Linking Data from Demographic and Agricultural Surveys to Examine the Drivers of Stunting and Wasting in Nigeria

Lessons Learned

June 2020

USAID
FROM THE AMERICAN PEOPLE

MEASURE Evaluation

WORKING PAPER

Linking Data from Demographic and Agricultural Surveys to Examine the Drivers of Stunting and Wasting in Nigeria: Lessons Learned

Emily H. Weaver, PhD, Siân Curtis, PhD, John Spencer, PhD, Gustavo Angeles, PhD

MEASURE Evaluation
University of North Carolina at Chapel Hill
123 West Franklin Street, Suite 330
Chapel Hill, NC 27516 USA
Phone: +1 919-445-9350
measure@unc.edu
www.measureevaluation.org

This publication was produced with the support of the United States Agency for International Development (USAID) under the terms of MEASURE Evaluation cooperative agreement AID-OAA-L-14-00004. MEASURE Evaluation is implemented by the Carolina Population Center, University of North Carolina at Chapel Hill in partnership with ICF International; John Snow, Inc.; Management Sciences for Health; Palladium; and Tulane University. Views expressed are not necessarily those of USAID or the United States government. WP-20-241
ISBN: 978-1-64232-262-0

ACKNOWLEDGMENTS

We would like to acknowledge the United States Agency for International Development (USAID) for its support of this activity. We are grateful for the input and technical direction received from Lindsey Anna and Anne Swindale of the USAID Bureau for Resilience and Food Security in conceptualizing and carrying out this work. We also are grateful for the programming support of Aiko Hattori and technical advice provided by Linda Adair. We thank MEASURE Evaluation's knowledge management team at the University of North Carolina at Chapel Hill for editorial, design, and production support.

CONTENTS

Acknowledgments 4
List of Figures 6
List of Tables 7
Abbreviations 8
Executive Summary 9
    Study Design 9
    Results 9
    Discussion 10
    Recommendations 11
    Conclusion 11
Introduction 12
    Study Motivation 12
    Study Objectives 13
Methods 14
    Selection of Country for Analysis 14
    Study Design 14
    Conceptual Framework 14
    Sample Frame 15
    Data Requirements, Collection, and Management 16
    Linking the Survey Datasets 17
    Variable Definitions 18
    Estimation Strategies 20
Results 22
    Multivariate Analysis 22
Discussion 31
    Limitations and Lessons Learned 31
    Aggregation and Linking of Data 32
    Variable Specifications 33
    Machine Learning 34
Recommendations 35
Conclusion 36
References 37
Appendix. A. Country Selection Process 38
    Declining Stunting and Declining Wasting 40
    Declining Stunting and No Change in Wasting 43
    Declining Stunting and Increasing Wasting 45
    No Change in Stunting and No Change in Wasting 46
    Increasing Stunting and No Change in Wasting 47
Appendix B. Complete Multivariate Results 51

LIST OF FIGURES

Figure 1. Framework for achieving optimum growth and development 15
Figure 2. Overlap of the NDHS and GHS-Panel surveys in 2008 and 2010, respectively 16
Figure 3. Depiction of nearest neighbor matching method 18
Figure 4. Surface for wasting; comparison of ArcGIS Kernel Density Estimation and PrevR 33
Figure A1. Wasting and stunting trends in Ethiopia 40
Figure A2. Wasting and stunting trends in Ghana 41
Figure A3. Wasting and stunting trends in Kenya 41
Figure A4. Wasting and stunting trends in Nepal 42
Figure A5. Wasting and stunting trends in Uganda 42
Figure A6. Wasting and stunting trends in Bangladesh 43
Figure A7. Wasting and stunting trends in Guatemala 43
Figure A8. Wasting and stunting trends in Honduras 44
Figure A9. Wasting and stunting trends in Senegal 44
Figure A10. Wasting and stunting trends in Nigeria 45
Figure A11. Wasting and stunting trends in Niger 46
Figure A12. Wasting and stunting trends in Zambia 46
Figure A13. Wasting and stunting trends in Mali 47

LIST OF TABLES

Table 1. List of variables for analysis of wasting and stunting 18

Table 2. Descriptive statistics of matched and matched + nearest neighbor samples, NDHS 2008 and GHS-Panel 2010/2011 22

Table 3. Selected multivariate results assessing drivers of wasting and stunting in Nigeria, NDHS 2008 and GHS-Panel 2010/2011 26

Table 4. Machine learning results for stunting 29

Table 5. Machine learning results for wasting 30

Table A1. Trends in stunting and wasting, DHS STATCompiler 1986–2017 38

Table A2. Identified food security databases 48

Table A3. Inventory of spatial data sets obtained for possible use in Nigeria 49

Table B1. Complete multivariate results assessing drivers of wasting and stunting in Nigeria, NDHS 2008 and GHS-Panel 2010/2011 51

ABBREVIATIONS

ARI	Acute respiratory infection
BFS	Bureau for Resilience and Food Security
BMGF	Bill and Melinda Gates Foundation
BMI	Body mass index
DHS	Demographic and Health Surveys
EA	Enumeration area
ETC	Extra trees classifier
FTF	Feed the Future
GHS-Panel	General Household Survey Panel
GIS	Geographic information system
LGA	Local government authority
ML	Machine learning
PSU	Primary sampling unit
RFC	Random forest classifier
SD	Standard deviation
USAID	United States Agency for International Development

EXECUTIVE SUMMARY

Stunting and wasting are still global issues, with an estimated 149 million children under five with stunted growth and 49 million children under five suffering from wasting worldwide. Wasting and stunting can have severe health effects on children and are therefore a major health concern for most low-middle income countries where stunting and wasting rates are highest (UNICEF/WHO/World Bank Group, 2019). Both stunting and wasting share underlying risk factors that derive from several different levels of influence. Existing studies focus on demographic and health indicators, such as those that are available in the Demographic and Health Surveys (DHS). However, additional influences on these outcomes are also agricultural and community-level indicators that are not included in conventional demographic and health surveys. Studies are needed to trial the linking of these data and to provide lessons learned for others seeking to do the same.

The increased availability of data from multiple sources in low- and middle-income countries in recent years, combined with advances in data science, have stimulated an increased interest in using existing data in innovative ways to bring new insights to population, health, and nutrition problems. MEASURE Evaluation was contracted to do just that—to conduct an analysis of publicly available secondary data using innovative linking methods to better understand a broader range of drivers of wasting and stunting, particularly in contexts with stagnant or increasing wasting levels and decreasing stunting trends. The study links data from the Nigeria DHS (NDHS) with a Living Standards Measurement Survey Integrated Survey on Agriculture (LSMS-ISA) that contains agricultural and community information. This study also sought to use machine learning to identify additional or unique patterns of indicators that influence stunting and wasting. Neither of these two methods are prevalent in current research; therefore, these analyses also serve as proof of concept for these two approaches and provide lessons learned for future research.

Study Design

This was a retrospective cross-sectional analysis of Nigerian data. The analysis used two sources of secondary data: the NDHS and the Nigeria General Household Survey-Panel (GHS-Panel), which is one of the World Bank's Living Standards Measurement Survey Integrated Surveys on Agriculture. Wasting, stunting, and underlying individual and household-level correlates were assessed using the 2008 NDHS, which was merged with the 2010/2011 GHS-Panel at the local government area (LGA) level. All GHS-Panel household and community indicators were aggregated to the LGA level before being merged with the NDHS at that level. Two approaches to merging were used: 1) a matched method where only LGAs that directly merged from the GHS-Panel with the NDHS were included, and 2) a nearest neighbor matching method, which pooled estimates from adjacent LGAs—"nearest neighbors"—to create an estimate for the GHS-Panel where a LGA could not be directly matched with the NDHS. Both multivariate analysis and machine learning methods were applied.

Results

The prevalence of stunting and wasting in the NDHS matched and nearest neighbor samples were approximately 44 percent and 14 percent, respectively . The multivariate results of demographic and health variables on stunting and wasting found consistent results with existing literature on correlates of wasting and stunting. When we compared correlates of wasting and stunting, we found geographic region, BMI of mother, and mother's education level to be common correlates.

When we added in indicators at the LGA level from the NDHS and GHS-Panel, including indicators for food availability, economic access to food, classroom education, women's empowerment, social safety nets, and health services, the results were less clear. Very few LGA level variables showed an association with wasting and stunting.

The machine learning approaches performed better for wasting than for stunting. Results were dependent on the machine learning approach with the ensemble methods identifying a different set of features than the single decision tree approach. Education level of the mother was overwhelmingly the most important feature for the decision tree classifier but not identified as a feature in the Random Forest Classification (RFC) or Extra Trees Classifier (ETC). RFC and ETC had a longer list of features identified and no one feature was overwhelmingly important, although BMI of the mother was important in the nearest neighbor RFC and the matched ETC.

Discussion

By linking the NDHS and GHS-Panel data we aimed to generate lessons learned for future analyses of nutrition outcomes using existing data and to establish a proof of concept for linking additional nutrition-sensitive data to common nutrition datasets such as the DHS. Overall, the additional indicators from the GHS-Panel did not add any meaningful additional insights on the drivers of stunting and wasting in this analysis. However, we encountered several challenges linking the NDHS and GHS-Panel datasets that limit the scientific strength of the results. Lessons learned include the following:

Temporality and availability of datasets. The GHS-Panel data were collected two years after the NDHS, so we had to treat the nutrition-sensitive variables from the 2010 GHS-Panel as proxies for the variables in 2008 when the NDHS was collected. This is a strong assumption given annual variation in agricultural and other conditions, which effectively introduces measurement error into the variables from the GHS-Panel data and undermines the substantive findings. The differential timing of data collection from different data sources is a significant challenge in using existing data from multiple surveys. In selecting countries for this study, we found that it was difficult to identify countries that had existing datasets for the same time periods that could be linked to provide both nutrition-sensitive and nutrition-specific variables for analysis.

Aggregation and linking of data. The NDHS and the GHS-Panel survey are independently sampled national surveys, so the data collected in each survey refer to different households. To address this issue we need to have a common unit in which to link the data. We linked the surveys spatially at the LGA level by aggregating data from the GHS-Panel survey to the LGA level and linking the aggregated LGA-level variables to the individual data from the DHS. The LGA was the lowest geographic level of aggregation feasible but an LGA is still a sizeable geographic area. Further, neither the DHS nor the GHS-Panel survey were designed to produce statistically reliable estimates at the LGA level; sometimes a single GHS-Panel cluster represented that entire LGA. This means that there is considerable sampling error associated with the aggregated GHS-panel variables.

Variable specifications. The conceptual framework we used to guide this analysis includes a wide range of different factors that are hypothesized to influence nutritional outcomes. The DHS and the GHS-Panel survey contain many questions relevant to different elements in the conceptual framework, but they did not always readily translate into variables that were usable for this analysis.

Use of machine learning for identifying drivers of stunting and wasting. The machine learning analysis did not add significant insight into the factors associated with stunting and wasting in this analysis, with different modeling approaches producing different results. However, this analysis employed the most basic random trees classifier and extra trees classifier. A best practice would be to refine the machine learning models by doing feature engineering as well as adjusting decision tree parameters that could improve the model. Model selection using machine learning in a context where different approaches produce varying results is important because it could easily lead to different programmatic implications.

Recommendations

Recommendations for future work linking data from different sources to explore drivers of nutritional outcomes based on the lessons learned from this study include:

1) We do not recommend linking data from two or more independent sample surveys at low levels of geographic disaggregation, such as the LGA. Linking appears more likely to work well when linking a sample survey such as DHS to a dataset that has complete geographic coverage, such as remote-sensed climate data (e.g., Tusting et al., 2020).
2) It could be worth further exploring linking sample survey data to modeled surfaces obtained from other data sources, but we recommend focusing on using one modeled surface initially to test the idea. Using modeled surfaces won't eliminate measurement error concerns for variables at low levels of geographic disaggregation and will work better for some variables than others depending on the precision of the models. It is time intensive to create modeled surfaces so using existing publicly available models will be more efficient. However, the existing models we are aware of tend to focus on outcome variables such as stunting, rather than the nutrition-specific independent variables that were the focus of this analysis.
3) For some research questions, linking data at higher levels of geographical disaggregation, such as a region, could be sufficient for meaningful analysis. For example, individual data from a DHS survey could be linked to aggregate variables at the regional level from an agricultural survey that provides adequate statistical precision for regional level estimates. Whether this is meaningful will depend on whether the regional level is granular enough to be relevant conceptually to the question of interest.
4) In all linked analysis the different data need to refer to the same time period, or to be otherwise meaningfully related temporally.

Conclusion

This study has generated learning on the process and challenges involved in bringing together data from different sectors to analyze the drivers of stunting and wasting in Nigeria. We identified several significant data challenges that undermine the credibility of the substantive findings and were unable to demonstrate that the proposed approach was feasible in addressing the questions presented as the primary focus of the case study. Similar challenges are likely to be present in other contexts. However, some of the approaches we explored for this analysis could be useful in other applications.

New data science methods are not a panacea for underlying data problems such as sparse or poor-quality data. They will work better for some applications than for others. Use of existing data requires compromises to accommodate the data available, which can undermine the credibility of the substantive results. More sophisticated methods can require substantial investments of time from highly specialized analysts, which will have significant cost implications. Using existing data will not necessarily be cheap and quick. Although we have more and better data now than we have ever had, and more and better ways of analyzing them, challenges remain to linking data for complex analyses.

INTRODUCTION

Stunting and wasting are still global issues, with an estimated 149 million children under five with stunted growth and 49 million children under five suffering from wasting worldwide. Wasting is generally thought of as an acute condition whereas stunting can be the result of repeated or chronic insults to health and can have more enduring implications for growth and development (Richard, Black, Gilman, Guerrant, Kang, et al., 2012). Wasting and stunting can have severe health effects on children and are therefore a major health concern for most low- and middle-income countries where stunting and wasting rates are highest (UNICEF/WHO/World Bank Group, 2019).

Although wasting and stunting represent different processes of malnutrition, both stunting and wasting share underlying risk factors such as low birth weight, small birth size, increasing child's age, male sex of child, prolonged duration of breastfeeding (>12 months), diarrheal episodes, low mother's and father's education, mother's age (<20 years), low mother's body mass index (BMI), unimproved source of drinking water, less wealth, and rural place of residence (Akombi, Agho, Hall, Wali, Renzaho, et al., 2017; Ricci and Becker, 1996). Other literature shows that factors specific to wasting are prenatal factors, intrauterine growth restriction, small body size, infections, and food insecurity (Martorell and Young, 2017; Motbainor, Worku, Kumie, 2015). Separate factors associated with stunting are food diversity (Motbainor, Worku, Kumie, 2015) and nutrition education or counseling (Panjwani and Heidkamp, 2017). Prevalence of stunting and wasting may also vary seasonally, possibly in part due to changes in feeding practices (Roba, O'Connor, Belachew, and O'Brien, 2016; Young and Marshak, 2018).

Conceptual frameworks for nutrition draw from multiple levels of influence on stunting and wasting and relate growth and development to variables that are not frequently in one dataset. Existing studies focus on demographic and health, or nutrition-specific, indicators, such as those that are available in the Demographic and Health Surveys (DHS). However, additional influences on these outcomes also include agricultural and community-level, or nutrition-sensitive, indicators that are not included in conventional demographic and health surveys. Studies examining drivers of stunting and wasting have not traditionally linked data from other surveys, such as the World Bank's Living Standards Measurement Study – Integrated Surveys on Agriculture (LSMS-ISA) or other harvest surveys that contain such agricultural and community-level information. Studies are needed to trial the linking of these data to better understand the combined influences of both types of indicators on stunting and wasting outcomes and to provide lessons learned for others seeking to do the same.

Study Motivation

The USAID Bureau for Resilience and Food Security (RFS)seeks to achieve greater programming efficiency and effectiveness to accelerate the reduction of wasting and stunting in all contexts by jointly addressing both manifestations of undernutrition. MEASURE Evaluation was contracted to conduct an analysis of publicly available secondary data using innovative statistical methods to better understand a broader range of drivers of wasting and stunting in contexts experiencing stagnate or increasing wasting levels amid decreasing stunting trends. MEASURE Evaluation identified two ways of approaching this request. First, this study links data from the DHS with a LSMS-ISA in Nigeria that contains agricultural and community information. Second, this study sought to use machine learning to identify additional or unique patterns of indicators that influence stunting and wasting. This study serves as a proof of concept for how to use and implement these two approaches while also providing lessons learned for future research.

