Supplementary document for "Spatial heterogeneity and co-occurrence of childhood undernutrition in India: a joint analysis of stunting, anaemia and low birthweight"

This section provides additional technical details of the trivariate copula geoadditive modelling framework used in the main analysis. The formulation follows the approach proposed by Filippou et al. (2019), extended to jointly model three binary outcomes and their spatially varying dependence structure.

Comparison of excluded and included observations

Of the 232,920 children in the NFHS-5 children's recode with non-missing anthropometric and biomarker data collection eligibility, 104,021 had complete information on all outcomes and covariates used in the main analysis and were retained in the analytic (complete-case) sample; the remaining 128,899 were excluded due to missing data on one or more anthropometric indicators or covariates. Supplementary Table S1 compares the demographic profile of excluded and included children on key characteristics.

Excluded children differed from included children on several characteristics, most notably household wealth and maternal education: excluded children were more likely to come from the poorest wealth quintile (31.1% vs. 22.4%) and to have mothers with no formal education (24.8% vs. 18.5%), and were somewhat more likely to reside in rural areas (81.4% vs. 77.7%). These differences were all statistically significant, reflecting the very large sample sizes involved, but corresponded to small effect sizes throughout (Cramer's V ranging from 0.04 to 0.12), comparable in magnitude to the co-occurrence associations reported in the main text. Because lower household wealth and lower maternal education are generally associated with a higher, not lower, risk of childhood undernutrition, this pattern of exclusion is more consistent with the analytic sample's prevalence estimates being conservative (i.e., a modest underestimate of the true population burden) than with the introduction of spurious associations among the three outcomes studied. We do not consider this pattern to materially threaten the study's central focus on the relative co-occurrence and spatial distribution of stunting, anaemia, and low birthweight, since these depend on associations between outcomes and covariates rather than on unconditional prevalence levels. Nonetheless, we note this as a limitation of the complete-case approach, and future work could explore multiple imputation or inverse-probability weighting for missingness to formally adjust for this differential pattern of exclusion.

Sensitivity analysis: measured versus recalled birthweight

Within the analytic sample, 63,933 children (61.5%) had birthweight recorded from a written health card, while 40,087 (38.5%) relied on maternal recall. Recalled birthweights showed greater digit heaping at round values (68.3% of recalled weights fell at multiples of 0.5 kg, compared with 59.0%

TABLE S1: Comparison of children excluded from the analytic sample (n = 128,899) versus children retained in the analytic sample (n = 104,020–104,021, depending on missingness in the specific characteristic shown) on key demographic characteristics. Excluded n's/percentages, Included n's/percentages, and pairwise test statistics comparing the two groups are shown.

Characteristic	Excluded, n (%)	Included, n (%)	Test statistic
Child sex			χ^2=352.4, p<0.001, V=0.039
Male	64,526 (50.1%)	56,139 (54.0%)	
Female	64,374 (49.9%)	47,881 (46.0%)	
Residence			χ^2=468.2, p<0.001, V=0.045
Urban	24,033 (18.6%)	23,166 (22.3%)	
Rural	104,867 (81.4%)	80,854 (77.7%)	
Wealth quintile			χ^2=3122.8, p<0.001, V=0.116
Poorest	40,120 (31.1%)	23,286 (22.4%)	
Poorer	30,962 (24.0%)	23,501 (22.6%)	
Middle	23,788 (18.5%)	21,295 (20.5%)	
Richer	19,537 (15.2%)	19,557 (18.8%)	
Richest	14,493 (11.2%)	16,381 (15.7%)	
Caste			χ^2=451.6, p<0.001, V=0.045
Scheduled caste	26,243 (22.5%)	21,605 (20.8%)	
Scheduled tribe	25,836 (22.2%)	21,282 (20.5%)	
Other backward classes	46,853 (40.2%)	42,240 (40.6%)	
Others	17,680 (15.2%)	18,893 (18.2%)	
Maternal education			χ^2=1981.4, p<0.001, V=0.092
No education	31,944 (24.8%)	19,266 (18.5%)	
Primary	17,696 (13.7%)	12,385 (11.9%)	
Secondary	63,849 (49.5%)	56,015 (53.9%)	
Higher	15,411 (12.0%)	16,354 (15.7%)	
Maternal age (years), Mean (SD)	26.94 (4.99)	27.78 (5.05)	t=40.3, p<0.001

of card-recorded weights), and low birthweight prevalence was modestly higher among children with recalled birthweight (17.4%) than among those with card-recorded birthweight (15.3%; χ^2 = 80.54, p<0.001). Supplementary Table S2 compares key unadjusted odds ratios for low birthweight in the full analytic sample against those obtained after restricting to children with card-recorded (measured) birthweight only.

TABLE S2: Sensitivity of key covariate associations with low birthweight to restricting the analytic sample to children with card-recorded (measured) birthweight only. Odds ratios (95% confidence intervals) are from separate unadjusted logistic regressions.

