BMJ Paediatrics Open

Statistical modelling of determinants of child stunting using secondary data and Bayesian networks: a UKRI Global Challenges Research Fund (GCRF) Action Against Stunting Hub protocol paper

Todd S Rosenstock [1], Barbaros Yet [2]

To cite: Rosenstock TS, Yet B. Statistical modelling of determinants of child stunting using secondary data and Bayesian networks: a UKRI Global Challenges Research Fund (GCRF) Action Against Stunting Hub protocol paper. *BMJ Paediatrics Open* 2024;**8**:e001983. doi:10.1136/bmjpo-2023-001983

► Additional supplemental material is published online only. To view, please visit the journal online (https://doi.org/10.1136/bmjpo-2023-001983).

Received 26 May 2023
Accepted 11 January 2024

[1]Bioversity International, Montpellier, France
[2]Department of Cognitive Science, Graduate School of Informatics, Middle East Technical University, Ankara, Turkey

Correspondence to
Dr Todd S Rosenstock; t.rosenstock@cgiar.org

ABSTRACT

Introduction Several factors have been implicated in child stunting, but the precise determinants, mechanisms of action and causal pathways remain poorly understood. The objective of this study is to explore causal relationships between the various determinants of child stunting.

Methods and analysis The study will use data compiled from national health surveys in India, Indonesia and Senegal, and reviews of published evidence on determinants of child stunting. The data will be analysed using a causal Bayesian network (BN)—an approach suitable for modelling interdependent networks of causal relationships. The model's structure will be defined in a directed acyclic graph and illustrate causal relationship between the variables (determinants) and outcome (child stunting). Conditional probability distributions will be generated to show the strength of direct causality between variables and outcome. BN will provide evidence of the causal role of the various determinants of child stunning, identify evidence gaps and support in-depth interrogation of the evidence base. Furthermore, the method will support integration of expert opinion/assumptions, allowing for inclusion of the many factors implicated in child stunting. The development of the BN model and its outputs will represent an ideal opportunity for transdisciplinary research on the determinants of stunting.

Ethics and dissemination Not applicable/no human participants included.

WHAT IS ALREADY KNOWN ON THIS TOPIC

⇒ The risk factors for child stunting include concerns ranging from inadequate nutrition to epigenetics, environment, gut health, climate change and more.
⇒ Child stunting presents itself from the complex interactions of many risk factors together.
⇒ The importance of specific risk factors for child stunting is subject to local conditions complicating actions against stunting.

WHAT THIS STUDY ADDS

⇒ A probabilistic approach to modelling child stunting that reflects current understanding of the risk factors and how they interact.
⇒ Identification of key determinants of child stunting with examples of how they vary across locations with high prevalence.
⇒ Information on data gaps needed to reduce the uncertainty around the determinants of child stunting and their interaction.

HOW THIS STUDY MIGHT AFFECT RESEARCH, PRACTICE OR POLICY

⇒ The directed acyclic graph may illuminate hidden relationships and biases among interacting factors, spurring interdisciplinary investigations.
⇒ The results may inform programmatic interventions and policies on factors or combinations of factors to address child stunting.

INTRODUCTION

Childhood stunting is associated with myriad proximal and distal factors. Historically, investigations into the aetiology of stunting were focused primarily on dietary intake, with inadequate consumption of nutritious food and low dietary diversity identified as the primary risk factors.[1–3] However, improving nutrition through intervention studies had inconsistent effects on child growth.[4 5] Attention then shifted to the impact of infectious diseases, such as diarrhoea, on child nutritional status. Research studies into improving hygiene through provision of clean water, toilets and handwashing with soap also showed variable results, offering conflicting evidence of the role of water, sanitation and hygiene in improving childhood growth.[6 7] It now appears that inadequate diets and infectious diseases can interact to contribute to stunting. Exposure to enteropathogens

among people living in settings with poor sanitation and hygiene can modify gut structure and function and lead to reduced digestion and absorption of nutrients.[8] These effects may be amplified in some individuals due to the expression of epigenetic traits in response to environmental triggers.[9] In addition, underlying the aetiology of stunting are multitude of other factors, including socioeconomic status, childcare practices, parental education, malaria, air pollution, etc.[10–15]