Study Objectives

MEASURE Evaluation conducted a cross-sectional analysis of stunting and wasting, assessing nutrition-sensitive and nutrition-specific influences, through the linking of two publicly available secondary datasets—the DHS and LSMS-ISA. The specific objectives of this study were:

1. To conduct a cross-sectional analysis using the Nigeria Demographic and Health Survey (2008 NDHS) to attempt to understand the various nutrition-specific and nutrition-sensitive influences on stunting and wasting.
2. To provide learning about the feasibility of replicating the analysis—including techniques for linking two secondary datasets and integration of machine learning approaches—for other countries.
3. To establish a proof of concept for the analysis approach.

METHODS

Selection of Country for Analysis

Initially, the goal of the analysis was to assess changes over time in the influences of nutrition-specific and nutrition-sensitive indicators on stunting and wasting. We conducted an initial exploratory descriptive analysis of trends in stunting and wasting in all DHS countries with available trend data. We examined patterns of stunting and wasting and set a 10 percent wasting threshold for inclusion criteria. Specifically, we looked for a country where stunting was decreasing, and where wasting was increasing or constant and at or above 10 percent prevalence. Additionally, we sought a country with other datasets available for linking agricultural data. Based on these factors, Nigeria was selected as the case for this study using the two most recent NDHS survey years, 2008 and 2013, and the 2010/2011 and 2012/2013 GHS-Panel surveys. See Appendix A for further details.

After selecting Nigeria as the case study and moving forward in the linking process, we learned two more pieces of information. First, we learned that there are questions about the quality of measurement of our key outcomes for the 2013 NDHS; thus, we were not able to use it in the study. Second, the 2018 NDHS was released, which showed a decline in wasting between 2013 and 2018. Based on these two factors, we decided to pursue a cross-sectional analysis using the 2008 NDHS to serve as a proof of concept for the linking and machine learning approaches. The remainder of this section describes methods for the 2008 cross-sectional analysis using the NDHS.

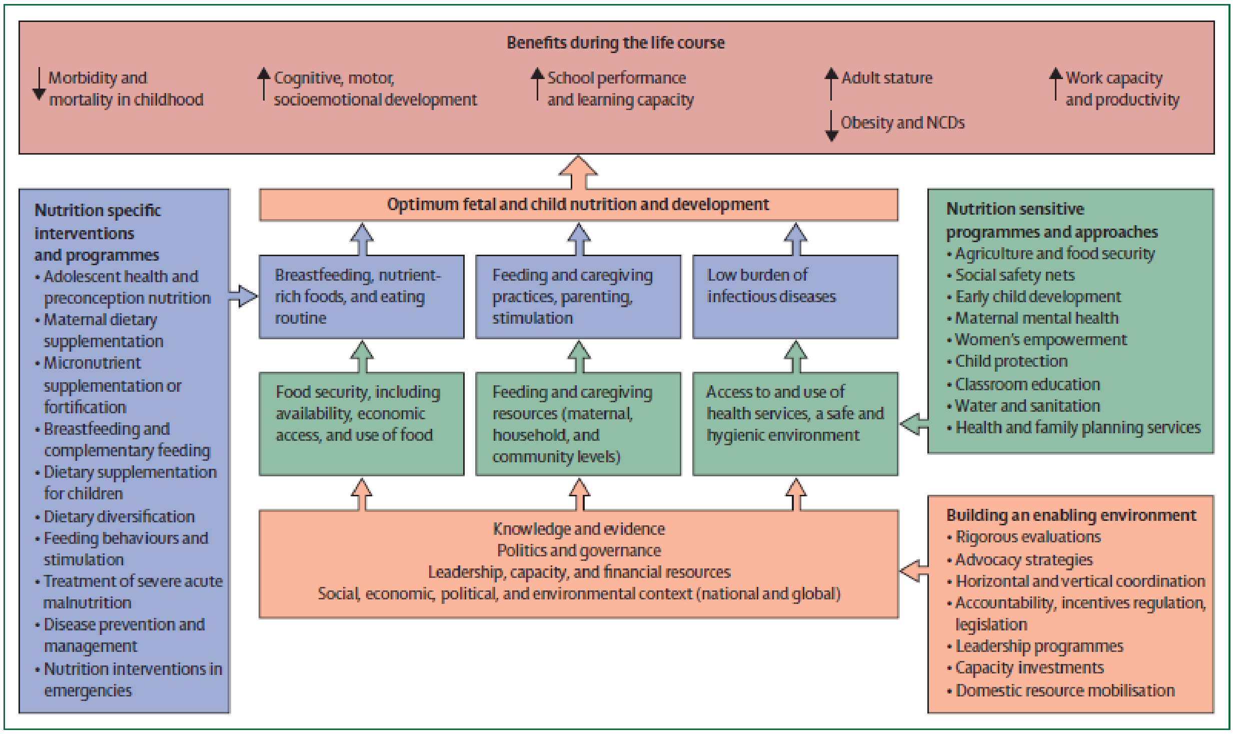
Study Design

This was a retrospective cross-sectional analysis of Nigerian data. The analysis used two sources of secondary data: the NDHS and the Nigeria General Household Survey-Panel (GHS-Panel), which is one of the World Bank's Living Standards Measurement Surveys. Wasting and stunting were assessed using the 2008 NDHS, which was merged with the 2010/2011 GHS-Panel. Both multivariate analysis and machine learning methods were applied.

Conceptual Framework

This selection of indicators for this study was guided by the conceptual framework depicted in Figure 1 (Black, Victora, Walker, Bhutta, and Christian, 2013). This framework shows the means to achieving optimum growth and development and includes dietary, behavioral, and health determinants and how they are affected by other underlying factors such as food security, caregiving resources, and environmental conditions. Economic and social conditions, national and global contexts, resources, and governance shape the underlying conditions depicted in the conceptual model. The model also identifies nutrition-specific and nutrition-sensitive interventions that can address immediate and underlying causes and determinants of malnutrition. An enabling environment supports interventions and programs to improve outcomes.

Figure 1. Framework for achieving optimum growth and development



Source: Black, et al. (2013).

Sample Frame

Nigeria Demographic and Health Survey, 2008

The study analyzed secondary data from the NDHS conducted in 2008. The 2008 NDHS aimed to assess the demographic and health status of the population. The survey collected information on background characteristics, nutritional status, fertility, mortality, and morbidity from a nationally representative sample of households, women ages 15–49, and children under age 5 residing in noninstitutionalized dwelling units. The survey used a stratified multistage sampling design. In the 2008 NDHS, a sampling frame was constructed from the 2006 Population Census of the Federal Republic of Nigeria and was stratified by state. Then, each stratum was divided into census enumeration areas (EAs), which served as primary sampling units (PSUs). PSUs were selected with probability proportional to the estimated size within each stratum and a fixed number of 45 households were selected per PSU, yielding a total sample of 38,522 households, containing 38,948 women ages 15–49.

The 2008 NDHS measured height and weight of all de facto children under five years old in households selected for the survey, regardless of their mother's eligibility/interview status, and all children under age five whose mothers were interviewed. Data on maternal characteristics was needed for this analysis, and therefore, the study focused on the children's sample consisting of those whose mothers were eligible and interviewed in the NDHS. This inclusion criteria yielded a total of 25,446 children under age five for this study.

Nigeria General Household Survey-Panel, 2010–2011

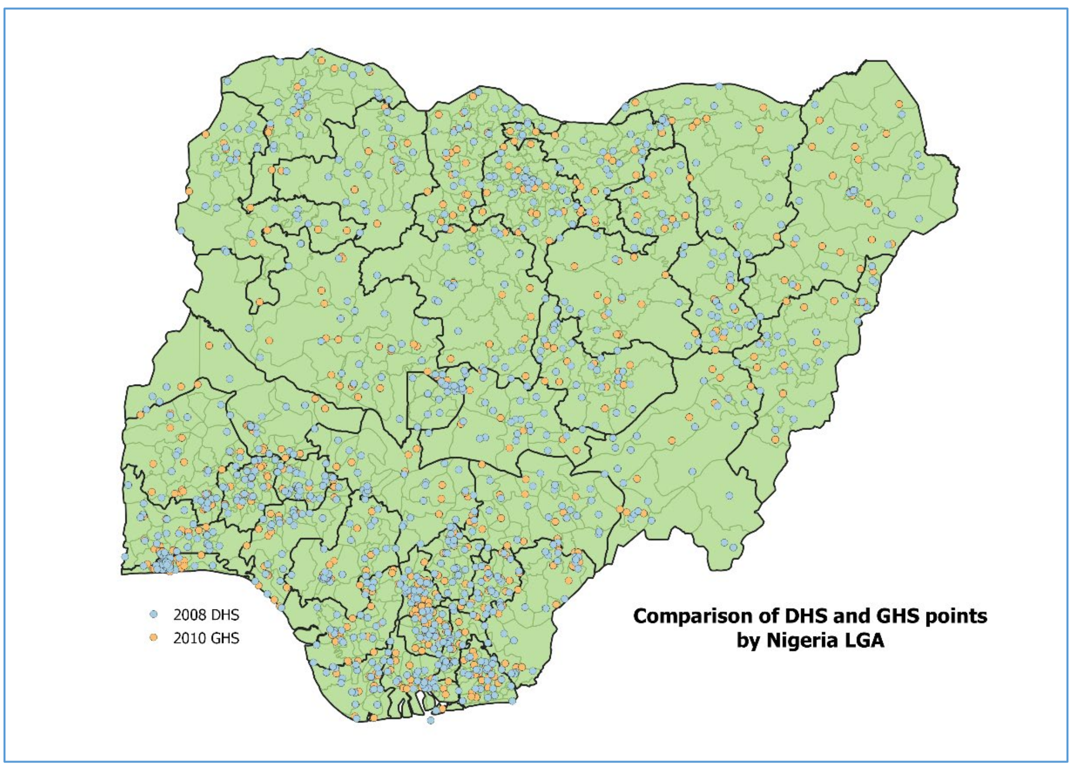
The GHS-Panel is a subset of 5,000 households derived from the GHS survey of 22,000 households throughout the country. The panel component covered multiple agricultural activities and aimed to link data from the agricultural sector with additional community and household behavior and other characteristics. The GHS-Panel was implemented by the Nigeria Bureau of Statistics and 2010 was the first wave of the panel. In addition to agriculture, the GHS-Panel gathered information on poverty and socioeconomic characteristics of households. Information was gathered during two stages for the survey—during a post-planting period and during a post-harvest period.

The GHS Panel survey used a two-stage stratified sample selection process. The first stage selected EAs with probability proportional to size in all 36 states and the Federal Capital Territory, Abuja. The sampling frame was also constructed from the 2006 Population Census of the Federal Republic of Nigeria. The second stage randomly sampled ten households per EA. Data from 4, 840 households were aggregated to 379 LGAs, which were available for linking with the NDHS.

Data Requirements, Collection, and Management

The requisite datasets are publicly available online by request from the DHS program and the Nigerian National Bureau of Statistics. The NDHS GIS dataset was also requested. These datasets were merged at the lowest administrative unit possible for analysis, which was the LGA level. Merging the datasets required that the data were collected, or could be aggregated, at the same levels. To assess the merging of these two datasets, we mapped the overlap of the NDHS and GHS-Panel survey to estimate the number of LGAs with households surveyed in both the NDHS and GHS-Panel (see Figure 2). Approximately 40 percent of the total LGAs had sampled units within both surveys.

Figure 2. Overlap of the NDHS and GHS-Panel surveys in 2008 and 2010, respectively



Linking the Survey Datasets

Linking these two datasets allowed for the use of additional nutrition-sensitive indicators that were not available in the NDHS alone. The NDHS and GHS-Panel did not use the same unique identifiers for LGAs, so a GIS approach was used to facilitate linking the datasets. An initial challenge was to find an accurate LGA boundary file to serve as our reference. These can be difficult to find because LGA boundaries can change frequently and are often not well defined.[1] We found that The Bill and Melinda Gates Foundation (BMGF) produced an LGA boundary file to support their vaccination program and that file is considered to be one of the more accurate options. The locations of the NDHS and GHS-Panel clusters were overlaid onto the BMGF LGA boundary file and a table was created that identified which LGA each cluster was located in, using a unique LGA ID included in the BMGF file. It was then possible to link the NDHS and GHS-Panel using that BMGF unique ID in STATA and Excel.

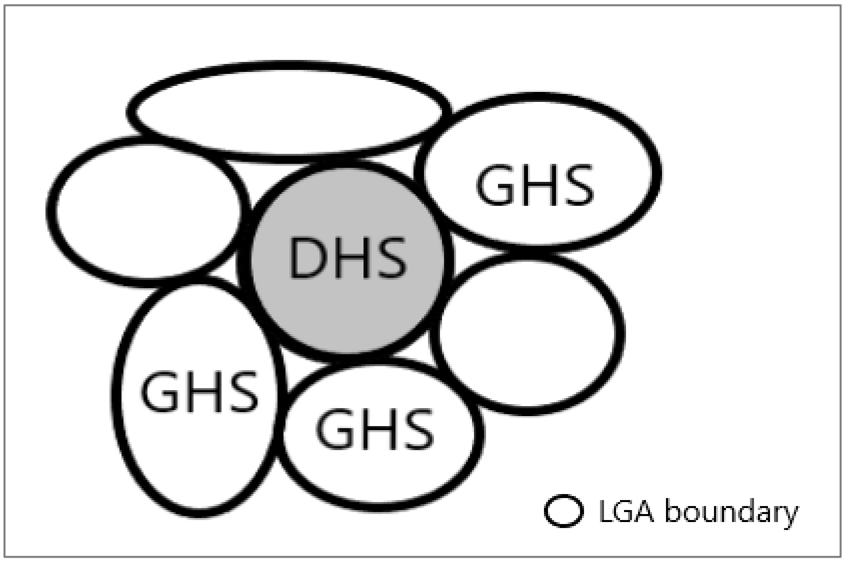
We used a two-step process to link aggregated LGA-level variables from the GHS-Panel to the NDHS clusters. For the NDHS clusters in LGAs that also had at least one GHS-Panel cluster, we were able to directly link the NDHS cluster to the GHS-Panel clusters in that LGA. This provided a sample of 7,831 children for analysis who had GHS-Panel data from the same LGA in which they were sampled in the NDHS.

For the NDHS clusters located in an LGA in which there was no GHS-Panel cluster we used a "nearest neighbor" matching approach to link NDHS clusters with GHS-Panel clusters in adjacent LGAs. Figure 3 shows a NDHS LGA surrounded by six other LGAs, three of which had a GHS-Panel cluster. In this case, the averages of the GHS-Panel LGA-level variables for the three neighboring GHS-Panel clusters would be linked to the NDHS cluster. An average for each LGA would then be computed and the average of the three LGAs would be used. Please refer to the section "Variable Definitions" below for definitions of the GHS-Panel LGA-level variables. The nearest neighbor matching increases the analysis sample to 14,192 children. The nearest neighbor matching has the benefits of increasing the sample size for analysis and reduces potential selection bias associated with restricting the sample to NDHS clusters that are located in an LGA that also has a GHS-Panel cluster. However. it assumes that the LGA-level variables from adjacent LGAs are good proxies for the LGA in which the NDHS cluster is located so introduces an additional level of potential measurement error. We ran both the multivariate analysis and the machine learning (ML) analysis on both the matched NDHS sample that links to a GHS-Panel cluster in the same LGA and the larger sample that also includes NDHS clusters linked to the GHS-Panel using nearest neighbor matching to test the sensitivity of our results to the different linking approaches.

---

[1] According to the lead mapper at BMGF, Nigerian LGA boundaries are often based on lines drawn on maps by hand as opposed to being precisely surveyed. As a result, when those lines are digitized in a GIS they may be off by several kilometers because the GIS has a higher level of precision than the paper map used when the line was drawn. BMGF visited nearly every settlement in the country and identified many locations that were supposed to be in one LGA but were in fact in another. They modified their LGA boundary file to match what they found on the ground.

