USAID
FROM THE AMERICAN PEOPLE

The Potential Value-Added to USAID of Open Access Data on Food Assistance for Nutrition: *An Analysis of Research Publications Archived on the REFINE Database*

A Report from the Food Aid Quality Review

PREPARED BY:

Lauren Thompson
Audrey Karabayinga
Beatrice Rogers
Patrick Webb

January 2021

This report was made possible by the generous support of the American people through the support of the United States Agency for International Development's Bureau for Humanitarian Assistance (USAID/BHA) and the legacy Office of Food for Peace (FFP) under the terms of Contract No. AID-OAA-C-16-00020, managed by Tufts University.

The contents are the responsibility of Tufts University and its partners in the Food Aid Quality Review (FAQR) and do not necessarily reflect the views of USAID or the United States Government.

The authors have no conflict of interest to declare.

January 2021

Recommended Citation
Thompson, Lauren; Karabayinga, Audrey; Rogers, Beatrice; Webb, Patrick. 2021. *The Potential Value-Added to USAID of Open Access Data on Food Assistance for Nutrition.* Report to USAID. Boston, MA: Tufts University

This document may be reproduced without written permission by including a full citation of the source.

For correspondence, contact:

Patrick Webb
Friedman School of Nutrition Science and Policy
Tufts University
150 Harrison Avenue
Boston, MA 02111
patrick.webb@tufts.edu

ACRONYMS

BCC	Behavior Change Communication
BHA	Bureau for Humanitarian Assistance
BMI	Body Mass Index
CMAM	Community-Based Management of Acute Malnutrition
DDL	Development Data Library
DMP	Data Management Plan
FAIR	Findability, Accessibility, Interoperability, Reusability
FAQR	Food Aid Quality Review
GAM	Global Acute Malnutrition
HAZ	Height-for-age Z-score
IPD	Individual Participant Data
LAZ	Length-for-age Z-score
LNS	Lipid-based Nutrient Supplement
WASH	Water, Sanitation and Hygiene
WAZ	Weight-for-age Z-score
WHZ	Weight-for-height Z-score
WLZ	Weight-for-length Z-score
MDD	Minimum Dietary Diversity
MUAC	Mid-Upper Arm Circumference
MAM	Moderate Acute Malnutrition
OSF	Open Science Framework
REFINE	Research Engagement on Food Interventions for Nutritional Effectiveness
RUTF	Ready-to-use Therapeutic Foods
SAM	Severe Acute Malnutrition
SES	Socio-Economic Status
SQ-LNS	Small Quantity Lipid-based Nutrient Supplementation
USAID	United States Agency for International Development

DEFINITIONS

Codebook: A document that describes the layout of the data within a dataset and offers an explanation as to what the data codes mean. A codebook defines the values associated with a given variable or response option.[1]

Open-Access Repository: An online database which makes the full text of items (or complete files) it contains freely and immediately available without any access restrictions.[2]

Metadata: Term used for data that describes other data. In this case, metadata is the information that describes the data contained in a specific dataset or file.

Syntax: The set of symbols and rules forming statements and expressions, in the context of computer programming.

Replication code: References the line-by-line code used to run the statistical analyses presented in a given publication of a scientific study, in the specific statistical software language used by the authors. This may permit replication of analyses by other analysists.

The FAIR guiding principles: A set of guidelines intended to support the findability, assessibility, interoperability and reusability of scholarly research findings.[3]

TABLE OF CONTENTS

EXECUTIVE SUMMARY

Open access data have the potential to enhance transparency of study results and facilitate reuse of data to inform further scientific research. Many research funding organizations support open access data sharing and have adopted frameworks such as the FAIR principles (Findability, Accessibility, Interoperability, Reusability) to guide good data stewardship practices among their partners.[3] Given the high cost associated with conducting research among vulnerable populations such as those experiencing the burden of undernutrition, leveraging, pooling and re-analyzing available data may be an efficient method to delve further into key questions.

The Food Aid Quality Review (FAQR) project at Tufts University Friedman School of Nutrition Science and Policy, in consultation with United States Agency for International Development (USAID), undertook a study aimed at identifying what databases associated with publications housed in the Research Engagement on Food Interventions for Nutritional Effectiveness database are accessible to any researcher. This desk study ran from June through December 2020. It was found that of 179 studies published and archived in REFINE over the last 5 years, only 15 (8.4% of this sample) had posted underlying datasets in a publicly accessible format. Where datasets were published, many were minimal, and they lacked accompanying documentation (codebooks, study protocols, data collection instruments, and replication codes) needed for interpretation.