Knowledge around the aetiology of child stunting is largely derived from observational studies through correlation of factors with measures of child growth (such as height-for-weight z-scores). Data alone cannot show causality. Drawing causal inferences requires both statistical inference and sound knowledge of the domain/subject area, otherwise the statistical outcomes can be biased[16] or even paradoxical.[17] The challenges of correlation studies are variables being included haphazardly in the model based on statistical associations without considering their causal influence on other factors. This is supported by an example from Hernán *et al* on the impact of folic acid supplementation on neural tube defects, where the authors suggest lowering the OR of the intervention by roughly 20%, from 0.80 to 0.65, taking into consideration a priori causal knowledge, as opposed to relying solely on statistical associations.[18] Hernán *et al* show that statistical strategies for variable selection and confounding recommend adjusting for a third variable representing whether the pregnancy ends in stillbirth or therapeutic abortion. However, a priori causal knowledge indicates that adjusting for this variable is likely to bias the causal effect estimates between folic acid supplementation and neural tube effects.

In addition, estimating causal effects requires the inclusion of confounding variables, which may be limited by availability of data, especially when using secondary data. Thus, it is not uncommon to see disregard of potentially important confounding variables in correlation models where there are no data, despite knowledge of the impact of the variable on the study's outcome. This procedure also leads to biased estimates due to unobserved confounders. Such data limitations are particularly challenging for research into child stunting, where very few or no datasets exist that include the multitude of determinants, and where some purported determinants are purely based on hypotheses and observation with only limited data.

An underlying causal model enables the determination of the causal relations that can be estimated and provides guidance in collecting additional information where data are lacking. Furthermore, recent studies of child stunting conducted in Bangladesh and India have adopted machine learning—random forests, gradient boosting and quantile regression—to learn from the data and discover further evidence of associations between stunting and various determinants.[19–21] However, these approaches also risk spurious predictions, especially when making similar predictions outside the study, because they do not exploit the causal understanding of their domain. Thus, there is a need to apply modelling frameworks that integrate causal knowledge and multiple types of data.[22]

Bayesian networks (BNs) by comparison offer an opportunity to build holistic causal models, based on the current understanding of the complex factors determining child stunting and their inter-relationship. Encoding a holistic model of causal relations is critical in assessing diverse intervention opportunities, to inform programmes and policies, and to identify evidence gaps in domain knowledge. Causal structure of these models can provide a basis for causal assumptions in future studies.

The aims of this study are to: (1) create a directed acyclic graph (DAG) summarising probable causal factors of child stunting at the whole child level, (2) parameterise BNs based on the DAG by using secondary data, and (3) assess the drivers of stunting in Indonesia, India and Senegal, and examine evidence gaps using BNs.

METHODS AND ANALYSIS

This study will develop a causal BN model using secondary data from national health surveys in India, Indonesia and Senegal to explore the determinants of child stunting at the whole child level. By using innovative methods to explore both the literature and secondary data on child stunting, this study will be able to capture and make explicit the interdisciplinary assumptions underpinning the different research workstreams of the UK Research and Innovation (UKRI) Global Challenges Research Fund (GCRF) Action Against Stunting Hub. The BN modelling process will consist of three inter-related and co-dependent steps: (1) develop a detailed causal structure of the whole child approach, (2) compile relevant secondary datasets, and (3) parameterise separate BNs for each country to bring together biological, social and environmental drivers of stunting in various contexts.

Model development

The model structure: the whole child DAG

The core of a BN is the DAG. As illustrated in figure 1, DAG is a graphical model illustrating the presumed linkages among risk factors, symptoms and outcomes. The interdisciplinary research knowledge at GCRF Action Against Stunting Hub will be used as a leverage to develop a DAG to show a causal model of child stunting using the whole child hypothesis. This method enables the identification of linkages among key drivers of stunting including: epigenetics, food, nutrition, cognition and learning environments. The DAG encoding the whole child hypothesis will be developed through mixed data and expert-driven approach.