Covariate	**Full sample (n = 104,020)**	**Measured-only (n = 63,933)**
Wealth (Ref: poorest)		
Poorer	0.90 (0.86-0.95)	0.90 (0.84-0.96)
Middle	0.80 (0.76-0.84)	0.80 (0.75-0.85)
Richer	0.73 (0.69-0.77)	0.71 (0.66-0.76)
Richest	0.66 (0.63-0.70)	0.66 (0.62-0.71)
Maternal education (Ref: no education)		
Primary	1.06 (1.00-1.13)	1.04 (0.96-1.12)
Secondary	0.89 (0.86-0.93)	0.88 (0.83-0.93)
Higher secondary	0.66 (0.63-0.70)	0.65 (0.60-0.70)
Perceived size at birth (Ref: very large)		
Larger than average	0.99 (0.91-1.07)	0.99 (0.90-1.09)
Average	0.79 (0.74-0.85)	0.80 (0.74-0.87)
Smaller than average	4.90 (4.54-5.29)	4.48 (4.08-4.93)
Very small	12.96 (11.63-14.45)	9.79 (8.54-11.22)

1 Trivariate copula geoadditive model

Let Y_{i1}, Y_{i2} and Y_{i3} denote the stunting, anemia and low birthweight status of the i^{th} child such that for disease j, $Y_{ij} = 1$ if the i^{th} child has the disease and 0 otherwise ($j = 1, 2, 3$). Here n denotes the size of the analysis-ready data ($n = 1{,}04{,}021$). In a trivariate copula framework, we represent the joint distribution function of the response vector (Y_1, Y_2, Y_3) as a copula function of the corresponding univariate distributions and the correlation matrix of the responses as follows

$$F_{Y_1,Y_2,Y_3}(y_{i1}, y_{i2}, y_{i3}|\boldsymbol{x}_{i1}, \boldsymbol{x}_{i2}, \boldsymbol{x}_{i3}) = C(F_1(y_{i1}|\boldsymbol{x}_{i1}), F_2(y_{i2}|\boldsymbol{x}_{i2}), F_3(y_{i3}|\boldsymbol{x}_{i3}); \Sigma(\vartheta))$$

Here F_{Y_1,Y_2,Y_3} is the trivariate cumulative distribution function of the response vector, $F_j(y_{ij}|\boldsymbol{x}_{ij}) = P(Y_{ij} \leq y_{ij}|\boldsymbol{x}_{ij})$ is the marginal cumulative distribution function of $Y_j, j = 1, 2, 3$ while $\boldsymbol{x}_{ij}$ denotes the covariate vector for Y_j. This formulation allows the joint distribution of the three outcomes to be decomposed into marginal probabilities and a dependence structure capturing their associations. As in (1), we employ a trivariate Gaussian copula function $C : [0,1]^3 \to [0,1]$ with dependence structure characterized by the 3×3 correlation matrix $\Sigma(\vartheta)$ given by

$$\Sigma(\vartheta) = \begin{pmatrix} 1 & \vartheta_{12} & \vartheta_{13} \\ \vartheta_{21} & 1 & \vartheta_{23} \\ \vartheta_{31} & \vartheta_{32} & 1 \end{pmatrix}$$

where ϑ_{jk} corresponds to the correlation coefficient between the j^{th} and k^{th} diseases where $j, k = 1, 2, 3$ with $j \neq k$. For the specific case of three binary responses as in ours, the copula specification can be re-written as

$$P(Y_{i1} = 1, Y_{i2} = 1, Y_{i3} = 1|\boldsymbol{x}_{i1}, \boldsymbol{x}_{i2}, \boldsymbol{x}_{i3}) = C(P(Y_{i1} = 1|\boldsymbol{x}_{i1}), P(Y_{i2} = 1|\boldsymbol{x}_{i2}), P(Y_{i3} = 1|\boldsymbol{x}_{i3}); \Sigma(\vartheta))$$

Using the latent variable representation of binary models, we can specify $P(Y_{ij} = 1|\boldsymbol{x}_{ij})$ as

$$P(Y_{ij} = 1|\boldsymbol{x}_{ij}) = P(Y_{ij}^* > 0|\boldsymbol{x}_{ij}) = 1 - F_j(-\eta_{ij}), \quad j = 1, 2, 3$$

where Y_{ij}^* is a continuous latent variable such that $Y_{ij}^* = \eta_{ij} + \epsilon_{ij}$ where η_{ij} is a additive predictor relating the marginal disease prevalence to the corresponding covariates and spatial effects. The flexibility of the copula approach lies in the ability to specify a wide class of parametric distributions for the copula distribution as well as the marginal response distribution. In the current study, we use a trivariate Gaussian copula which is given by

$$\Phi_3(\Phi^{-1}(F_1(\eta_{i1})), \Phi^{-1}(F_2(\eta_{i2})), \Phi^{-1}(F_3(\eta_{i3})))$$

where Φ^{-1} is the quantile function of the standard normal distribution while $F_j(\eta_{ij})$ is the cumulative distribution function of an univariate logistic distribution. This leads to the logistic model specification for the marginal models for each disease indicator i.e

$$F_j(\eta_{ij}) = \frac{\exp(\eta_{ij})}{1 + \exp(\eta_{ij})}, \quad j = 1, 2, 3$$