An in-depth assessment of the publications from this sample that were archived with open access datasets showed significant heterogeneity among study designs, which hindered prospects for conducting meta-analyses with pooled data derived from these studies. Further, these studies collected a wide range of experimental and program-level exposure variables, and there was much heterogeneity among the coding structures of key outcome variables. Therefore, this review concluded that access to 'open access' data is quite limited at present, and thus the potential for pooled secondary analyses is also limited. This indicates that robust guidance and infrastructure for data sharing is needed in order to make open access truly meaningful.

Key recommendations identified for funding organizations, including USAID, focus on approaches to improve the FAIRness of open data. Funding organizations should provide guidance on the following:

1. **Prospective planning for data sharing**
   - Require a detailed data management plan (DMP) specifying data sharing procedures.
   - Include in DMPs: data collection methods, defined roles for research staff assigned to data archival, and plans for long-term accessibility of data.
   - Provide guidance and example DMPs, a mechanism for evaluation of DMPs, and training for evaluators.
   - Include in participant consent forms a statement of intent to openly share de-identified data.

2. **Curation of data for dissemination, repository deposition, and inclusion of accompanying documentation**
   - Specify timeframe to archive underlying data following article publication.
   - Ensure datasets are comprehensive and de-identified.
   - Archive documentation needed for interpretation of dataset contents:

    a. Complete codebooks/data dictionaries with variable names, descriptions, response options/values and value labels, and criteria for indicator calculation and data exclusion.
    b. Data collection instruments with items corresponding to dataset variables and response options.
    c. Syntax used to run analyses to allow for reproduction of results (replication code).
    d. Study protocols detailing when and how datasets were collected (including time points).
- Clarify preferred location for archiving data:
    a. Require partners to archive complete datasets and accompanying documentation in a publicly accessible database.
    b. Provide a list of preferred open access repositories for archiving of data.
- Monitor data sharing compliance:
    a. Confirm datasets have been appropriately archived by the deadline and follow up with non-compliant partners.

REFINE OPEN ACCESS DATA STUDY RECOMMENDATIONS

The Potential Value-Added to USAID of Open Access Data on Food Assistance for Nutrition: An Analysis of Publications Archived on the REFINE Database review, conducted by Tufts University's Food Aid Quality Review (FAQR) team, found that most underlying datasets of recent studies related to food assistance for nutrition are not open access. Funding organizations are progressively requiring that their partners make data open access and recommend that partners adhere to the **FAIR** principles that require open data to be **findable, accessible, interoperable** and **reusable**.[1] Making research data publicly available can increase transparency, accessibility and maximization of existing data and facilitate innovation with data reuse. Datasets that are open access are usually **findable** and **accessible** but not always **interoperable** and **reusable**. These key recommendations identify what is needed to improve the **FAIR**ness of datasets to make a significant impact on the advancement of food assistance for nutrition.

Findable **Accessible** **Interoperable** **Reusable**

RECOMMENDATIONS

1. Prospective Planning for Data Sharing

- Require a detailed data management plan (DMP) specifying data sharing procedures.
- Include in DMPs: data collection methods, defined roles for research staff assigned to data archival, and plans for long-term accessibility of data.
- Provide guidance and example DMPs, a mechanism for evaluation of DMPs, and training for evaluators.
- Include in participant consent forms a statement of intent to openly share de-identified data.

2. Curation of Data for Dissemination, Repository Deposition, and Inclusion of Accompanying Documentation

- Specify timeframe to archive underlying data following article publication.
- Ensure datasets are comprehensive and de-identified.
- Archive documentation needed for interpretation of dataset contents:
  - Complete codebooks/data dictionaries with variable names, descriptions, response options/values and value labels, and criteria for indicator calculation and data exclusion.
  - Data collection instrument with items corresponding to dataset variables and response options.
  - Syntax used to run analyses to allow for reproduction of results (replication code).
  - Study protocols detailing when and how datasets were collected (including time points).
- Clarify preferred location for archiving data:
  - Require partners to archive complete datasets and accompanying documentation in a publicly accessible database.
  - Provide a list of preferred open access repositories for archiving of data.
- Monitor data sharing compliance:
  - Confirm data has been appropriately archived by the deadline and follow up with non-compliant partners.