Descriptions of factors determining child stunting will be compiled from published, peer-reviewed literature using Google Scholar and Web of Science. The

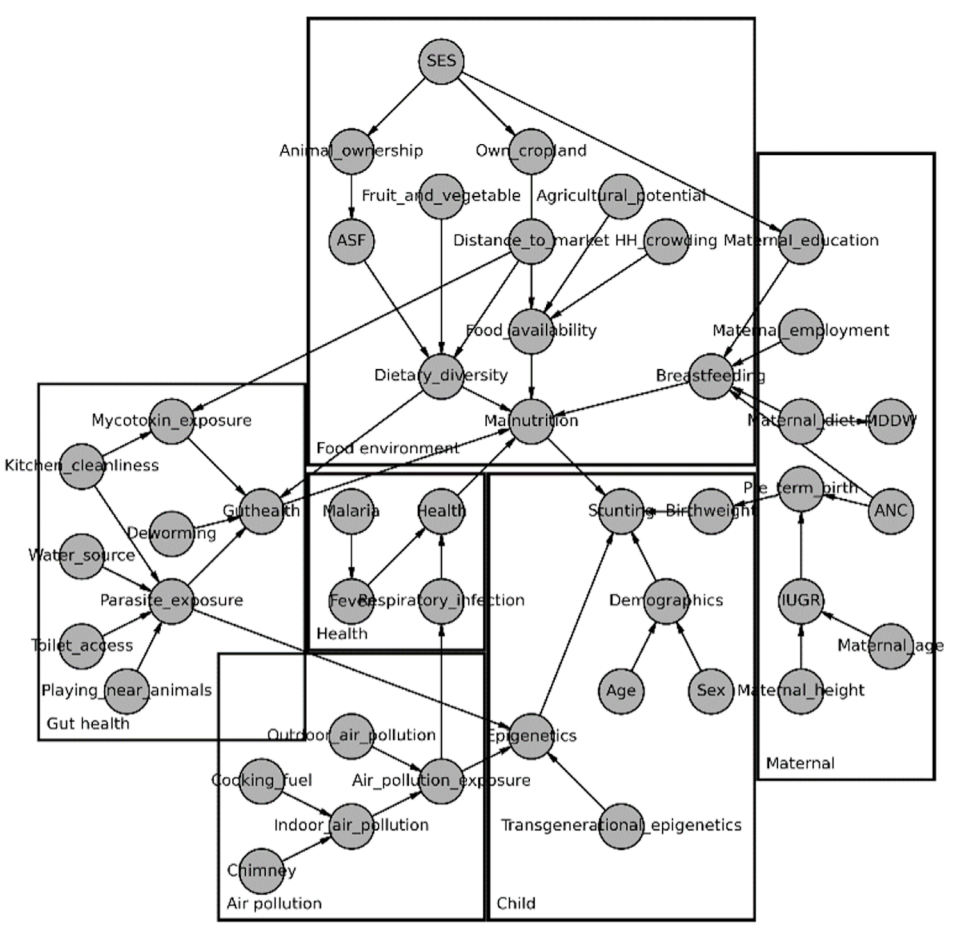

Figure 1 A depiction of directed acyclic graph for child stunting: visualising the whole child approach. ANC, antenatal care; ASF, animal source food; HH, household; IUGR, intrauterine growth restriction; MDDW, minimum dietary diversity for women; SES, socioeconomic status.

literature will be searched using combinations of the following keywords: child stunting and Boolean operators combining food environment, diets, nutrition, home environment, epigenetics, cognition and learning environment. Both primary research articles and reviews will be considered. Relationships described in the studies will be recorded in a Direct Edge Index (DEI).[23] DEI is a list which identifies the affecting factor, the affected factor and the type of relationship between the two. An example of DEI is depicted in table 1. Findings from studies will be added to the DEI until no new factors or relationships are uncovered. In contrast to Evidence Gap Maps, the DEI literature review is not intended to be systematic or exhaustive. Rather, the review simply aims to limit bias by ensuring that major factors identified in the literature inform the construction of the DAG.

Next, the factor labels used in the DEI will be standardised across the studies. For example, *height-for-age* and *length-for-age* are both measures interchanged across studies to indicate child stunting. Such repetitions will

Table 1 Examples of edges compiled in the Directed Edge Index based on the literature review

Study	Origin of edge	Termination of edge	Bidirectional
Prado *et al*[32]	Dietary diversity	Diarrhoea incidence	No
Mosites *et al*[33]	Gestational age at birth	Stunting	No
Onubi *et al*[34]	Probiotics	Increased weight	No
Beal *et al*[15]	Supplementary feeding	Weight gain	No
Galway *et al*[35]	Distance to nearest road	Fruit and vegetable consumption	No
Oh *et al*[36]	Micronutrient supplementation	Preterm birth	No
Webb-Girard *et al*[37]	Agricultural training	Nutrient-rich agriculture	No
Vilcins *et al*[38]	Dirt floor	Infection	No

be eliminated within the standardised list of factors and relationships, and very closely related factors will be synthesised. The resulting DEI will then become the reference point for developing the DAG. The description of supporting and conflicting evidence for each relation in the DEI will be documented in an evidence base.[24] In the DAG, the DEI's standardised factors become 'nodes' and the relationships between the factors described in the DEI become 'edges', drawn as lines between nodes indicating the direction of the effect (figure 1 and online supplemental table 1). We will begin the process of developing the DAG in this way to reduce participant bias in structuring the model.[25]