Figure 3. Depiction of nearest neighbor matching method



Variable Definitions

For this study, we used two analytic approaches—multivariate analysis and ML. Both approaches used variables from the conceptual framework in an empirical model of determinants of wasting and stunting at the country level (Table 1). These variables included components of the nutrition-specific and nutrition-sensitive programs and approaches available in the 2008 NDHS (e.g., breastfeeding, economic access, safe and hygienic environment) and 2010/2011 GHS-Panel (e.g., measures of food accessibility, infrastructure, and social safety net programs). Individual and household-level variables were drawn from the NDHS and most community-level nutrition-sensitive variables were drawn from the GHS-Panel and aggregated to the LGA level. Several household-level variables were aggregated to the LGA level from the NDHS as well, specifically those concerning women's empowerment.

Table 1. List of variables for analysis of wasting and stunting

Variable	Level	Definition	Dataset
Outcomes			
Nutritional indicator	Child	Whether child is stunted or wasted using the WHO standard (<2SD)	NDHS
Nutrition-specific indicators			
Breastfeeding	Child	Child is never breastfed, still breastfeeding or weaned	NDHS
Care during pregnancy	Mother	Took iron for the latest birth in the last 5 yrs	NDHS
		Took vitamin A or micronutrients after the latest birth in the last 5 yrs	
		Was de-wormed during pregnancy for the latest birth in the last 5 yrs (dichotomous)	
		Received tetanus vaccination per recommendations during pregnancy for the latest birth in the last 5 yrs (dichotomous)	
		Whether or not she took anti-malaria drug during pregnancy for the latest birth in the last 5 yrs	

Variable	Level	Definition	Dataset
Infectious disease	Household	Any child had diarrhea, ARI symptoms, or fever in the last two weeks	NDHS
Dietary diversity	Child/ Household/ LGA*	Number of food groups consumed by child ages 6–23 months	NDHS
Nutrition-sensitive indicators			
Body mass index	Mother	Weight in kilograms divided by height in meters squared	NDHS
Family planning	Mother	Current use of family planning (dichotomous)	NDHS
Education	Mother	Highest level of education (categorical per NDHS definition)	NDHS
Employment	Mother	Mother is engaged in agricultural work (dichotomous)	NDHS
Empowerment	Mother	Mother participates in all decision making (dichotomous) Currently married women disagree with all of the reasons justifying wife-beating (dichotomous)	NDHS
Wealth index	Household	Wealth quintile (per DHS definition)	NDHS
Water and sanitation	Household	HH has improved drinking water source based on DHS definition (dichotomous) HH has improved sanitation (categorical)	NDHS
Age dependency ratio	Household	Number of children less than five plus adults over age 60 per woman age 15–49	NDHS
Livestock ownership	Household	HH owns any livestock	NDHS
Any land ownership	Household	HH owns any land	NDHS
Shocks	LGA	Proportion of households with three or more negative shocks in the last 12 months	GHS-Panel
Primary school†	LGA	Primary school in the LGA	GHS-Panel
Secondary school†	LGA	Secondary school in the LGA	GHS-Panel
Nursery†	LGA	Nursery in the LGA	GHS-Panel
Health center†	LGA	Health center in the LGA	GHS-Panel
Microfinance institution†	LGA	Microfinance institution in the LGAs	GHS-Panel
Bank†	LGA	Bank in the LGAs	GHS-Panel
Food availability	LGA	LGAs where dairy, eggs, or vitamin A-rich foods are available at the market	GHS-Panel
Social safety nets	LGA	Percentage of households that report receiving any assistance aggregated to LGA level, proxy for safety net availability	GHS-Panel
Other variables			
Age of child	Child	Age in years	NDHS
Gender	Child	Male or female	NDHS
Age of mother	Mother	Five-year age group of mother	NDHS
Region	LGA	Region of LGA	NDHS
Household location	Household	Urban/rural status of household location	NDHS

† dichotomous for matched LGAs, proportion of clusters with infrastructure for nearest neighbor LGAs.
* For this indicator, a value for children ages 6–23 months was estimated. For children in the same household as the child age 6–23 months, the sibling value was used. For households without a child age 6–23 months, the LGA average of the indicator value was imputed.

Several indicators were derived for the analysis that were not based on standard definitions. A categorical measure of breastfeeding was created to indicate whether the child was still breastfeeding, weaned, or never breastfed. To estimate the burden of infectious disease, a category from the conceptual framework, we created a dichotomous variable that indicated whether any child in the household had an episode of diarrhea, fever, or respiratory infection in the last two weeks. An index of dietary diversity was created using the seven standard food groups defined in the global standard indicator of minimum acceptable diet for children ages 6–23 months. Because data for these food groups are only collected for children between ages 6–23 months old, we imputed the value from each household, where available, to other children in the same household. Where no child age 6–23 months lived in the household, we imputed an LGA average of dietary diversity for the child (or children). Finally, we sought to create a measure of food availability based on the same standard food groups. However, we found that four of the seven food groups were almost ubiquitous in the surveyed communities; thus, we used availability of the three other food groups directly in the models that provided some variability of availability in the surveyed communities. These food groups were dairy, eggs, and vitamin A-rich foods.

Estimation Strategies

Multivariate Analysis

We used the following model to examine the determinants of wasting and stunting on children under age five:

$$Y_{ijkm} = \beta_0 + \beta_1 X_{ijkm} + \beta_2 IndBeh_{ijkm} + \beta_3 C_{jkm} + \beta_4 NutSpec_{jkm} + \beta_5 NutSens_{jkm} + \beta_6 NutSens_{km} + \lambda_m + \varepsilon_{ijkm} \quad (1)$$

where,

Y_{ijkm}:	Outcome of interest (wasting or stunting) for individual i in cluster j located in LGA k of region m
X_{ijkm}:	Individual and household background characteristics (e.g., age, gender, mother's education, "dependency ratio")
$IndBeh_{ijkm}$:	Individual nutrition-related behaviors (e.g., breastfeeding, mother's use of health services during pregnancy, and whether the child consumes a diverse diet)
C_{jkm}:	Cluster-level characteristics, aggregated from the NDHS (non-programmatic community variables that influence nutrition, e.g., urban/rural)
$NutSpec_{jkm}$:	Nutrition-specific interventions in cluster j located in LGA k of region m, from the NDHS (e.g., infectious disease in the household)
$NutSens_{jkm}$:	Nutrition-sensitive interventions in cluster j located in LGA k of region m, aggregated from the GHS-Panel, aggregated from the NDHS (e.g., beliefs about wife-beating, household decision making)
$NutSens_{km}$:	Nutrition-sensitive interventions in LGA k of region m, aggregated from the GHS-Panel (e.g., household shocks, presence of schools, banks, food availability)
λ_m:	Unobserved characteristics of region m, specified as region-level dummies
ε_{ijkm}:	Error term

The outcome variables of interest (stunting and wasting) were specified using dummy variables indicating if the child had that specific nutritional condition—that is, if the child was stunted or if the child was wasted. The regions *m* were specified as relevant groups of states which might have had substantive differences in nutritional influences between them, specifically, North, Northeast, Northwest, South, Southeast, Southwest, East, and West.

Model (1) was estimated with estimation methods applied on data from the 2008 NDHS merged with LGA-level aggregated data from the 2010/2011 GHS-Panel. Standard errors were corrected for clustering at the NDHS cluster level. In this model, the estimated coefficients, the $\beta's$, are interpreted as the contribution of the corresponding individual/household/LGA characteristics to the nutritional status outcome (wasting or stunting) of children between 6–59 months old.

Model (1) was estimated for stunting and for wasting, separately, and also stratified by urban and rural status of the NDHS households.

Machine Learning

ML is an approach well suited for identifying relationships in multiple, diverse datasets such as those that would be used to understand stunting and wasting. It is an approach that can find structures in data, which can be used to identify previously unrecognized patterns and relationships between variables. To help improve our understanding of the patterns behind stunting and wasting, an exploratory analysis using ML techniques was conducted. ML identifies correlation, not causation, and therefore the findings from the analysis identifies relationships between variables but no conclusions about causality can be drawn. This is in contrast to multivariate analysis techniques which do allow for conclusions of causality under certain circumstances.

There are many different ML approaches and techniques. However, due to the nature of the data included in this study, this analysis used semi-supervised ML algorithms. Semi-supervised ML approaches use a mix of records where the outcome (stunted/nonstunted or wasted/nonwasted) is known and unknown to the model as it is being trained.[2] ML algorithms were applied to the linked datasets to identify relationships between stunting/wasting and variables from the conceptual framework. The ML algorithm culled variables that do not have a relationship to stunting/wasting and evaluated the importance of the variables that do.

Random Forest Ensemble (RFE) methods were the ML approach used in the analysis. RFE was selected over other methods because it can perform well with the data limitations that existed and could be executed with a minimal amount of computing resources (human and technological). RFE methodology uses adaptive learning to combine several algorithms that decrease variance and bias to improve predictions. Three different random forest approaches were employed:

- **Decision Tree Classifier**: A basic classifier. A series of decision rules are used to determine classifications
- **Random Forest Classifier (RFC)**: An ensemble learning technique where a collection of decision trees are generated to find the best classifiers.
- **Extra Trees Classifier (ETC)**: Similar to RFC, in that it is an ensemble technique, but there is a random component to the creation of decision trees that is not present in RFC.

[2] The algorithm uses 70 percent of the data to develop (or train) the model then tests the fit of the model by using it to predict the outcome for the remaining 30 percent of the data and then comparing the predicted outcome with the actual outcome.

RESULTS

Multivariate Analysis

Table 2 provides descriptive statistics for the matched and matched + nearest neighbor samples. The prevalence of stunting and wasting in the samples was approximately 44 and 14 percent, respectively. The majority of the sample population lived in the northern regions in rural areas. Seventy-percent of children under age five were weaned and about 28 percent were still breastfeeding. Very few children had never been breastfed. Mothers' BMI was about 22.5 and fewer than 11 percent of mothers used family planning. Almost half of mothers had no education and one-fifth worked in agriculture. Although about half of the sample households in both samples had an improved source of water, over half of households in both samples had unimproved sanitation facilities. A small percentage of LGAs had markets; however, eggs and dairy were available in most locations with a market.

While most demographic and health indicators are very similar across the two samples, small differences exist across the samples, primarily regarding LGA level infrastructure characteristics (e.g., school and bank availability, and access to food at markets) that are mainly derived from the GHS-Panel dataset.

Table 2. Descriptive statistics of matched and matched + nearest neighbor samples, NDHS 2008 and GHS-Panel 2010/2011

	Matched (percentage)	Matched + Nearest Neighbor (percentage)
Nutritional status		
Stunted	44.2	44.8
Wasted	13.9	14.2
Sex of child (NDHS)		
Female	50.3	50.4
Male	49.7	49.6
Mother's age (NDHS)		
Age 15–19 years	4.5	4.4
Age 20–24 years	17.8	18.3
Age 25–29 years	29.3	29.0
Age 30–34 years	22.4	21.9
Age 35–39 years	15.3	15.2
Age 40–44 years	7.4	7.8
Age 45–49 years	3.3	3.3
Household location (NDHS)		
Urban	28.8	28.5
Rural	71.2	71.5
Geographic region (NDHS)		
North	18.3	17.9
Northeast	21.6	22.5
Northwest	21.9	24.7
Southeast	11.4	9.2
South	14.2	12.1
Southwest	12.6	13.7

	Matched (percentage)	Matched + Nearest Neighbor (percentage)
Nutrition-specific factors		
Dietary diversity score (NDHS)	2.8	2.8
Household infection, last two weeks (NDHS)	34.6	34.0
Child's age (NDHS)		
Age 6–11 months	12.9	12.8
Age 12–23 months	22.7	22.2
Age 24–35 months	21.2	21.4
Age 36–47 months	23.0	22.9
Age 48–59 months	20.2	20.7
Breastfeeding status (NDHS)		
Never breastfed	0.9	1.0
Still breastfeeding	28.5	28.8
Weaned	70.6	70.2
Nutrition-specific interventions and programs		
Mother's care during last pregnancy (NDHS)		
Received vitamin A	25.4	24.3
Took malaria drugs	19.2	18.5
Took de-worming medication	10.6	10.1
Took iron supplements	55.6	54.9
Protected from tetanus	45.9	44.8
Nutrition-sensitive factors		
BMI of mother (NDHS)	22.6	22.4
Mother's use of family planning (NDHS)	10.5	10.5
HH has improved water source (NDHS)	49.5	49.5
HH sanitation (NDHS)		
Unimproved	54.9	52.6
Improved, shared	18.9	19.0
Improved, not shared	26.2	28.4
Food availability		
Market in LGA (GHS-Panel)	13.4	10.2
Food availability at market (GHS-Panel)		
Dairy available at market in LGA	93.9	79.4
Eggs available at market in LGA	93.3	78.3
Vitamin A-rich foods available at market in LGA	75.3	62.2
Economic access to food		
Wealth quintiles (NDHS)		
First quintile	23.8	24.2

	Matched (percentage)	Matched + Nearest Neighbor (percentage)
Second quintile	22.5	23.0
Third quintile	20.0	19.9
Fourth quintile	17.5	18.0
Fifth quintile	16.1	14.8
Household age dependency ratio (NDHS)	1.6	1.6
Family owns livestock (NDHS)	63.4	65.0
Family owns land (NDHS)	72.3	73.5
Mother works in agriculture (NDHS)	22.3	21.0
Bank in LGA (GHS-Panel)	59.5	50.1
Microfinance institution in LGA (GHS-Panel)	89.6	75.8
Percent of households with 3+ shocks is last year (LGA) (GHS-Panel)	31.7	26.7
Classroom education		
Mother's education level (NDHS)		
No education	43.3	46.4
Primary	24.8	24.4
Secondary	26.2	23.8
Higher than secondary	5.7	5.3
Nursery in LGA (GHS-Panel)	55.9	45.8
Primary school in LGA (GHS-Panel)	84.1	71.4
Secondary school in LGA (GHS-Panel)	66.9	55.2
Women's empowerment (NDHS)		
Joint household decision-making (LGA)	21.3	21.3
Disagrees with wife-beating (LGA)	52.3	52.4
Social safety nets		
Safety net used in LGA (GHS-Panel)	14.3	12.0
Health services		
Health clinic in LGA (GHS-Panel)	30.8	24.5
Private health clinic in LGA (GHS-Panel)	61.6	50.7
Total N	7,831	14,192

Table 3 presents selected multivariate results for variables that had significant effects in at least one model (see Table B1 for full results). Because existing studies have examined the relationship between wasting and stunting mainly using surveys including demographic and health indicators, we first ran a standard linear probability model of wasting and stunting on variables from the NDHS for the sample of NDHS clusters that were in the same LGA as a GHS-Panel cluster (i.e., the matched sample). These results found geographic region, BMI of mother, mother's occupation in agriculture, and mother's education to be significantly associated with wasting (Table 3, panel 1). We also found that male sex, region, household infection in the last two weeks, BMI of mother, mother's education level, household wealth, and household age dependency ratio were significantly associated with stunting (Table 3, Panel 4). These results are generally consistent with existing literature on correlates of wasting and stunting. When we compared correlates of wasting and stunting, we found geographic region, BMI of mother, and mother's education level to be common correlates.

We then added in nutrition-sensitive indicators at the LGA level from the NDHS and GHS-Panel, including indicators for food availability, economic access to food, classroom education, women's empowerment, social safety nets, and health services (Table 3, panels 2 and 5). We found that eggs being available at a market in the LGA, a proxy for food availability, was additionally significantly associated with higher prevalence of wasting. None of the LGA level nutrition-sensitive variables were significantly associated with stunting.

Results from the matched with nearest neighbor sample are found in Table 3, panels 3 and 6. The increased sample size increases the power of the model to identify statistically significant effects. However, we also introduce additional measurement error in the LGA-level variables because of the nearest neighbor matching approach. For some variables, the effect size does not change much compared to the model for the matched sample but they become statistically significant in this model, likely due to the increased power associated with the larger sample size—for example, male sex for the wasting models (panels 1–3) and dietary diversity score for the stunting models (panel 4–6). For several other variables, both the effect size and statistical significance changed relative to the models for the smaller matched sample. Examples of these types of changes are vitamin A after the last pregnancy (panel 6), dairy available at market in LGA (panel 3), and existence of a primary school in LGA (panel 3). The effects of measurement error are hard to predict and, in these cases, multiple factors (e.g., selection, measurement error, and sample size) could be interacting to produce the different results.