[1] WILKINSON, M., DUMONTIER, M., AALBERSBERG, I. *ET AL.* THE FAIR GUIDING PRINCIPLES FOR SCIENTIFIC DATA MANAGEMENT AND STEWARDSHIP. *SCI DATA* 3, 160018 (2016). HTTPS://DOI.ORG/10.1038/SDATA.2016.18

1. INTRODUCTION

In recent years, the value of increasing availability and open accessibility of data to the public has drawn attention, particularly in the field of global public health and nutrition, but also for programmatic use. In this information age, leveraging large amounts of data (Big Data analysis) is increasingly important for informing global public health policy and decision-making.[4] It is becoming widely understood that sharing de-identified, individual-level health research data has the potential to enable full interpretation of study results and critical appraisal via replication of analyses, thereby enhancing transparency in study reporting.[5] Further, effective reuse (mining) of research data may facilitate innovation and discovery through the generation of novel research questions and contribute to powerful secondary and meta-analyses.[5-8]

Many researchers and research funders, including the Bill & Melinda Gates Foundation and the United States Agency for International Development (USAID), have developed statements supporting increasing the availability of health research data to facilitate "equitable, ethical, and efficient" use of such data in the interest of the public.[7,9] A number of these funders and organizations have encouraged frameworks such as the FAIR Data Principles to guide good data stewardship practices. At a high level, these principles encourage making data available with a unique global identifier (similar to a DOI for publications) to make it easy to locate, registering or indexing in a searchable resource, accompanying datasets with rich, clear, and accessible documentation (e.g., codebooks), all of which allow interoperability and reuse of the data.[8]

The Food Aid Quality Review (FAQR) project at Tufts University, funded by USAID's Bureau for Humanitarian Assistance (BHA), provides actionable recommendations to USAID and its partners on approaches to improve nutrition among vulnerable populations, for whom the direct distribution of food assistance products for nutrition can make a significant impact. During Option Year 2, FAQR undertook a study of open data which endeavored to leverage open access primary research (as opposed to project evaluation) datasets from published scientific studies housed on the Research Engagement on Food Interventions for Nutritional Effectiveness (REFINE) website at https://refinenutrition.org/. REFINE is a knowledge-sharing mechanism that seeks to improve the effectiveness of food-supported interventions in emergency and non-emergency contexts. This website aggregates and organizes information from ongoing clinical trials and published, peer-reviewed scientific journal articles that consider advancements in research on food assistance for nutrition in one comprehensive database.

Donors and practitioners can expect considerable benefits from maximizing the use of data collected from the field in rigorous ways and made available to other researchers beyond the original study.[10] These benefits would include more robust effect estimates for interventions using food aid and food assistance products, with a view to informing improved (effective, cost-effective) investments in program designs and implementation.

The aim of this review was to assess the landscape of open access data in the sphere of research on food assistance for nutrition, and to identify potential research questions that could be addressed with pooled meta-analyses. Specific goals of the review were:

1. Locate and obtain open access primary datasets from published research studies catalogued in the REFINE database in the last 5 years;
2. Compare study designs and open access datasets to identify research questions with the potential to be addressed by pooled/meta-analyses;
3. Compile and harmonize datasets in a standard format conducive to such analyses; and
4. Provide insight and recommendations for future data-sharing practices and policies to be considered by USAID and other research funders.

2. METHODS

2.1. SECTION I: STUDY COMPARISON

2.1.1. DATA COLLECTION PROCESS

The manuscripts used in this project were limited to those included in the REFINE database (available at https://refinenutrition.org/). REFINE is a knowledge-sharing platform that houses information from ongoing clinical trials and published, peer-reviewed scientific journal articles. The selection process for publication inclusion on REFINE is based on the following criteria:

1. The study is designed to address one or more specific nutrition conditions in young children and women in a humanitarian or development context.
2. The study is designed to employ or investigate a specialized, nutritionally-enhanced food assistance product, or an ingredient intended for use in a nutritionally enhanced food assistance product to prevent or treat the potential causes or outcomes of undernutrition. The study should apply a rigorous intervention design addressing the efficacy or effectiveness of the nutritionally-enhanced food assistance product on the predetermined study outcome (further elaboration below).
3. The study reports outcomes that measure the impact of the study food on one or more of the following: birth weight, head circumference, weight gain, height gain, weight-for-age (WAZ), height/length-for-age (HAZ/LAZ), weight-for-height/length (WHZ/WLZ), mid-upper arm circumference (MUAC), lean body mass, cognitive development and/or cognition, "recovery" from malnutrition (as defined by the study investigators), mortality, default from the interventional nutrition program, nutritional intake, and/or product acceptability.