The DAG developed will then be presented to an interdisciplinary team of experts at the UKRI GCRF Hub and country experts for review along with the supporting evidence base. An interview will be conducted with a domain expert in each key driver of stunting including epigenetics, food, nutrition, cognition and learning environments. The results of DEI will be presented to each domain expert, and feedback will be collected as a relevant part of the DEI with respect to their area of expertise. Experts may suggest adding or removing variables and adding, removing or reversing edges during the interview. Suggestions that do not conflict with DEI or other experts will be incorporated in the DAG and recorded in the evidence base. When DEI and expert suggestions conflict, a second domain expert will review the suggestion and the majority of opinion will be reflected in the DAG. Note that the evidence base will document both conflicting and supporting evidence corresponding to the DAG.

This iterative engagement process makes explicit the team's abundant tacit expertise to further develop the DAG. The DAG will facilitate graphical presentation of evidence on the determinants of child stunting and their inter-relationship. The DAG also acts as a boundary object for cross-disciplinary work, the development of a common understanding across diverse teams, and communication with stakeholders and the scientific community. Similar research in different countries could in the future interrogate the DAG's structure to determine what changes in factors (nodes) or relationships (edges) are needed to capture the contextual circumstances affecting child stunting in each country. The model's underlying structure will be general for the determinants of child stunting but informed by the knowledge of experts in India, Indonesia and Senegal.

Secondary data: Demographic and Health Surveys in India, Indonesia and Senegal

The objective of the data and evidence compilation is to parameterise the BN model illustrated by DAG, which will be used to explore the drivers of child stunting. BN offers a suitable medium to identify the parts of a model with scarce data and to combine multiple types of evidence—raw data and meta-analyses—to learn model parameters.[26] Given that no single study to date has captured all the factors implicated in child stunting that could be identified in the DAG, this approach is crucial to reconcile data across studies, qualitative beliefs and 'gaps' where limited data are available.

Secondary data are the most immediately useful information to estimate the conditional probability distributions of the interactions shown by nodes and edges in the DAG. DAG-relevant data from the Demographic and Health Surveys (DHS), with a focus on India, Indonesia and Senegal, will be compiled. These datasets will include information on dietary intake, anthropometry, access to healthcare, stunting and the home environment. We expect to parameterise most of the model based on data from thousands of individual children across the three countries. In addition, ancillary data will be compiled from other publicly available datasets on agriculture and air pollution to support model parameterisation. In particular, we will match the closest spatial particulate matter 2.5 air pollution data and Global Agro-Ecological Zones (GAEZ) agricultural production and gap data for child's birth year with the household Global Positioning System coordinates in the DHS dataset.

Some nodes in the DAG may have missing data in the DHS and ancillary datasets, and some nodes may be unobserved or essentially unmeasurable. We will use BN abstraction operations and expectation–maximisation (EM) algorithm to deal with the unobserved variables and missing data. First, we will remove unobserved and unmeasurable variables with BN abstraction operations to create a simplified DAG that matches with the available data. BN abstraction operations are graphical operations on DAG that remove or merge variables without disturbing the independence assumptions in the rest of the DAG.[27–29] Since BN abstraction operations are algorithmic, the link between the original and simplified DAG remains clear. After this step, we will have a simplified DAG that has corresponding data from DHS, pollution and GAEZ datasets. The evidence base will highlight the evidence gap by documenting the link between the original and simplified DAG that is parameterised. Second, we use the EM algorithm to learn the conditional probability distributions of the simplified DAG. The EM algorithm learns BN parameters from missing data by iteratively estimating the expectations of the missing instances and computing the conditional probability distributions based on these expectations. An evidence base for the BN model will document the datasets and methods used to estimate the conditional probability distributions in different parts of the model. The evidence base will be available online for domain experts to browse and review the underlying evidence of the BN model.