We use the logistic link due to the simplicity of interpretation of the effect of the covariates on the odds of the disease probabilities. The general form of the additive predictor η_{ij} is as follows

$$\eta_{ij} = \boldsymbol{Z}'_{ij}\boldsymbol{\beta} + \sum_{k=1}^{K} s_k(\boldsymbol{U}_{kij}), \quad i = 1, 2, ..., n; j = 1, 2, 3$$

where $\boldsymbol{Z}_{ij}$ is the vector of binary and categorical covariates corresponding to the i^{th} child and j^{th} disease while $\boldsymbol{U}_{kij}$ constitute the corresponding set of continuous covariates. The continuous covariates are modeled using smooth functions $s_k(.)$ parameterized using thin plate regression splines (2; 3). For the purpose of our study, we incorporated four distinct predictor effects in the marginal model specification, namely linear effects for the binary and categorical covariates, splines for continuous covariates, structured spatial effects to model the similarity in the disease-prevalence among children residing in adjacent regions while unstructured spatial effects were used to model the disease-correlation between children located in the same region. The marginal model corresponding to the i^{th} child and j^{th} disease $(i = 1, 2, ..., n; j = 1, 2, 3)$ is as follows

$$\begin{aligned}
\text{logit}(P(Y_{ij} = 1)) &= \beta_{0j} + \beta_{1j}\text{Water} + \beta_{2j}\text{Toilet} + \beta_{3j}\text{Prenataldoc} + \beta_{4j}\text{Shortstature} + \beta_{5j}\text{Caesarean} \\
&+ \beta_{6j}\text{Child.Gender} + \beta_{7j}\text{Rural} + \beta_{8j}\text{MUnderweight} + \beta_{9j}\text{MEdu1} + \beta_{10j}\text{MEdu2} \\
&+ \beta_{11j}\text{MEdu3} + \beta_{12j}\text{HWealth2} + \beta_{13j}\text{HWealth3} + \beta_{14j}\text{HWealth4} + \beta_{15j}\text{HWealth5} \\
&+ \beta_{16j}\text{Caste2} + \beta_{17j}\text{Caste3} + \beta_{18j}\text{Caste4} + \beta_{19j}\text{Birthsize2} + \beta_{20j}\text{Birthsize3} \\
&+ \beta_{21j}\text{Birthsize4} + \beta_{22j}\text{Birthsize5} + \beta_{23j}\text{MAnemia2} + \beta_{24j}\text{MAnemia3} \\
&+ \beta_{25j}\text{MAnemia4} + \beta_{26j}\text{Headsex} + s_{1j}(\text{Childage}) + s_{2j}(\text{Agefirstbirth}) \\
&+ s_{3j}(\text{Breastfed}) + s_{4j}(\text{State, structured}) + s_{5j}(\text{State, unstructured})
\end{aligned}$$

Here "MUnderweight", "MEdu", "MAnemia" and "HWealth" correspond to the categories of maternal underweight, maternal education, maternal anemia levels and household wealth quintiles respectively. β_{mj} $(m = 1, 2, ..., 26)$ are the effect parameters for the binary and categorical covariates, s_{kj} $(k = 1, 2, 3)$ are smooth effect-functions for the continuous covariates while s_{4j} and s_{5j} represent the structured and unstructured spatial effects respectively. Structured spatial effects were modeled using Markov random field smoothers (4). We further model the three correlation parameters $(\vartheta_{12}, \vartheta_{13}, \vartheta_{23})$ representing the pairwise dependence among the disease indicators with respect to the structured random effects to capture their spatial heterogeneity across state boundaries i.e

$$\vartheta_{jk} = s(\text{State, unstructured}), \qquad j, k = 1, 2, 3; \quad j \neq k.$$

Doing so enables us to study the spatial heterogeneity in the dependence structure of the responses across the states and union territories of India. Finally, model-based estimates were used to generate joint probability maps and spatial correlation maps, as presented in the main manuscript.

References

[1] Filippou P, Kneib T, Marra G, Radice R. A trivariate additive regression model with arbitrary link functions and varying correlation matrix. Journal of Statistical Planning and Inference. 2019;199:236-48.

[2] Eilers PH, Marx BD. Flexible smoothing with B-splines and penalties. Statistical science. 1996;11(2):89-121.

[3] Wood SN. Generalized additive models: an introduction with R. chapman and hall/CRC; 2017.

[4] Rue H, Held L. Gaussian Markov random fields: theory and applications. Chapman and Hall/CRC; 2005.