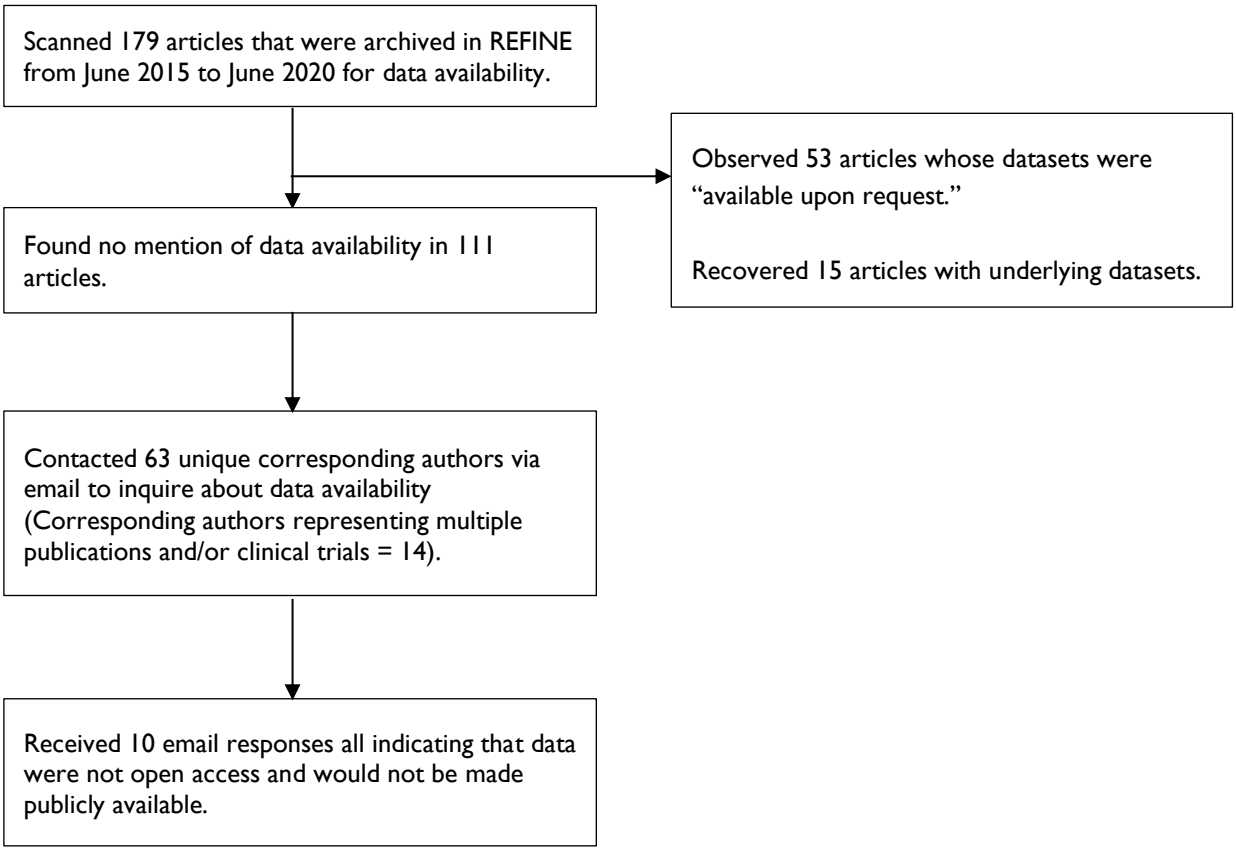
To identify publications suitable for inclusion in REFINE, a periodic manual keyword search (shown in **Table 1** below) is employed in relevant clinical trial registry webpages, and MEDLINE through PubMed (https://pubmed.ncbi.nlm.nih.gov). Included in the analysis were scientific articles published and archived by REFINE between June 2015 and June 2020. The articles were limited to publications detailing the results of a clinical trial or interventional study employing a specialized, nutritionally-enhanced food assistance product for the treatment or prevention of malnutrition.