Data analysis

This desk study will use two software programs to conduct the analysis. Descriptive statistics, preparatory analysis and parameter learning will occur in R.[30] Descriptive statistics will summarise information on the factors in the model derived from the national health surveys. Parameter

learning will define the conditional probability distribution parameters of each variable in BNs. The BN will be built in PyAgrum V.10.[31] PyAgrum is a Python library for BN modelling.

Patient and public partnership strategy

This study will use only secondary and published evidence with no direct involvement of patients and members of the public; thus, no patient and public consultations will be held. However, the study's concept was discussed with experts in the study countries (India, Indonesia and Senegal), and they will be involved in dissemination of the findings.

The findings of the study will be published in peer-reviewed journals. Results will also be disseminated through a blog post. All data, models and codes used will be available upon completion of the project in a public repository.

Contributors TSR conceived of the study. TSR and BY designed the study. TSR and BY drafted the protocol.

Funding The work was supported by UK Research and Innovation (UKRI) under its Global Challenges Research Fund (GCRF; reference: MR/S01313X/1).

Competing interests None declared.

Patient and public involvement Patients and/or the public were not involved in the design, or conduct, or reporting, or dissemination plans of this research.

Patient consent for publication Not applicable.

Ethics approval Not applicable/no human participants included.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement No data are available.

ORCID iDs
Todd S Rosenstock http://orcid.org/0000-0002-1958-9500
Barbaros Yet http://orcid.org/0000-0003-4058-2677

REFERENCES

1 Krasevec J, An X, Kumapley R, *et al*. Diet quality and risk of Stunting among infants and young children in Low- and middle-income countries. *Matern Child Nutr* 2017;13 Suppl 2(Suppl 2):e12430.
2 Afshin A, Sur PJ, Fay KA, *et al*. Health effects of dietary risks in 195 countries, 1990–2017: a systematic analysis for the global burden of disease study 2017. *The Lancet* 2019;393:1958–72.
3 Bhandari N, Bahl R, Taneja S. Effect of Micronutrient supplementation on linear growth of children. *Br J Nutr* 2001;85 Suppl 2:S131–7.
4 Pedraza DF, Rocha ACD, Sales MC. Micronutrient deficiencies and linear growth: a systematic review of observational studies. *Cien Saude Colet* 2013;18:3333–47.
5 Tam E, Keats EC, Rind F, *et al*. Micronutrient supplementation and Fortification interventions on health and development outcomes among children under-five in Low- and middle-income countries: A systematic review and meta-analysis. *Nutrients* 2020;12:289.
6 Humphrey JH, Mbuya MNN, Ntozini R, *et al*. Independent and combined effects of improved water, sanitation, and hygiene, and improved complementary feeding, on child Stunting and anaemia in rural Zimbabwe: a cluster-randomised trial. *Lancet Glob Health* 2019;7:e132–47.
7 Dangour AD, Watson L, Cumming O, *et al*. Interventions to improve water quality and supply, sanitation and hygiene practices, and their effects on the nutritional status of children. *Cochrane Database Syst Rev* 2013:CD009382.
8 Prendergast AJ, Kelly P. Interactions between intestinal pathogens, Enteropathy, and malnutrition in developing countries. *Curr Opin Infect Dis* 2016;29:229–36.
9 Indrio F, Martini S, Francavilla R, *et al*. Epigenetic matters: the link between early nutrition, Microbiome, and long-term health development. *Front Pediatr* 2017;5:178.
10 Sinharoy SS, Clasen T, Martorell R. Air pollution and Stunting: a missing link. *Lancet Glob Health* 2020;8:e472–5.
11 Christian P, Lee SE, Donahue Angel M, *et al*. Risk of childhood Undernutrition related to small-for-gestational age and Preterm birth in Low- and middle-income countries. *Int J Epidemiol* 2013;42:1340–55.
12 Danaei G, Andrews KG, Sudfeld CR, *et al*. Risk factors for childhood Stunting in 137 developing countries: A comparative risk assessment analysis at global, regional, and country levels. *PLoS Med* 2016;13:e1002164.
13 Jackson BD, Black RE. A literature review of the effect of malaria on Stunting. *J Nutr* 2017;147:2163S–2168S.
14 Wirth JP, Rohner F, Petry N, *et al*. Assessment of the WHO Stunting framework using Ethiopia as a case study. *Matern Child Nutr* 2017;13:e12310.
15 Beal T, Tumilowicz A, Sutrisna A, *et al*. A review of child Stunting determinants in Indonesia. *Matern Child Nutr* 2018;14:e12617.
16 Pearl J, Glymour M. *Causal Inference in Statistics: A Primer*. Wiley, 2016.
17 Hernán MA, Clayton D, Keiding N. The Simpson's paradox unraveled. *Int J Epidemiol* 2011;40:780–5.
18 Hernán MA, Hernández-Díaz S, Werler MM, *et al*. Causal knowledge as a prerequisite for confounding evaluation: an application to birth defects epidemiology. *Am J Epidemiol* 2002;155:176–84.
19 Mansur M, Afiaz A, Hossain MS. Sociodemographic risk factors of under-five Stunting in Bangladesh: assessing the role of interactions using a machine learning method. *PLoS ONE* 2021;16:e0256729.
20 Khan JR, Tomal JH, Raheem E. Model and variable selection using machine learning methods with applications to childhood Stunting in Bangladesh. *Informatics for Health and Social Care* 2021;46:425–42.
21 Fenske N, Burns J, Hothorn T, *et al*. Understanding child Stunting in India: A comprehensive analysis of socio-economic, nutritional and environmental determinants using additive Quantile regression. *PLoS ONE* 2013;8:e78692.
22 Pearl J. The seven tools of causal inference, with reflections on machine learning. *Commun ACM* 2019;62:54–60.
23 Puga JL, Krzywinski M, Altman N. Bayesian networks. *Nat Methods* 2015;12:799–800.
24 Yet B, Perkins ZB, Tai NRM, *et al*. Clinical evidence framework for Bayesian networks. *Knowl Inf Syst* 2017;50:117–43.
25 Ferguson KD, McCann M, Katikireddi SV, *et al*. Evidence synthesis for constructing directed Acyclic graphs (ESC-Dags): a novel and systematic method for building directed Acyclic graphs. *Int J Epidemiol* 2020;49:322–9.
26 Yet B, Perkins ZB, Rasmussen TE, *et al*. Combining data and meta-analysis to build Bayesian networks for clinical decision support. *J Biomed Inform* 2014;52:373–85.
27 Yet B, Marsh DWR. Compatible and incompatible abstractions in Bayesian networks. *Knowledge-Based Systems* 2014;62:84–97.
28 Shachter RD. Probabilistic inference and influence diagrams. *Operations Research* 1988;36:589–604.
29 Lauritzen SL. The EM algorithm for graphical Association models with missing data. *Computational Statistics & Data Analysis* 1995;19:191–201.
30 R Core Team. R: A language and environment for statistical computing Vienna, Austria [R Foundation for Statistical Computing]. 2021. Available: https://www.R-project.org
31 Ducamp G, Gonzales C, Wuillemin PH. aGrUM/pyAgrum: a Toolbox to build models and Algorithms for probabilistic graphical models in python. 2020. In International Conference on Probabilistic Graphical Models; PMLR,