Table 3. Selected multivariate results assessing drivers of wasting and stunting in Nigeria, NDHS 2008 and GHS-Panel 2010/2011

		Wasting			Stunting		
		Panel 1	Panel 2	Panel 3	Panel 4	Panel 5	Panel 6
	Level of indicator	NDHS (n=7,979)	Matched NDHS+GHS (7,831)	Matched + Nearest Neighbor matching NDHS + GHS (14,192)	NDHS (n=7,979)	Matched NDHS + GHS (7,831)	Matched + Nearest Neighbor matching NDHS + GHS (14,192)
Sex of child							
Female	CH	--	--	--	--	--	--
Male	CH	0.012	0.009	**0.013***	**0.039****	**0.040****	**0.044****
Region							
North	HH	--	--	--	--	--	--
Northeast	HH	**0.100****	**0.118****	**0.105****	-0.003	0.011	0.023
Northwest	HH	**0.053***	**0.067****	**0.067****	0.049	0.046	**0.061***
Southeast	HH	**0.033***	**0.040***	0.028	**-0.190****	**-0.189****	**-0.164****
South	HH	-0.001	-0.001	0.000	**-0.098****	**-0.095***	**-0.088****
Southwest	HH	**0.032***	0.027	0.006	**-0.059***	**-0.077***	**-0.059***
Nutrition-specific factors							
Dietary diversity score	CH/HH/ LGA	0.003	0.001	-0.001	0.009	0.009	**0.010***
Household infection, last 2 weeks	HH	0.015	0.009	0.010	**0.035***	**0.037***	**0.024***
Child's age (NDHS)							
Age 6–11 months	CH	--	--	--	--	--	--
Age 12–23 months	CH	0.301	0.417	0.167	-0.251	-0.206	-0.232
Age 24–35 months	CH	0.209	0.202	0.026	-0.318	-0.339	-0.360
Age 36–47 months	CH	0.066	-0.007	-0.083	-0.350	-0.359	-0.358
Age 48–59 months	CH	0.342	0.344	0.072	-0.477	-0.495	-0.336
Breastfeeding status							
Never breastfed	CH	--	--	--	--	--	--

	Level of indicator	Wasting			Stunting		
		Panel 1	Panel 2	Panel 3	Panel 4	Panel 5	Panel 6
		NDHS (n=7,979)	Matched NDHS+GHS (7,831)	Matched + Nearest Neighbor matching NDHS + GHS (14,192)	NDHS (n=7,979)	Matched NDHS + GHS (7,831)	Matched + Nearest Neighbor matching NDHS + GHS (14,192)
Still breastfeeding	CH	-0.279	-0.292	-0.306	0.049	0.047	-0.082
Weaned	CH	-0.157	-0.194	-0.258	-0.048	-0.046	-0.145
Age*Breastfeeding status							
Age 12–23 months * still breastfeeding	CH	0.215	0.167	0.221	-0.101	-0.220	0.030
Age 12–23 months * weaned	CH	0.079	0.057	0.146	-0.043	-0.162	0.071
Age 24–35 months * still breastfeeding	CH	0.245	0.256	0.329	0.066	0.066	**0.262***
Age 24–35 months * weaned	CH	0.123	0.164	0.258	0.079	0.083	0.245
Age 36–47 months * still breastfeeding	CH	0.445	0.445	0.390	0.185	0.261	**0.418***
Age 36–47 months * weaned	CH	0.154	0.182	0.241	0.176	0.244	**0.289***
Age 48–59 months * still breastfeeding	CH	0.438	0.458	0.315	-0.319	-0.313	-0.031
Age 48–59 months * weaned	CH	0.274	0.310	0.217	-0.110	-0.116	0.132
Nutrition-specific interventions and programs							
Mother's care during last pregnancy							
Received vitamin A	MO	-0.002	-0.002	0.007	-0.017	-0.019	**-0.028***
Took malaria drugs	MO	-0.002	-0.005	0.006	-0.023	-0.024	**-0.029***
Protected from tetanus	MO	-0.021	-0.023	**-0.018***	-0.004	-0.001	-0.006
Nutrition-sensitive factors							
BMI of mother	MO	**-0.003***	**-0.003***	**-0.003****	**-0.006****	**-0.006****	**-0.008****
Food availability							
Dairy available at market in LGA	LGA	--	-0.056	**-0.080***	--	0.045	**0.058***
Eggs available at market in LGA	LGA	--	**0.101****	**0.091****	--	-0.029	-0.016

	Level of indicator	Wasting			Stunting		
		Panel 1	Panel 2	Panel 3	Panel 4	Panel 5	Panel 6
		NDHS (n=7,979)	Matched NDHS+GHS (7,831)	Matched + Nearest Neighbor matching NDHS + GHS (14,192)	NDHS (n=7,979)	Matched NDHS + GHS (7,831)	Matched + Nearest Neighbor matching NDHS + GHS (14,192)
Vitamin A-rich foods available at market in LGA	LGA	--	-0.026	**-0.030***	--	-0.024	-0.030
Economic access to food							
Wealth quintiles							
First quintile	HH	--	--	--	--	--	--
Second quintile	HH	-0.017	-0.019	-0.011	-0.039	-0.035	-0.014
Third quintile	HH	-0.018	-0.019	**-0.038***	**-0.075****	**-0.069***	**-0.039***
Fourth quintile	HH	-0.032	-0.030	**-0.034***	**-0.098****	**-0.095****	**-0.077****
Fifth quintile	HH	0.000	-0.001	-0.009	**-0.153****	**-0.149****	**-0.111****
Household age dependency ratio		-0.004	-0.004	0.000	**-0.024***	**-0.024***	**-0.014***
Mother works in agriculture	MO	**-0.035***	**-0.028***	**-0.036****	0.003	0.003	0.004
Classroom education							
Mother's education level							
No education	MO	--	--	--	--	--	--
Primary	MO	**-0.043****	**-0.039***	**-0.042****	-0.026	-0.024	-0.012
Secondary	MO	**-0.056****	**-0.051****	**-0.053****	**-0.072****	**-0.070****	**-0.048***
Higher than secondary	MO	**-0.085****	**-0.079****	**-0.077****	**-0.135****	**-0.136****	**-0.108****
Nursery in LGA	LGA	--	-0.047	-0.051	--	0.007	-0.007
Primary school in LGA	LGA	--	0.027	**0.042***	--	-0.002	-0.002
Secondary school in LGA	LGA	--	0.011	-0.002	--	-0.006	0.010
R^2		.334	.332	.334	.464	.464	.468

* significant at the p≤0.05; ** significant at the p≤0.001

Note: CH = child, MO = mother, HH = household, LGA = local government authority.

Machine Learning

In this analysis, multiple Random Forest Techniques were used to identify the most important factors correlated with stunting and wasting at the individual level in tandem with multivariate analysis.

Children classified as stunted made up 40 percent of the records in the matched sample. This represents a fairly balanced dataset which can sometimes limit the effectiveness of some ML techniques. That seemed to be true in this case. No matter which ML approach was used or which of the two analysis samples was used, the ML model had only around a 60 percent accuracy rate (Table 4). This means that it correctly classified an individual's stunting status 60 percent of the time, barely better than chance.

Children classified as being wasted made up 14 percent of the records. The ML approaches performed better for wasting than for stunting; each approach correctly classified children about 85 percent of the time (Table 5). There was little difference in the results between the two analysis samples. While this is much better than the stunting results, it still is not particularly strong.

The ML analysis identified the **feature importance** of the variables. Feature importance is a quantification of which variables contributed the most to the decision tree model. In this analysis, the most important features varied based on the type of random forest approach used and the use of nearest neighbor matching (Tables 4 and 5). Education level of the mother was overwhelmingly the most important feature for the Decision Tree Classifier. Random Forest Classification and Extra Trees Classifier (ETC) had a longer list of features identified and no one feature was overwhelmingly important, although BMI of mother was important in the nearest neighbor RFC and the matched ETC models. This heterogeneity in RFC and ETC feature importance is perhaps expected because RFC and ETC are more complex and typically perform better than the Decision Tree Classifier.

Table 4. Machine learning results for stunting

	Matched	**Nearest Neighbor**
	Percent	**Percent**
Decision Tree Classifier	0.63 accuracy	0.63 accuracy
Age of child	42%	40%
Education level of mother	40%	40%
HH wealth	9%	11%
Random Forest Classifier	0.61 accuracy	0.64 accuracy
BMI of mother	21%	15%
Extra Trees Classifier	0.61 accuracy	0.63 accuracy
Disagrees with wife-beating (LGA)	11%	--
Joint household decision-making (LGA)	11%	--
Child dietary score	11%	--
HH age dependency ratio	11%	--
Age of child	--	8%

Table 5. Machine learning results for wasting

	Matched	Nearest Neighbor
	Percent	Percent
Decision Tree Classifier	0.85 accuracy	0.85 accuracy
Education level of mother	65%	65%
HH wealth	11%	11%
Mother works in agriculture	9%	11%
Random Forest Classifier	0.85 accuracy	0.85 accuracy
BMI of mother	--	24%
Disagrees with wife-beating (LGA)	10%	10%
Joint decision-making (LGA)	10%	10%
Child dietary diversity score	10%	10%
HH age dependency ratio	10%	10%
Extra Trees Classifier	0.84 accuracy	0.84 accuracy
BMI of mother	20%	--
Disagrees with wife-beating (LGA)	12%	12%
Joint decision-making (LGA)	12%	12%
Child dietary diversity score	12%	12%
HH age dependency ratio	11%	11%

DISCUSSION

This study aimed to analyze the drivers of stunting and wasting from a cross-sectional analysis of the 2008 Nigeria DHS, while incorporating additional nutrition-specific and nutrition-sensitive variables from the 2010–2011 Nigeria GHS-Panel survey. Linking these datasets allowed us to explore a wider range of potential drivers than is typically available in a single data source such as the DHS. In particular, agriculture-related variables are relevant to nutrition outcomes but are typically collected in focused agriculture surveys. By linking the NDHS and GHS-Panel data, we aimed to understand nutrition-sensitive and nutrition-specific drivers of stunting and wasting while also generating lessons learned for future analyses of nutrition outcomes using existing data. We also aimed to establish a proof of concept for linking additional nutrition-sensitive data to common nutrition datasets such as the DHS. Additionally, we compared traditional multivariate analysis with ML approaches to see whether ML provided any further insights over multivariate analysis.

The results of the multivariate analysis were broadly consistent with previous analyses of DHS data on stunting and wasting in the significant determinants identified. The LGA-level nutrition-sensitive variables added from the GHS-Panel were generally not statistically significant, and when they were the results did not necessarily align with expectations from the theoretical framework used to inform the analysis (e.g., market variable effects were in the opposite direction than expected). The significance of individual GHS-Panel variables also varied according to the model specification, whether using matched or nearest neighbor approaches. Overall, the additional nutrition-sensitive indicators from the GHS-Panel did not add any meaningful additional insights on the drivers of stunting and wasting in this analysis.

The ML analyses performed better for wasting than for stunting, but even for wasting the results were not particularly strong and were not significantly better than logistic regression. The feature importance of variables was highly dependent on the ML method employed and were not particularly meaningful conceptually.

The disappointing results from these analyses reflect at least in part the significant data challenges that we encountered when linking data from two surveys. Since this was an exploratory proof-of-concept analysis, we made several pragmatic decisions in defining variables and linking the datasets to allow the analysis to move forward within the defined time frame and resource envelope. These decisions allowed us to generate learning about the process but significantly limited the scientific strength of the substantive analysis. Consequently, this study was able to meet its second and third objectives better than its first one. Below we summarize the data limitations and learning generated about the process and the implications for the substantive results.

Limitations and Lessons Learned

Temporality and Availability of Datasets

The GHS-Panel data were collected two years after the NDHS, so we had to treat the nutrition-sensitive variables from the 2010–2011 GHS-Panel as proxies for the variables in 2008 when the NDHS was collected. This is a strong assumption given annual variation in agricultural and other conditions, which effectively introduces measurement error into the variables from the GHS-Panel data and undermines the substantive findings from this analysis. The differential timing of data collection from different data sources is a significant challenge in using existing data from multiple surveys. In selecting countries for this study, we found that it was difficult to identify countries that had existing datasets for the same time periods that could be linked to provide both nutrition-sensitive and nutrition-specific variables for analysis.

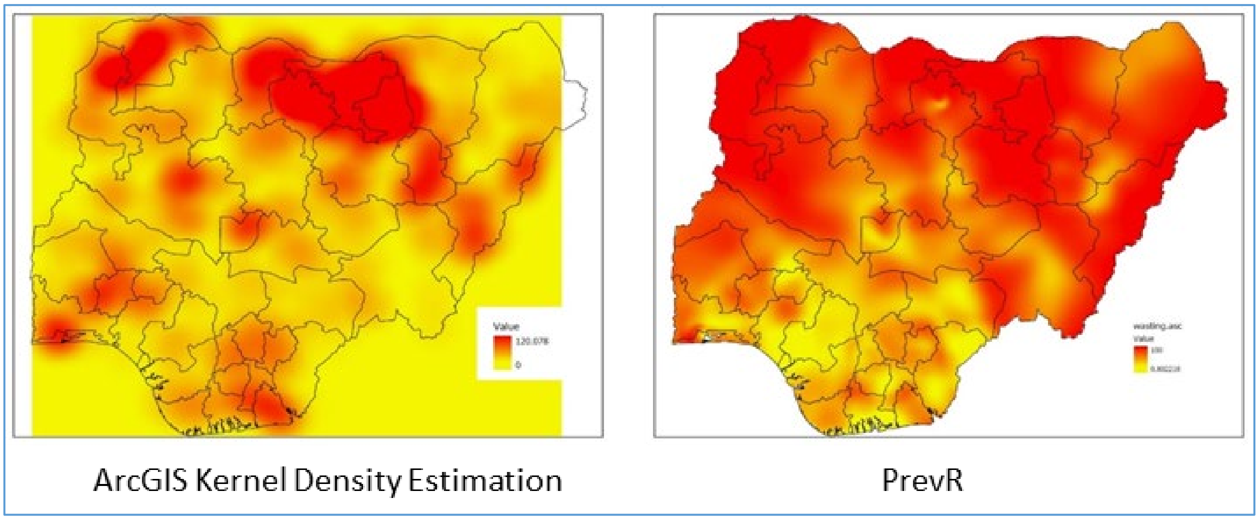
We sought to identify additional datasets for Nigeria to bring in other relevant factors to the analysis, such as climate variables. Some national climate data are available, although all were more recent than 2008. In addition, climate is highly correlated with specific regions in Nigeria, which is also correlated with cultural and economic factors that are expected to have a strong impact on stunting and wasting. Given these concerns over the temporality of relevant climate data with the NDHS, and of potential confounding between climatic and other cultural and economic factors in the Nigerian context, in addition to given timelines, we decided not to pursue including climate datasets in this analysis. There are relevant global and continental datasets available from around 2008; with additional time and resources these datasets may add value to this analysis. For example, in a recent paper, Tusting, et al. (2020) linked climate data collected through remote sensing to DHS data across Africa to examine the association of environmental temperature with growth faltering.

Aggregation and Linking of Data

The Nigeria DHS and the GHS-panel survey are independently sampled national surveys, so the data collected in each survey refer to different households. To address this issue, we need to have a common unit at which to link the data. We linked the surveys spatially at the LGA level by aggregating data from the GHS-Panel survey to the LGA level and linking the aggregated LGA-level variables to the individual data from the NDHS. The LGA was the lowest geographic level of aggregation feasible, but an LGA is still a sizeable geographic area. Further, neither the NDHS nor the GHS-Panel survey were designed to produce statistically reliable estimates at the LGA level; sometimes a single GHS-Panel cluster represented that entire LGA. This means that there is considerable sampling error associated with the aggregated GHS-panel variables. If LGA measures of nutrition-sensitive indicators from the GHS-Panel data do not well-represent the local conditions faced by the NDHS households, then the conclusions drawn will not be internally valid. Even when using the LGA level for aggregation only about 40 percent of LGAs had both a NDHS and a GHS-Panel cluster, so for those NDHS clusters that were located in an LGA that did not contain at least one GHS-Panel cluster we had to aggregate GHS-Panel data from adjacent clusters to link to the NDHS cluster. This further exacerbates the internal validity issues that arise from assuming that the aggregated GHS-Panel variables represent the conditions faced by NDHS households. Skiles, et al. (2013) found that linking independent samples from health facility surveys and DHS surveys resulted in substantial misclassification errors in aggregated facility variables linked to DHS clusters, resulting in biases in analysis. The challenges faced in this analysis reinforce their conclusion that linking data from independently sampled surveys is problematic.

