learning algorithms: Scoping review," *Media Publikasi Promosi Kesehatan Indonesia (MPPKI)*, vol. 7, no. 11, pp. 2616–2623, 2024, doi: 10.56338/mppki.v7i11.6289.
[17] S. Mandal, J. Behera, B. Shit, and S. Paul, "Regional disparity of anemia among children and its determinants: A study of National Family Health Survey-5," *Indian Journal of Public Health*, vol. 68, no. 3, pp. 450–453, 2024, doi: 10.4103/ijph.ijph_943_23.
[18] A. C. Mathew *et al.*, "Prevalence and determinants of anemia among school going children in the state of Tamil Nadu, India: Applications of two-level logistic regression model," *International Journal of Community Medicine and Public Health*, vol. 10, no. 12, pp. 4743–4750, 2023, doi: 10.18203/2394-6040.ijcmph20233773.
[19] A. J. Meitei, A. Saini, B. B. Mohapatra, and S. Das, "Predicting child anaemia in the North-Eastern states of India: A machine learning approach," *International Journal of System Assurance Engineering and Management*, vol. 13, pp. 2949–2962, 2022, doi: 10.1007/s13198-022-01765-4.
[20] V. Preethi, V. Hemalatha, N. Arlappa, and K. V. Radhakrishna, "Trends and predictors of severe and moderate anaemia among children aged 6–59 months in India: An analysis of three rounds of National Family Health Survey (NFHS) data," *BMC Public Health*, vol. 24, Art. no. 2824, 2024, doi: 10.1186/s12889-024-20328-9.
[21] L. Upadhye and S. P. Ram, "Application of machine learning algorithm in identification of anaemia diseases," in *Computational Intelligence and Data Analytics*, R. Buyya, S. M. Hernandez, R. M. R. Kovvur, and T. H. Sarma, Eds. Lecture Notes on Data Engineering and Communications Technologies, vol. 142. Singapore: Springer, 2023, pp. 85–96, doi: 10.1007/978-981-19-3391-2_8.
[22] K. Yadav *et al.*, "Prevalence and determinants of anemia due to micronutrient deficiencies among children aged 12–59 months in India—Evidence from Comprehensive National Nutrition Survey, 2016–18," *PLOS Global Public Health*, vol. 4, no. 1, Art. no. e0002095, 2024, doi: 10.1371/journal.pgph.0002095.