32 Prado EL, Yakes Jimenez E, Vosti S, *et al*. Path analyses of risk factors for linear growth faltering in four prospective cohorts of young children in Ghana, Malawi and Burkina Faso. *BMJ Glob Health* 2019;4:e001155.
33 Mosites E, Dawson-Hahn E, Walson J, *et al*. Piecing together the Stunting puzzle: a framework for attributable factors of child Stunting. *Paediatr Int Child Health* 2017;37:158–65.
34 Onubi OJ, Poobalan AS, Dineen B, *et al*. Effects of Probiotics on child growth: a systematic review. *J Health Popul Nutr* 2015;34:8:8.:.
35 Galway LP, Acharya Y, Jones AD. Deforestation and child diet diversity: A Geospatial analysis of 15 sub-Saharan African countries. *Health Place* 2018;51:78–88.
36 Oh C, Keats EC, Bhutta ZA. Vitamin and mineral supplementation during pregnancy on maternal, birth, child health and development outcomes in Low- and middle-income countries: A systematic review and meta-analysis. *Nutrients* 2020;12:491.
37 Girard AW, Self JL, McAuliffe C, *et al*. The effects of household food production strategies on the health and nutrition outcomes of women and young children: a systematic review. *Paediatr Perinat Epidemiol* 2012;26 Suppl 1:205–22.
38 Vilcins D, Sly PD, Jagals P. Environmental risk factors associated with child Stunting: A systematic review of the literature. *Ann Glob Health* 2018;84:551–62.