Table 1: MEDLINE Keyword Search for Articles for Inclusion in the REFINE Database

REFINE KEYWORD SEARCH
(“malnutrition OR malnourished OR undernutrition OR undernourished OR stunting OR stunted OR wasted OR wasting OR underweight) AND ("therapeutic food" OR "supplementary food" OR "supplemental food" OR "food supplement" OR "nutrient supplement" OR "lipid-based nutrient supplement" OR "lipid based nutrient supplement" OR "ready to use" OR "ready-to-use" OR "ready to use food" OR "ready-to-use-food" OR RUF OR RUTF OR LNS OR "fortified-blended food" OR "fortified blended food" OR "corn-soy blend" OR "corn soy blend" OR CSB) NOT (disease OR disorder OR syndrome OR diabetes OR cancer OR fibrosis OR surgery OR ADHD OR stroke OR schizophrenia OR dysplasia OR hypertension OR autism OR sleep OR encephalopathy OR hemodialysis OR patients OR "dietetic" OR "older adults" OR "nursing home")

In total, 179 articles were identified which met the initial criteria; these were then reviewed for mention of data availability, typically in the form of a “data availability statement”. The process of identifying studies with publicly available datasets is outlined in **Figure 1**. Of these, 15 articles had publicly available underlying datasets; 53 noted that their data were only available upon request; and the remaining 111 had no data availability statements or mention of data availability. For the articles that did not mention data availability, the team conducted a search by title and DOI in relevant open access databases, including Harvard Dataverse (https://dataverse.harvard.edu/), Open Science Framework (OSF) (https://osf.io/q5j8g/), Dryad (https://datadryad.org/stash), and Figshare (https://figshare.com/), World Bank (https://data.worldbank.org/), USAID Development Data Library (DDL) (https://data.usaid.gov/), in addition to a general Google search.

Figure 1: Process of Identifying Studies with Publicly Available Datasets.



For the 111 articles whose datasets were not found using this method, corresponding authors were contacted to inquire about data availability in an open access database. Since multiple articles were published using data from a single clinical trial, if a single corresponding author was linked to multiple studies, they were contacted only once. The FAQR team contacted 63 corresponding authors; a total of 10 responses were received. All responses indicated that the underlying datasets for the respective publication(s) were *not* publicly available and would not be made available "in the near future". As a result, a total of just 15 primary datasets were recovered and used.

The public accessibility of the primary dataset was highly dependent on the scientific journal in which the corresponding article was published. Eleven of the recovered datasets were linked to articles published in a PLOS Journal,[11-21] which has a robust data-sharing policy requiring authors to make "all data necessary to replicate the study's findings publicly available, without restriction, at the time of publication."[22] Two datasets had corresponding articles published in the Lancet Journal.[23,24] The remaining two datasets corresponded to articles published in a British Medical Journal (BMJ) (i.e., BMJ Open and BMJ Paediatrics).[25,26] The specific data sharing policies for each journal mentioned are summarized in **Table 2**.