We used a simple approach to aggregate data from the GHS-Panel survey to the LGA level as a starting point for the analysis. A more sophisticated approach would be to use statistical models to generate spatial surfaces that allow for more geographically granular estimates of variables. Such models have been used to generate spatial surfaces for a number of variables by the DHS program (Gething, et al., 2015; Mayala, et al., 2019), the Institute for Health Metrics and Evaluation (Osgood-Zimmerman, et al., 2018), and other groups such as WorldPop at the University of Southampton (https://www.worldpop.org/). In principle, it is possible to generate modeled values of the nutrition-sensitive variables of interest from the GHS-Panel survey for the geographic locations of the DHS clusters and then use those modeled values for the analysis. However, surfaces are time-intensive to create and sensitive to specific methods and the parameters used when generating the surface. For example, we generated two different surfaces for wasting from the 2008 NDHS data using two different interpolation methods with similar parameters and the results were not consistent (Figure 4). Typically, a variety of surface parameters are tested and evaluated prior to use of the surface in analysis; doing this for multiple different variables would be time intensive. A further consideration is that spatial interpolation methods are most suitable for variables that display strong spatial correlation. Not all the variables of interest in the GHS-Panel survey for this analysis will be strongly spatially correlated, so spatially interpolated surfaces will make more sense for some variables of interest than for others. Finally, modeled estimates are still subject to uncertainty so using modeled estimates does not eliminate measurement error in the GHS-Panel variables linked to the NDHS data; the sparser the data, the greater the uncertainty will be in the modeled estimates. Surfaces arguably add their greatest value when combined with either other surfaces or with covariates at a scale that is finer than the GHS-Panel variables. There may be value in creating surfaces for selected GHS-Panel variables to improve linkage with the NDHS, but this would require significant further investment in time and resources and the costs of creating the surfaces for this case study may outweigh their benefit given other data limitations.

Figure 4. Surface for wasting; comparison of ArcGIS Kernel Density Estimation and PrevR



Variable Specifications

The conceptual framework we used to guide this analysis includes a wide range of different factors that are hypothesized to influence nutritional outcomes. The NDHS and the GHS-Panel survey contain many questions relevant to different elements in the conceptual framework, but they did not always readily translate into variables that were usable for this analysis. For example:

- Some questions apply only to certain subgroups in the samples, so they cannot be readily included in an analysis that includes the full sample. For example, in the DHS current breastfeeding status is only asked for children under two years old, and dietary diversity is only typically asked for children ages 6–23 months old. We had to make some compromises to create derived variables for this analysis.
- Some variables we wanted to include in the analysis based on the conceptual framework have only been included in later waves of the DHS surveys, such as micronutrient supplementation for children or supplementary feeding with a ready-to-use supplementary food, so these could not be included in this analysis.

- The GHS-Panel survey had two time points (post-planting and post-harvest) and only included certain questions in each wave. Variables to measure household-level food security were only available post-harvest, which is when food is most abundant, and showed little evidence of food security problems and limited variation; thus, they were of limited value for this analysis.
- There were multiple questions in the GHS-Panel survey on nutrition-sensitive factors such as crop diversity, household receipt of remittances, financial inclusion, and livelihood diversification among others, but it was challenging to aggregate them in a meaningful way for this analysis. Significant investment of time would be required to test different specifications of aggregated variables to determine which specification, if any, would be most meaningful and viable for analysis. Given time constraints, we did not include variables that were complex to aggregate in these models.

All statistical analyses require choices in how variables and models are defined. The analysis we conducted could certainly be improved by refining the variable and model specifications in various ways but doing so would require further time investment. Given the significant data limitations already described, the costs of further investment in refining the variables and models for this analysis would likely outweigh the benefits. In other analyses, more attention to variable specifications would be warranted but there will be limitations in the extent to which relevant questions can be operationalized into meaningful variables for specific analyses.

Machine Learning

The ML analysis did not add significant insight into the factors associated with stunting and wasting in this analysis. However, this analysis employed the most basic RFC and ETC. Best practice would be to refine the ML models by doing feature engineering, as well as adjusting decision tree parameters that could improve the model. Feature engineering is using domain knowledge to develop variables that might serve the model better than raw data. An example would be using data on types of livestock kept by the household to create a variable, such as "keeps poultry," instead of individual variables such as "keeps chickens," "keeps ducks," etc. (similar to refining the variable specifications discussed above). However, given the other data limitations described and the disappointing initial results, we did not feel that further investment of time to refine the models would add an appreciable amount of learning value.

Machine learning is an empirical approach that uses correlation within the data to predict outcomes. As such, it is not designed to identify causal relationships, so it is perhaps not surprising that it was not a good fit for this application in assessing the drivers of stunting and wasting. ML may have more potential to address data gaps by generating improved model estimates of variables for use in multivariate analysis (see also the previous section on using models to improve linking of datasets).

RECOMMENDATIONS

Below we summarize recommendations for future work linking data from different sources to explore drivers of nutritional outcomes based on the lessons learned from this study.

1) We do not recommend linking data from two or more independent sample surveys at low levels of geographic disaggregation such as the LGA. Linking appears more likely to work well when linking a sample survey such as DHS to a dataset that has complete geographic coverage, such as remote sensed climate data (e.g., Tusting et al., 2020). Such an approach would, however, likely limit the range of additional variables that could be included to better understand drivers of the outcomes being examined..
2) It could be worth further exploring linking sample survey data to modeled surfaces obtained from other data sources, but we recommend focusing on using one modeled surface initially to test the idea. Using modeled surfaces won't eliminate measurement error concerns for variables at low levels of geographic disaggregation and will work better for some variables than others depending on the precision of the models. It is time-intensive to create modeled surfaces so using existing publicly available models will be more efficient. However, the existing models we are aware of tend to focus on outcome variables such as stunting, rather than the nutrition-specific independent variables which were the focus of this analysis.
3) For some research questions, linking data at higher levels of geographical disaggregation, such as a geographical region, could be sufficient for meaningful analysis. For example, individual data from a DHS survey could be linked to aggregate variables at the regional level from an agricultural survey that provides adequate statistical precision for regional level estimates. Whether this is meaningful will depend on whether the regional level is granular enough to be relevant conceptually to the question of interest.

In all linked analysis the different data need to refer to the same time period, or to be otherwise meaningfully related temporally.

CONCLUSION

The increased availability of data from multiple sources in low- and middle-income countries in recent years, combined with advances in data science, have stimulated an increased interest in using existing data in innovative ways to bring new insights to population, health, and nutrition problems. This study has generated learning on the process and challenges involved in bringing together data from different sectors to analyze the drivers of stunting and wasting in Nigeria. We identified several significant data challenges that undermine the credibility of the substantive findings and were unable to demonstrate that the proposed approach was feasible in this case study. Similar challenges are likely to be present in other contexts. Some of the approaches we explored for this analysis could be useful in other applications, however.

New data science methods are not a panacea for underlying data problems such as sparse or poor quality data. Use of existing data requires compromises to accommodate the data available, which can undermine the credibility of the substantive results. More sophisticated methods can require substantial investments of time from highly specialized analysts, which will have significant cost implications. Using existing data will not necessarily be cheap and quick either. Although we have more and better data now than ever before, in addition to more and better ways of analyzing them, challenges remain to linking data for complex analyses and to applying ML in this context.

REFERENCES

Akombi, B.J., Agho, K.E., Hall, J.J., Wali, N., Renzaho, A.M.N., & Merom, D. (2017). Stunting, Wasting and Underweight in Sub-Saharan Africa: A Systematic Review. *International Journal of Environmental Research and Public Health*, *14*(8).

Richard, S.A., Black, R.E., Gilman, R.H., Guerrant, R.L., Kang, G., Lanata, C.F., Mølbak, K., et al. (2012). Wasting is associated with stunting in early childhood. *The Journal of Nutrition,* 142(7): 1291–1296.

Black, R.E., Victora, C.G., Walker, S.P., Bhutta, Z.A., Christian P, et al. (2013). Maternal and child undernutrition and overweight in low-income and middle-income countries. *Lancet*; 382: 427–51.

Gething, P., Tatem, A., Bird, T., & Burgert-Brucker, C.R. (2015). Creating Spatial Interpolation Surfaces with DHS Data DHS Spatial Analysis Reports No. 11. Rockville, Maryland, USA: ICF International.

Martorell, R., & Young, M.F. (2012). Patterns of stunting and wasting: potential explanatory factors. *Advances in nutrition (Bethesda, Md.)*, *3*(2): 227–233.

Mayala, B.K., Dontamsetti, T., Fish, T.D., & Croft, T.N. (2019). Interpolation of DHS Survey Data at Subnational Administrative Level 2. DHS Spatial Analysis Reports No. 17. Rockville, Maryland, USA: ICF. https://dhsprogram.com/pubs/pdf/SAR17/SAR17.pdf

Motbainor, A., Worku, A., & Kumie, A. (2015). Stunting Is Associated with Food Diversity while Wasting with Food Insecurity among Underfive Children in East and West Gojjam Zones of Amhara Region, Ethiopia. *Plos One*, *10*(8), e0133542.

Osgood-Zimmerman A., Millear, A.I., Stubbs, R.W., Shields, C., Pickering, B.V., Earl, L., Graetz, N., Kinyoki, D.K., Ray, S.E., Bhatt, S., Browne, A.J., Burstein, R., Cameron, E., Casey, D.C., Deshpande, A., Fullman, N., Gething, P.W., Gibson, H.S., Henry, N.J., Herrero, M., Krause, L.K., Letourneau, I.D., Levine, A.J., Liu, P.Y., Longbottom, J., Mayala, B.K., Mosser, J.F., Noor, A.M., Pigott, D.M., Piwoz, E.G., Rao, P., Rawat, R., Reiner, R.C., Smith, D.L., Weiss, D.J., Wiens, K.E., Mokdad, A.H., Lim, S.S., Murray, C.J.L., Kassebaum, N.J., Hay, S.I. (2018). Mapping child growth failure in Africa between 2000 and 2015**.** *Nature*, 28 Feb 2018. doi:10.1038/nature25760. https://www.nature.com/articles/nature25760.pdf

Panjwani, A., & Heidkamp, R. (2017). Complementary Feeding Interventions Have a Small but Significant Impact on Linear and Ponderal Growth of Children in Low- and Middle-Income Countries: A Systematic Review and Meta-Analysis. *The Journal of Nutrition*, *147*(11): 2169S–2178S.

Ricci, J.A., & Becker, S. (1996). Risk factors for wasting and stunting among children in Metro Cebu, Philippines. *The American Journal of Clinical Nutrition*, *63*(6): 966–975.

Roba, K.T., O'Connor, T.P., Belachew, T., & O'Brien, N.M. (2016). Variations between post- and pre-harvest seasons in stunting, wasting, and Infant and Young Child Feeding (IYCF) practices among children 6-23 months of age in lowland and midland agro-ecological zones of rural Ethiopia. *The Pan African medical journal*, *24*: 163.

Skiles, M.P., Burgert, C., Curtis, S.L., Spencer, J. (2013). Geographically linking population and facility surveys: Methodological considerations. *Popul Health Metr*. 2013:11(1):14. doi: 10.1186/1478-7954-11-14.

Tusting, L.S., Bradley, J., Bhatt, S., Gibson, H.S., Weiss, D.J., Shenton, F.C., & Lindsay, S.W. (2020). Environmental temperature and growth faltering in African children: a cross-sectional study. *Lancet PlanetHealth* 2020; 4:e116-23. https://www.thelancet.com/journals/lanplh/article/PIIS2542-5196(20)30037-1/fulltext

UNICEF/ WHO/World Bank Group. (2019). Levels and trends in child malnutrition: Joint Child Malnutrition Estimates Key findings of the 2019 edition. Accessed June 24, 2019. Retrieved from https://www.who.int/nutgrowthdb/jme-2019-key-findings.pdf?ua=1

Young, A., Marshak, A. (2018) Persistent Global Acute Malnutrition. Retrieved from https://fic.tufts.edu/assets/FIC-Publication-Persistent-Global-Acute-Malnutrition_web_2.26s.pdf

APPENDIX. A. COUNTRY SELECTION PROCESS

As a first step in the country selection process, we examined all countries with multiple rounds of DHS data to identify trends in stunting and wasting patterns across countries. Trends for wasting and stunting were assessed using the DHS STATCompiler[3] over the past several survey waves. For the 57 countries with more than one measurement over time, countries were categorized into whether wasting and stunting were declining, constant, or increasing.[4] The results of this assessment are summarized below and in Table A1. Seventeen countries with only one measurement in time were excluded from the assessment.

Table A1 shows fifty-seven countries with available DHS data classified by their trends in wasting and stunting. Of particular interest are countries where stunting is declining but wasting is either constant or increasing, and where wasting prevalence exceeds the crisis threshold of 10 percent. Countries with declining stunting and constant wasting include Bangladesh, Burkina Faso, Comoros, Dominican Republic, Guatemala, Honduras, Jordan, and Senegal. Countries with declining stunting and increasing wasting include Armenia, Brazil, Egypt, India, Madagascar, Morocco, Nigeria, and Timor-Leste. Seven out of twelve, or 58 percent, of Feed the Future (FTF) countries have either constant or increasing wasting trends, while twenty-one out of fifty-seven, or 37 percent, of total countries represented have either constant or increasing wasting trends. FTF countries with 10 percent or greater prevalence of wasting include Bangladesh, Niger, Nigeria, and Mali. Just three countries show an increasing trend in stunting—Burundi, Mali, and Sierra Leone. Four countries exhibit constant trends in both stunting and wasting, specifically Guinea, Niger, Yemen, and Zambia. Countries with greater than 10 percent prevalence of wasting appear in bold in the table and FTF countries are denoted with an asterisk. Following Table A1 are graphical depictions of these trends for FTF countries.

We also sought to select a country case study with other datasets available for linking nutrition-sensitive variables for the analysis. Tables A2 and A3 provide inventories of the search for additional datasets and what was identified for this work. Nigeria was selected as the country with the desired trend, decreasing stunting and increasing or constant wasting, with at least a 10 percent prevalence of wasting, and with other datasets available for possible use.

Table A1. Trends in stunting and wasting, DHS STATCompiler 1986–2017

		Wasting**		
		Declining	**No change**	**Increasing**
Stunting	**Declining**	Albania	**Bangladesh***	Armenia
		Bolivia	**Burkina Faso**	Brazil
		Cambodia	**Comoros**	Egypt
		Cameroon	Dominican Republic	**India**
		Chad	Guatemala*	**Madagascar**
		Colombia	Honduras*	**Morocco**
		Congo	Jordan	**Nigeria***
		Congo, Democratic Republic	Senegal*	**Timor-Leste**
		Eritrea		
		Ethiopia*		
		Gabon		

[3] https://dhsprogram.com/data/STATcompiler.cfm

[4] For this exploratory analysis, the assessment of whether the trend is declining, constant, or increasing was based on visual inspection of graphs of the trends and not on statistical significance tests. Statistical analysis of trend data would require additional information from DHS.

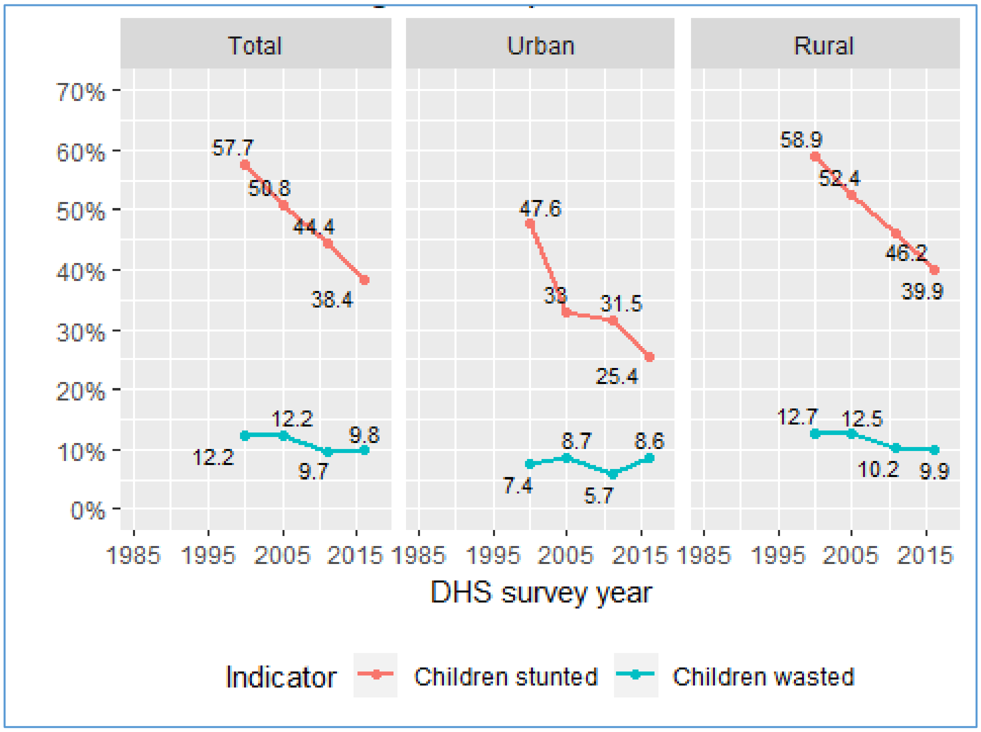
		Wasting**		
		Declining	**No change**	**Increasing**
		Ghana*		
		Haiti		
		Kazakhstan		
		Kenya*		
		Kyrgyz Republic		
		Lesotho		
		Liberia		
		Malawi		
		Maldives		
		Namibia		
		Nepal*		
		Nicaragua		
		Pakistan		
		Peru		
		Rwanda		
		Tajikistan		
		Tanzania		
		Togo		
		Turkey		
		Uganda*		
	No change	Benin	Guinea	
		Cote d'Ivoire	**Niger***	
		Mozambique	**Yemen**	
			Zambia	
	Increasing	Burundi	**Mali***	
		Sierra Leone		

Note: Several countries with only one time point were excluded from this analysis, including Angola, Azerbaijan, Central African Republic, Eswatini, **Gambia**, Guyana, **Mauritania**, Moldova, Myanmar, Paraguay, **Sao Tome and Principe**, South Africa, **Sri Lanka**, Thailand, Trinidad and Tobago, Tunisia, and **Uzbekistan.**

*Feed the Future target country. **Countries with greater than 10 percent wasting prevalence are in bold.

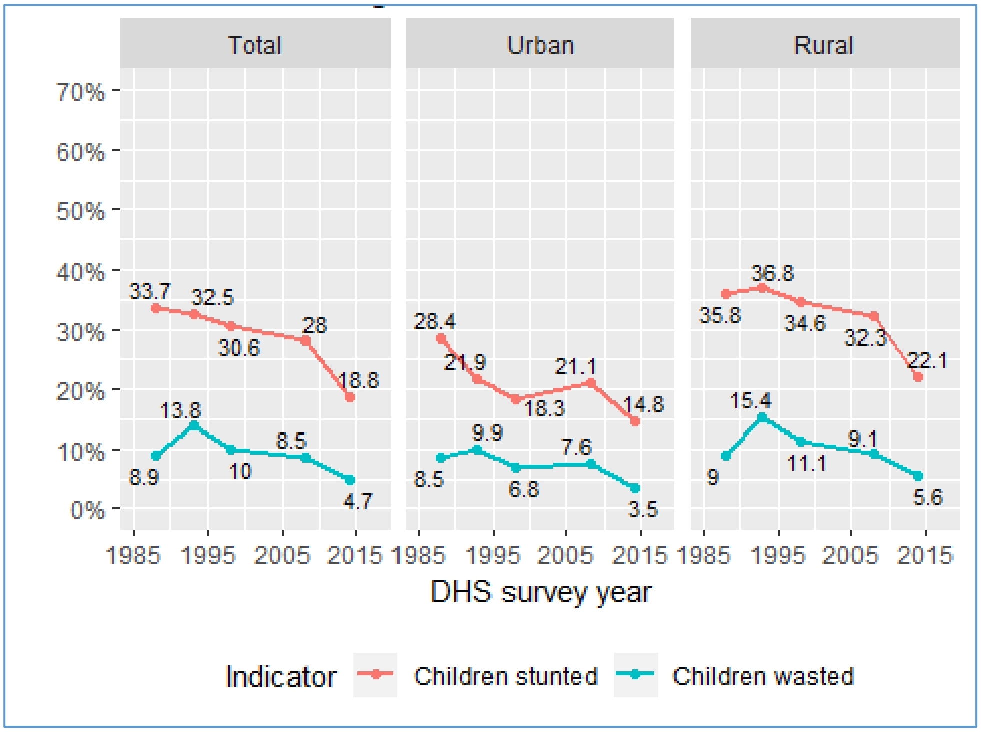
Declining Stunting and Declining Wasting

Figure A1. Wasting and stunting trends in Ethiopia



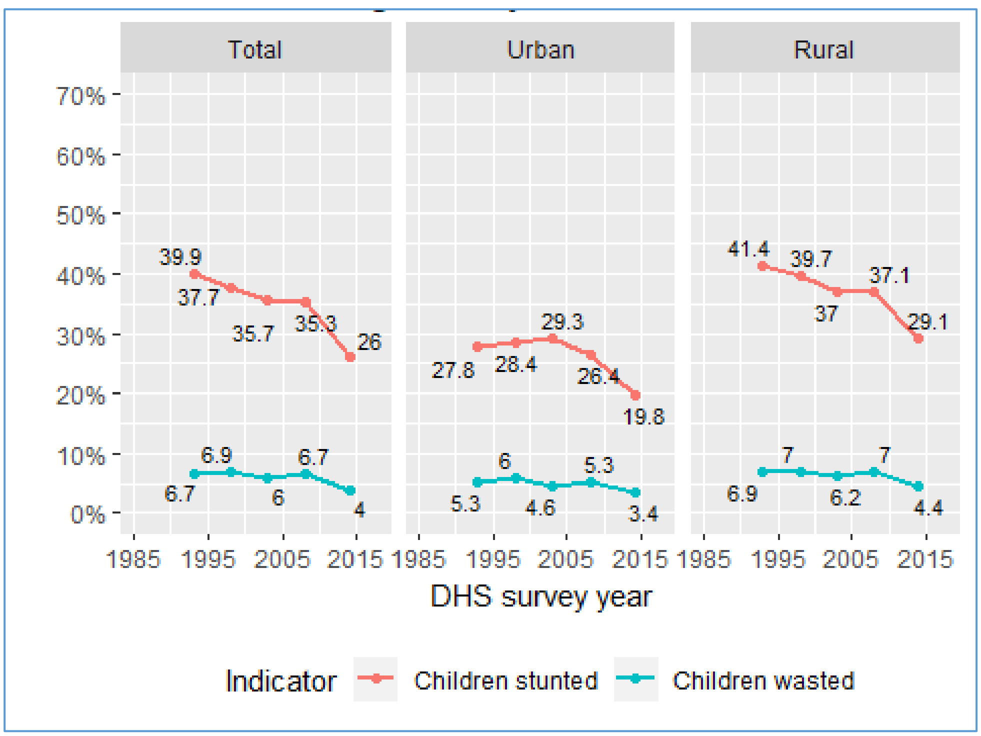
Note: Denominator is children under 5 in the sample household (HH).

Figure A2. Wasting and stunting trends in Ghana



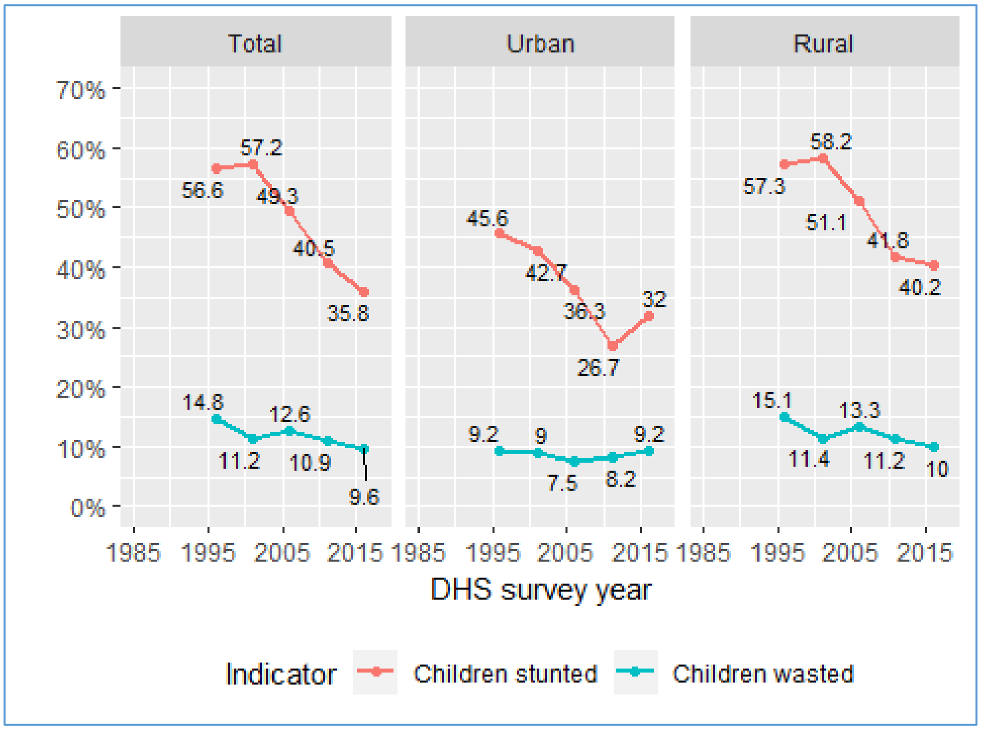
Note: Denominator for 1988–1998 varies; denominator for 2003–2014 is children under 5 in the sample HH.

Figure A3. Wasting and stunting trends in Kenya



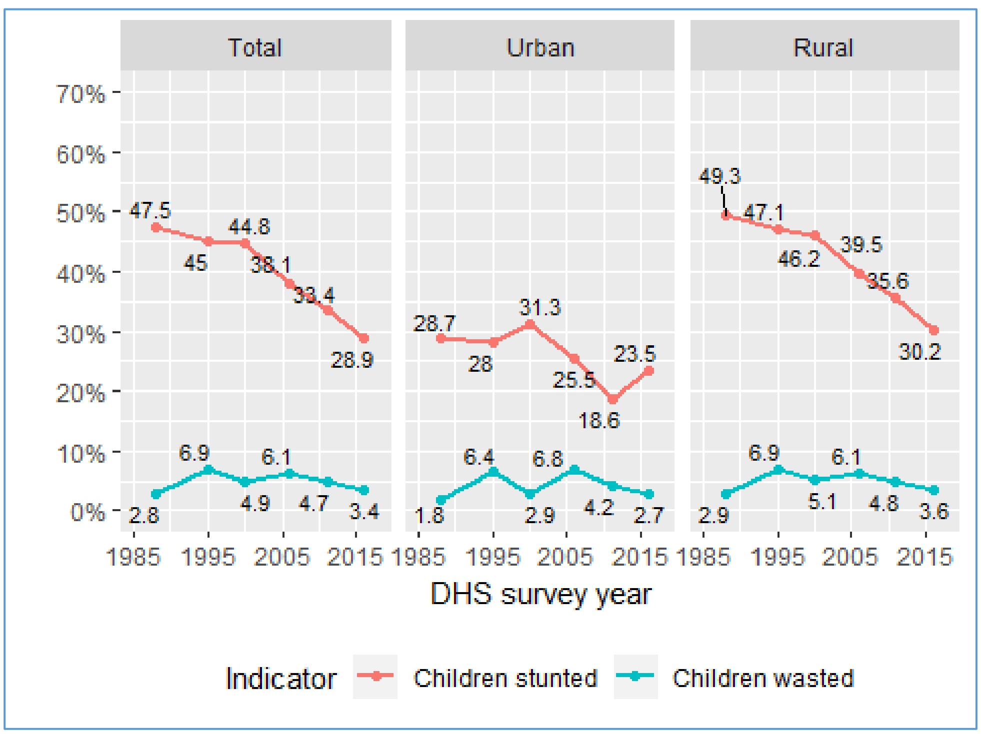
Note: Denominator for 1988–1998 varies; denominator for 2003–2014 is children under 5 in the sample HH.

Figure A4. Wasting and stunting trends in Nepal



Note: Denominator for 1996–2001 varies; denominator for 2006–2016 is children under 5 in the sample HH.

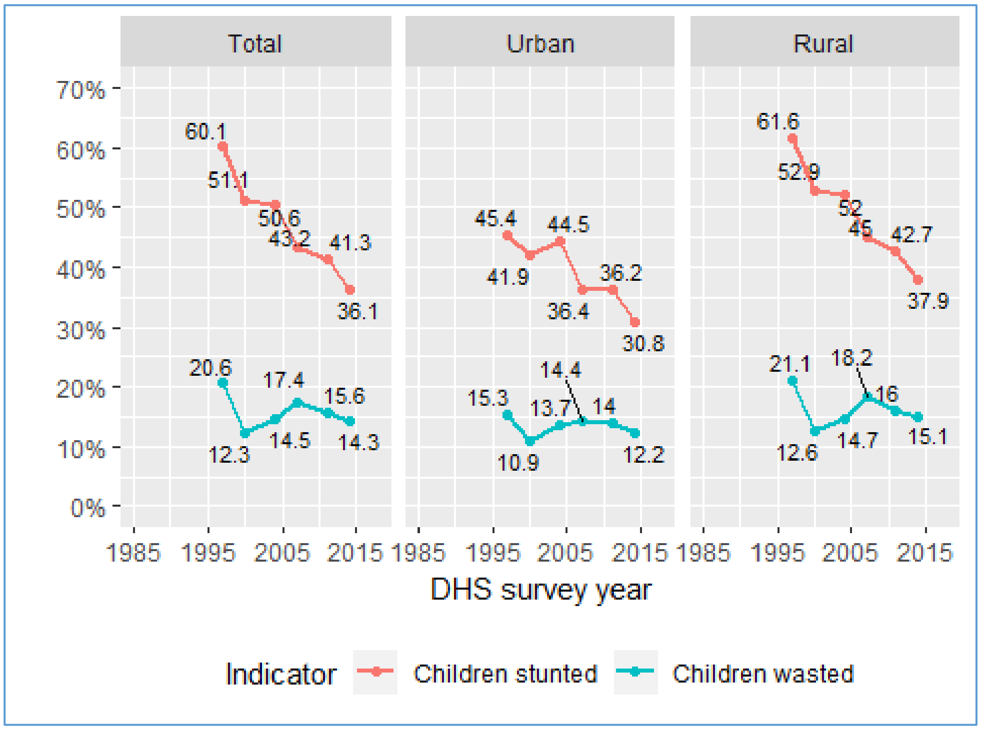
Figure A5. Wasting and stunting trends in Uganda



Note: Denominator for 1988–2001 varies; denominator for 2006–2016 is children under 5 in the sample HH.

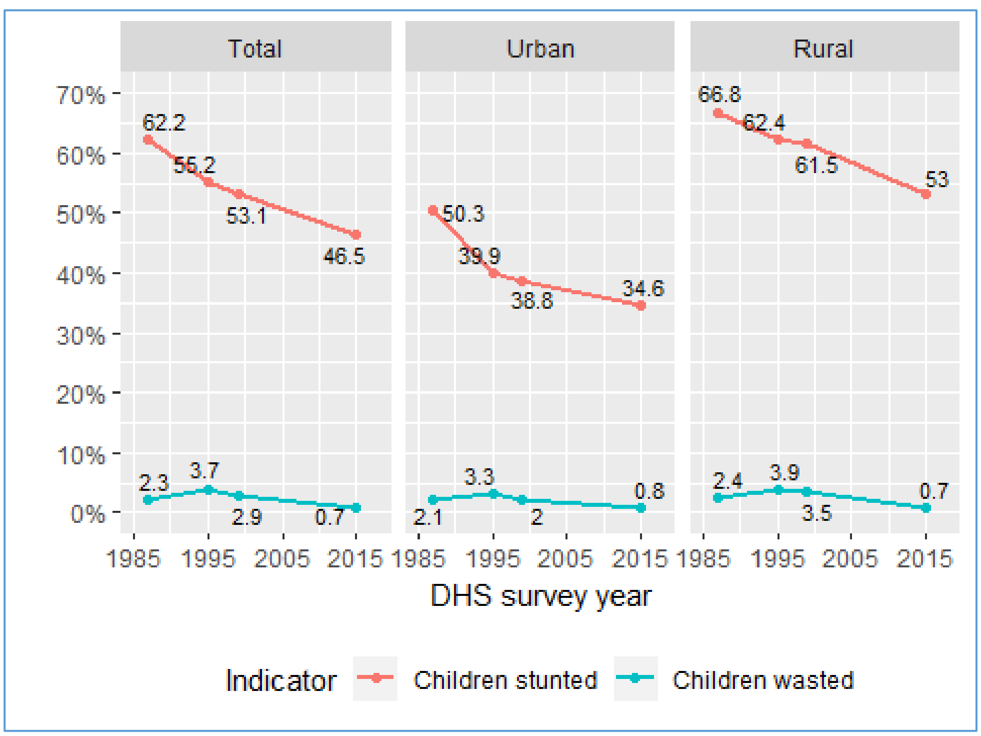
Declining Stunting and No Change in Wasting

Figure A6. Wasting and stunting trends in Bangladesh



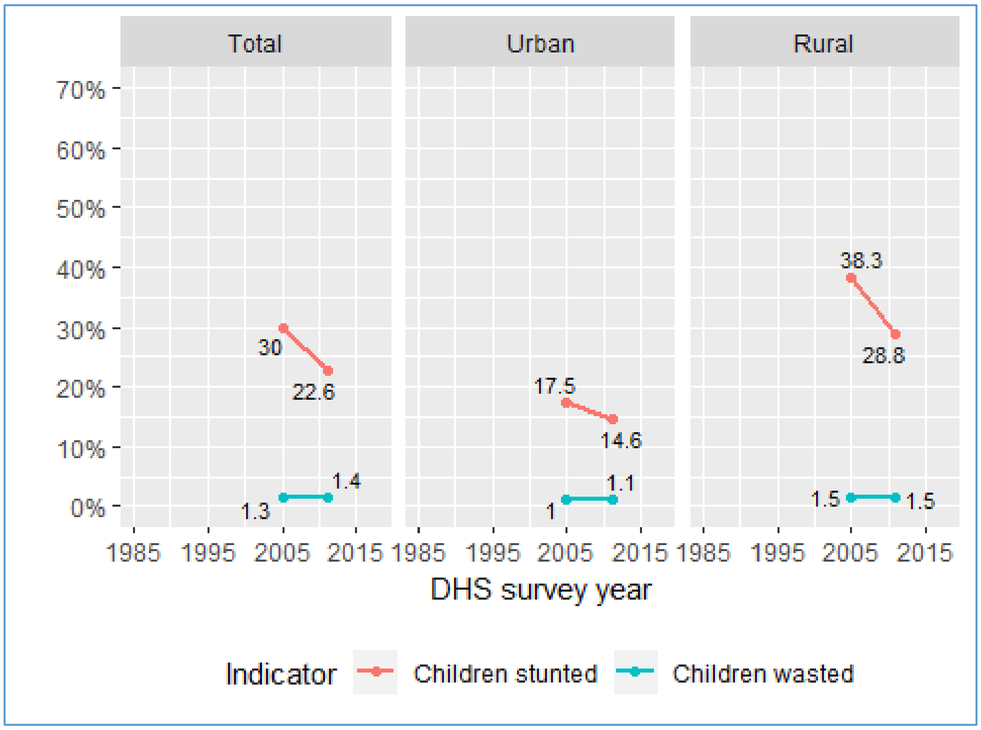
Note: Denominator for 1996–2000 varies; denominator for 2007–2014 is children under 5 in the sample HH.

Figure A7. Wasting and stunting trends in Guatemala



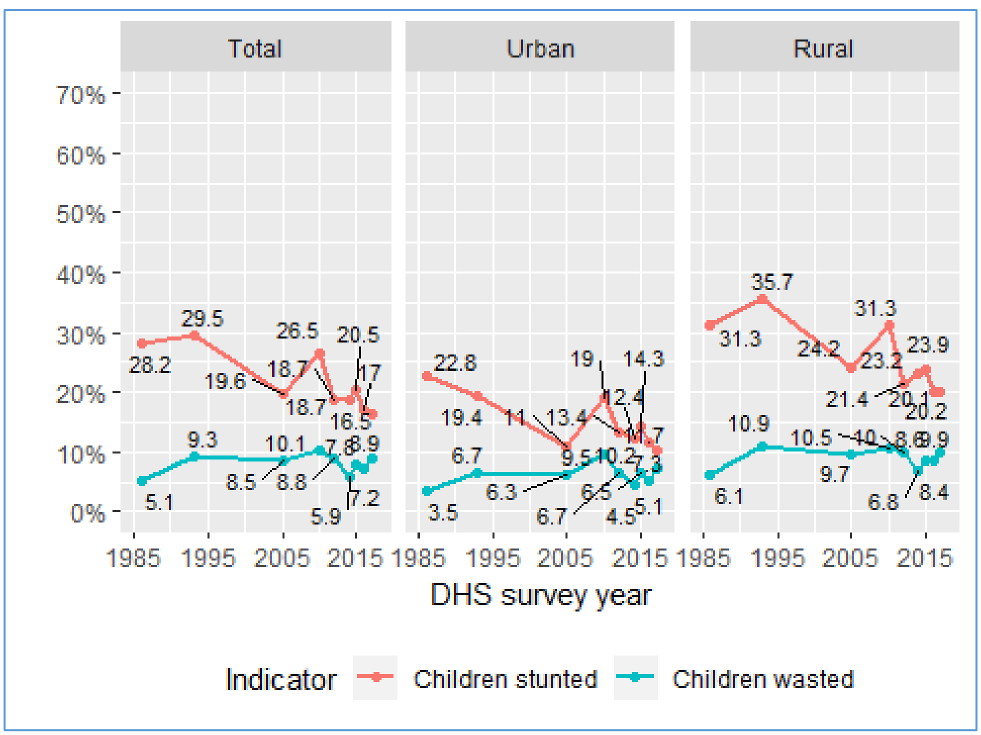
Note: Denominator for 1987–1999 varies; denominator for 2014–2015 is children under 5 in the sample HH.

Figure A8. Wasting and stunting trends in Honduras



Note: Denominator is children under 5 in the sample HH.

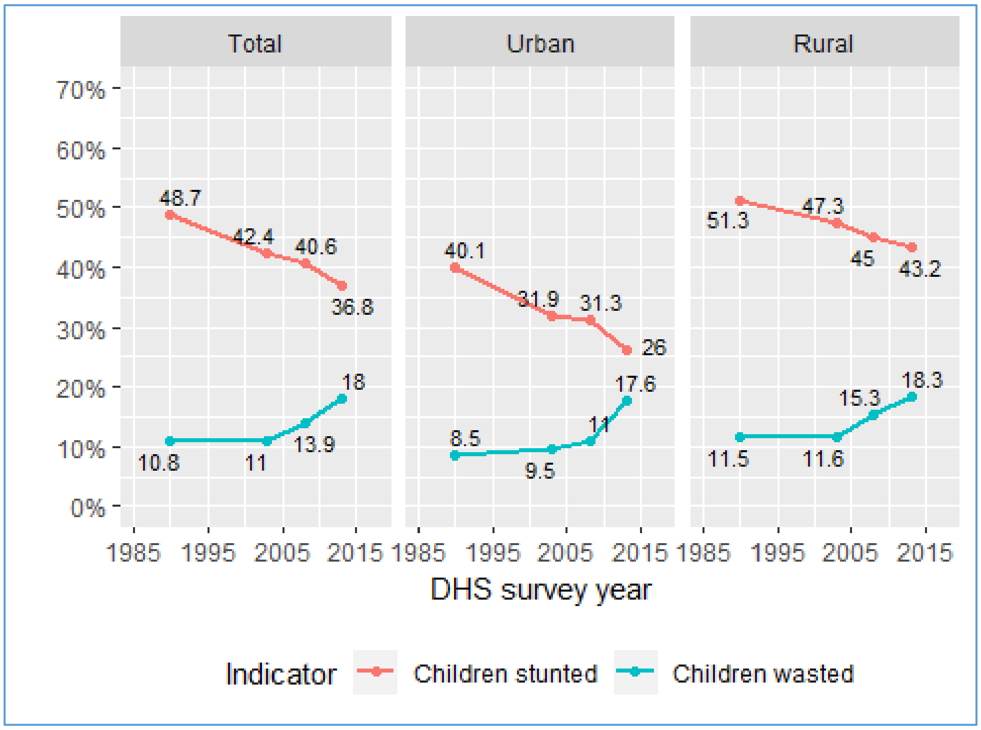
Figure A9. Wasting and stunting trends in Senegal



Note: Denominator for 1986–1993 varies; denominator for 2010–2017 is children under 5 in the sample HH.

Declining Stunting and Increasing Wasting

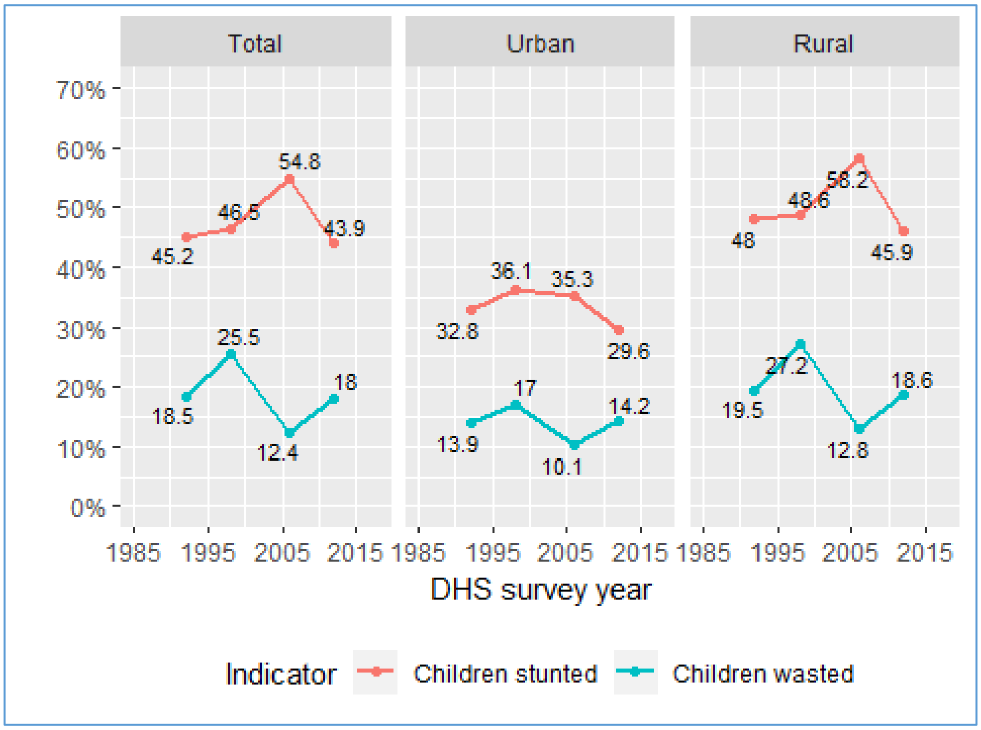
Figure A10. Wasting and stunting trends in Nigeria



Note: Denominator for 1990–1999 varies; denominator for 2003–2013 is children under 5 in the sample HH.

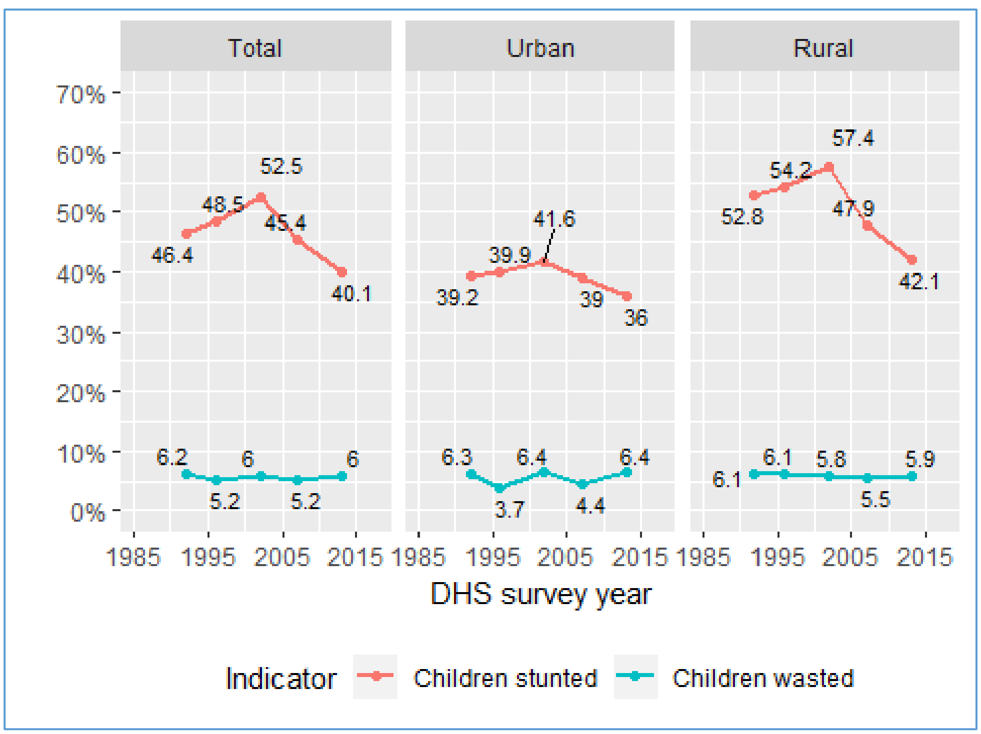
No Change in Stunting and No Change in Wasting

Figure A11. Wasting and stunting trends in Niger



Note: Denominator for 1992–2006 varies; denominator for 2012 is children under 5 in the sample HH.

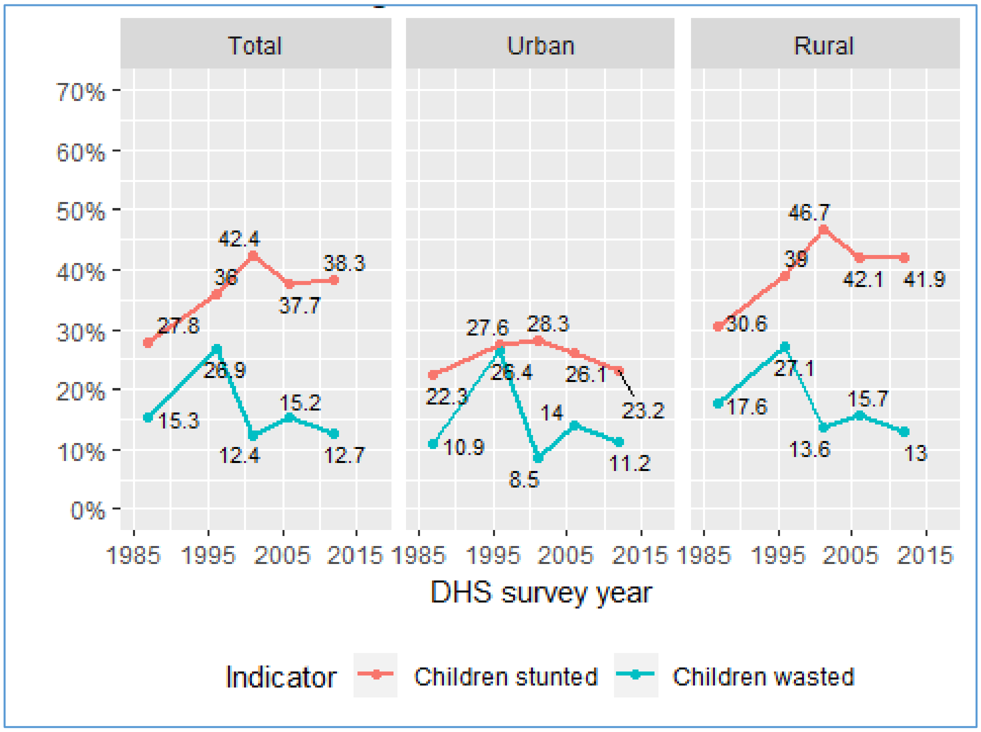
Figure A12. Wasting and stunting trends in Zambia



Note: Denominator for 1992–1996 varies; denominator for 2001–2014 is children under 5 in the sample HH.

Increasing Stunting and No Change in Wasting

Figure A13. Wasting and stunting trends in Mali



Note: Denominator for 1987–2006 varies; denominator for 2012–2013 is children under 5 in the sample HH.

Table A2. Identified food security databases

	Geography	Year	Notes
Food consumption scores	Subnational for 38 countries	2006–2015	The Food Consumption Score (FCS) dataset is based on the FCS indicator, which assigns a food security score based on food consumption and diets. This data is available subnationally for 38 countries.
Food insecurity mapping	Sudan, Ethiopia, Somalia, Uganda, Kenya	2013–2016	Most likely food security outcome for January-March Most likely food security outcome for April–June (May have limited utility)
Global Food Prices Database	Global	~2003	This dataset contains Global Food Prices data from the World Food Programme covering foods such as maize, rice, beans, fish, and sugar for 76 countries and some 1,500 markets. It is updated weekly but contains, to a large extent, monthly data. The data goes back as far as 1992 for a few countries, although many countries started reporting from 2003 or thereafter.
Eastern Africa food prices data	1st admin unit	Varies by country ~2002	Food and input prices of selected Association for Strengthening Agricultural Research in Eastern and Central Africa (ASARECA) member countries,
Food supply	National	Varies	Food supply – total quantity (kg/capita/yr) • Cereals, Meat, Fish, Eggs, Milk, etc. Total energy: • Energy from cereals, meat, fish/ seafood, etc. Protein/fat supply: • Protein/fat from cereals, meat, fish, starchy roots, etc.

Table A3. Inventory of spatial data sets obtained for possible use in Nigeria

Title	Timeframe	Scale	Description
Road network	2012	National	Road network taken from Digital Chart of World. Roads are in three categories: • Paved federal roads • Other paved roads • Other roads/tracks
Poverty – $2 a day	2015	1km^2	Worldpop Alpha version 2010 estimates of proportion of people per grid square living in poverty, as defined by USD $1.25 a day and $2 a day thresholds, and associated uncertainty metrics.
Poverty $1.25 a day	2015	1km^2	Worldpop Alpha version 2010 estimates of proportion of people per grid square living in poverty, as defined by USD $1.25 a day and $2 a day thresholds, and associated uncertainty metrics.
Global exposure	2015	1km^2	The GAR15 global exposure database is based on a top-down approach where statistical information, including socioeconomic, building type, and capital stock at a national level, are transposed onto the grids of 5x5 or 1x1 using geographic distribution of population data and gross domestic product (GDP) as proxies. This database includes estimation on the economic value of the exposed assets, as well as their physical characteristics in urban and rural agglomerations, including estimation of population.
Electrical grid	2018	1km^2	Building electrical grid maps begins by taking monthly images from the VIIRS satellite, and creating a composite. We then apply a custom image processing filter to remove background and reflected light, and identify locations that consistently demonstrate night-time lighting. These then serve as a proxy for the existence of grid electricity. Using known electrical grids as templates, based on data available from energydata.info, we employ a custom algorithm to connect the communities and infer grid paths based on their likelihood to follow roads, avoid water, and follow the shortest paths possible. Dataset created by Facebook
High resolution population	2018	30m^2?	Using a mixture of machine learning techniques, high-resolution satellite imagery, and population data, hundreds of millions of structures distributed across vast areas were mapped and then used to extrapolate and calculate local population density. Dataset created by Facebook

High resolution population from European Commission	1975, 1990, 2000, 2015	250m, 1km	This spatial raster dataset depicts the distribution and density of population, expressed as the number of people per cell. Residential population estimates for target years 1975, 1990, 2000, and 2015 provided by CIESIN GPWv4 were disaggregated from census or administrative units to grid cells, informed by the distribution and density of built-up as mapped in the Global Human Settlement Layer (GHSL) global layer per corresponding epoch.
Urban centers	1975, 1990, 2000, 2015	Global	The database represents the global status on urban centers in 2015 by offering cities' locations, their extent (surface, shape), and describing each city with a set of geographical, socioeconomic and environmental attributes, many of them dating back 25 or even 40 years in time. Urban centers are defined in a consistent way across geographical locations and over time; attributes on elevation, biome, climate, soil, river basin, temperature, population, built-up area, GDP, development, accessibility/ remoteness, greenness, pollution, flood exposure, earthquake risk, storm surge, heatwave, land use efficiency, and open spaces are included.
Built-up grid	1975, 1990, 2000, 2014	38m, 250m, 1km	The database includes multitemporal information layer on built-up presence as derived from Landsat image collections, and shows growth that occurred between epochs: • land not built-up in any epoch • built-up from 2000 to 2014 epoch • built-up from 1990 to 2000 epoch • built-up from 1975 to 1990 epoch • built-up up to 1975 epoch
Health facility locations	2019	National	Location of health facilities obtained from the Global Healthsites Mapping Project. Locations derived from institutional sharing and social media/crowdsourcing.

APPENDIX B. COMPLETE MULTIVARIATE RESULTS

Table B1. Complete multivariate results assessing drivers of wasting and stunting in Nigeria, NDHS 2008 and GHS-Panel 2010/2011

	Level of indicator	Stunting			Wasting		
		Panel 1	Panel 2	Panel 3	Panel 4	Panel 5	Panel 6
			Matched	Nearest Neighbor matching		Matched	Nearest Neighbor matching
		NDHS	NDHS+GHS	NDHS+GHS	NDHS	NDHS+GHS	NDHS+GHS
Sex of child (NDHS)							
Female	CH	--	--	--	--	--	--
Male	CH	**0.039****	**0.040****	**0.044****	0.012	0.009	**0.013***
Mother's age (NDHS)							
Age 15–19 years	MO	--	--	--	--	--	--
Age 20–24 years	MO	0.013	0.013	0.025	0.002	-0.005	-0.013
Age 25–29 years	MO	-0.010	-0.012	0.007	0.001	-0.003	-0.003
Age 30–34 years	MO	0.011	0.010	0.016	-0.007	-0.011	-0.004
Age 35–39 years	MO	-0.034	-0.033	-0.011	0.013	0.005	0.002
Age 40–44 years	MO	-0.019	-0.016	-0.028	0.025	0.018	0.012
Age 45–49 years	MO	0.000	-0.005	-0.014	-0.035	-0.044	-0.035
Household location (NDHS)							
Urban	HH	--	--	--	--	--	--
Rural	HH	0.007	0.009	0.024	0.009	0.001	0.004
Region (NDHS)							
North	HH	--	--	--	--	--	--
Northeast	HH	-0.003	0.011	0.023	**0.100****	**0.118****	**0.105****
Northwest	HH	0.049	0.046	**0.061***	**0.053***	**0.067****	**0.067****
Southeast	HH	**-0.190****	**-0.189****	**-0.164****	**0.033***	**0.040***	0.028
South	HH	**-0.098****	**-0.095***	**-0.088****	-0.001	-0.001	0.000
Southwest	HH	**-0.059***	**-0.077***	**-0.059***	**0.032***	0.027	0.006
Nutrition-specific factors							
Dietary diversity score (NDHS)	CH/HH/LGA	0.009	0.009	**0.010***	0.003	0.001	-0.001

	Level of indicator	Stunting			Wasting		
		Panel 1	Panel 2	Panel 3	Panel 4	Panel 5	Panel 6
			Matched	Nearest Neighbor matching		Matched	Nearest Neighbor matching
		NDHS	NDHS+GHS	NDHS+GHS	NDHS	NDHS+GHS	NDHS+GHS
Household infection, last 2 weeks (NDHS)	HH	**0.035***	**0.037***	**0.024***	0.015	0.009	0.010
Child's age (NDHS)							
Age 6–11 months	CH	--	--	--	--	--	--
Age 12–23 months	CH	0.301	0.417	0.167	-0.251	-0.206	-0.232
Age 24–35 months	CH	0.209	0.202	0.026	-0.318	-0.339	-0.360
Age 36–47 months	CH	0.066	-0.007	-0.083	-0.350	-0.359	-0.358
Age 48–59 months	CH	0.342	0.344	0.072	-0.477	-0.495	-0.336
Breastfeeding status (NDHS)							
Never breastfed	CH	--	--	--	--	--	--
Still breastfeeding	CH	0.049	0.047	-0.082	-0.279	-0.292	-0.306
Weaned	CH	-0.048	-0.046	-0.145	-0.157	-0.194	-0.258
Age*Breastfeeding status (NDHS)							
Age 12–23 months * still breastfeeding	CH	-0.101	-0.220	0.030	0.215	0.167	0.221
Age 12–23 months * weaned	CH	-0.043	-0.162	0.071	0.079	0.057	0.146
Age 24–35 months * still breastfeeding	CH	0.066	0.066	**0.262***	0.245	0.256	0.329
Age 24–35 months * weaned	CH	0.079	0.083	0.245	0.123	0.164	0.258
Age 36–47 months * still breastfeeding	CH	0.185	0.261	**0.418***	0.445	0.445	0.390
Age 36–47 months * weaned	CH	0.176	0.244	**0.289***	0.154	0.182	0.241
Age 48–59 months * still breastfeeding	CH	-0.319	-0.313	-0.031	0.438	0.458	0.315
Age 48–59 months * weaned	CH	-0.110	-0.116	0.132	0.274	0.310	0.217
Nutrition-specific interventions and programs							
Mother's care during last pregnancy (NDHS)							
Received vitamin A	MO	-0.017	-0.019	**-0.028***	-0.002	-0.002	0.007
Took malaria drugs	MO	-0.023	-0.024	**-0.029***	-0.002	-0.005	0.006
Took de-worming medication	MO	-0.008	-0.007	0.005	-0.011	-0.011	-0.006
Took iron supplements	MO	0.005	0.001	0.012	-0.015	-0.010	-0.011

	Level of indicator	Stunting			Wasting		
		Panel 1	Panel 2	Panel 3	Panel 4	Panel 5	Panel 6
			Matched	Nearest Neighbor matching		Matched	Nearest Neighbor matching
		NDHS	NDHS+GHS	NDHS+GHS	NDHS	NDHS+GHS	NDHS+GHS
Protected from tetanus	MO	-0.004	-0.001	-0.006	-0.021	-0.023	**-0.018***
Nutrition-sensitive							
Access to and use of health services and a safe hygienic environment (NDHS)							
BMI of mother	MO	**-0.006****	**-0.006****	**-0.008****	**-0.003***	**-0.003***	**-0.003****
Mother's use of family planning	MO	0.014	0.018	-0.002	0.001	0.000	0.003
HH has improved water source	MO	0.009	0.011	0.009	-0.001	-0.002	0.005
HH sanitation							
Unimproved	HH	--	--	--	--	--	--
Improved, shared	HH	0.021	0.022	0.008	0.004	0.003	-0.012
Improved, not shared	HH	-0.015	-0.016	-0.014	-0.005	-0.005	-0.017
Food availability (GHS-Panel)							
Market in LGA	LGA	--	0.003	-0.005	--	-0.029	-0.009
Food availability at market							
Dairy available at market in LGA	LGA	--	0.045	**0.058***	--	-0.056	**-0.080***
Eggs available at market in LGA	LGA	--	-0.029	-0.016	--	**0.101****	**0.091****
Vitamin A-rich foods available at market in LGA	LGA	--	-0.024	-0.030	--	-0.026	**-0.030***
Economic access to food							
Wealth quintiles (NDHS)							
First quintile	HH	--	--	--	--	--	--
Second quintile	HH	-0.039	-0.035	-0.014	-0.017	-0.019	-0.011
Third quintile	HH	**-0.075****	**-0.069***	**-0.039***	-0.018	-0.019	**-0.038***
Fourth quintile	HH	**-0.098****	**-0.095****	**-0.077****	-0.032	-0.030	**-0.034***
Fifth quintile	HH	**-0.153****	**-0.149****	**-0.111****	0.000	-0.001	-0.009
Household age dependency ratio (NDHS)	HH	**-0.024***	**-0.024***	**-0.014***	-0.004	-0.004	0.000
Family owns livestock (NDHS)	HH	0.014	0.012	0.001	-0.002	-0.003	0.006

	Level of indicator	Stunting			Wasting		
		Panel 1	Panel 2	Panel 3	Panel 4	Panel 5	Panel 6
			Matched	Nearest Neighbor matching		Matched	Nearest Neighbor matching
		NDHS	NDHS+GHS	NDHS+GHS	NDHS	NDHS+GHS	NDHS+GHS
Family owns land (NDHS)	HH	0.021	0.017	0.008	-0.005	-0.008	-0.011
Mother works in agriculture (NDHS)	MO	0.003	0.003	0.004	**-0.035***	**-0.028***	**-0.036****
Bank in LGA (GHS-Panel)	LGA	--	0.010	0.008	--	-0.022	-0.021
Microfinance institution in LGA (GHS-Panel)	LGA	--	0.027	0.033	--	0.000	0.004
Percent of households with 3+ shocks is last year (LGA) (GHS-Panel)	LGA	--	-0.009	0.011	--	-0.002	-0.008
Nutrition-sensitive programs and approaches (proxies)							
Classroom education							
Mother's education level (NDHS)							
No education	MO				--	--	--
Primary	MO	-0.026	-0.024	-0.012	**-0.043****	**-0.039***	**-0.042****
Secondary	MO	**-0.072****	**-0.070****	**-0.048***	**-0.056****	**-0.051****	**-0.053****
Higher than secondary	MO	**-0.135****	**-0.136****	**-0.108****	**-0.085****	**-0.079****	**-0.077****
Nursery in LGA (GHS-Panel)	LGA	--	0.007	-0.007	--	-0.047	-0.051
Primary school in LGA (GHS-Panel)	LGA	--	-0.002	-0.002	--	0.027	**0.042***
Secondary school in LGA (GHS-Panel)	LGA	--	-0.006	0.010	--	0.011	-0.002
Women's empowerment (NDHS)							
Joint household decision-making (LGA)	LGA	--	-0.096	-0.026	--	0.090	-0.009
Disagrees with wife-beating (LGA)	LGA	--	0.025	-0.057	--	-0.075	0.013
Social safety nets (GHS-Panel)							
Safety net used in LGA	LGA	--	0.010	0.007	--	-0.010	-0.019
Health services (GHS-Panel)							
Health clinic in LGA	LGA	--	0.019	0.010	--	0.008	0.010
Private health clinic in LGA	LGA	--	-0.016	-0.019	--	0.047	0.055

WORKING PAPER

MEASURE Evaluation
University of North Carolina at Chapel Hill
123 West Franklin Street, Suite 330
Chapel Hill, NC 27516 USA
Phone: +1 919-445-9350
measure@unc.edu
www.measureevaluation.org

This publication was produced with the support of the United States Agency for International Development (USAID) under the terms of MEASURE Evaluation cooperative agreement AID-OAA-L-14-00004. MEASURE Evaluation is implemented by the Carolina Population Center, University of North Carolina at Chapel Hill in partnership with ICF International; John Snow, Inc.; Management Sciences for Health; Palladium; and Tulane University. Views expressed are not necessarily those of USAID or the United States government. WP-20-241

ISBN: 978-1-64232-262-0