Table 2: Data Sharing Policies of Selected Scientific Journals

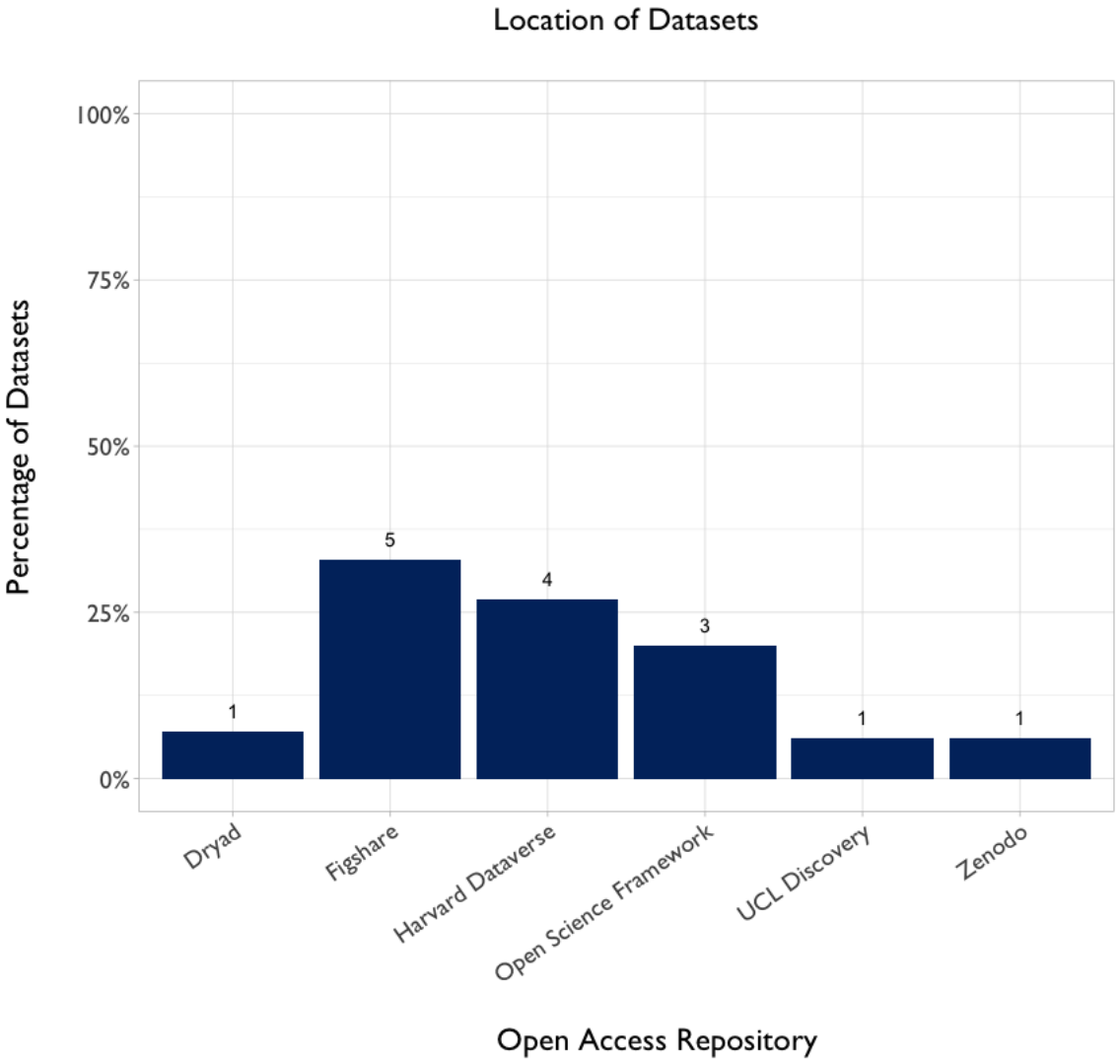
JOURNAL	DATA SHARING POLICY	% OF PUBLIC DATASETS FROM ARTICLES PUBLISHED IN JOURNAL
PLOS[a]	• "PLOS journals require authors to make all data necessary to replicate their study's findings publicly available without restriction at the time of publication." • Authors must share minimal dataset. • Deposition in a data repository strongly recommended.	11 of 15 (67%)
Lancet[b]	• From September 21st, 2020, all submitted manuscripts must contain a "data sharing statement." • This statement should include information on whether data will be made available, what data will be made available, whether additional documents will be available, in addition to when and where data will be archived. • The statement must also include information regarding by what "access criteria" data will be shared. • No minimum dataset required.	2 of 15 (13%)
British Medical Journal (BMJ)[c]	• It is "strongly encouraged" that supporting data be "made available as soon as possible, wherever legally and ethically possible." • It is encouraged that as much data be shared as possible but "at least the minimum data to reproduce the results" be presented in the article.	2 of 15 (13%)

a. "Data Availability", PLOS One, located at https://journals.plos.org/plosone/s/data-availability#loc-minimal-data-set-definition
b. "Information for Authors", Lancet, located at https://www.thelancet.com/pb-assets/Lancet/authors/tl-info-for-authors.pdf
c. "Data Sharing", BMJ, located at https://authors.bmj.com/policies/data-sharing/

2.1.2. OPEN ACCESS DATASET SOURCE AND DOCUMENTATION

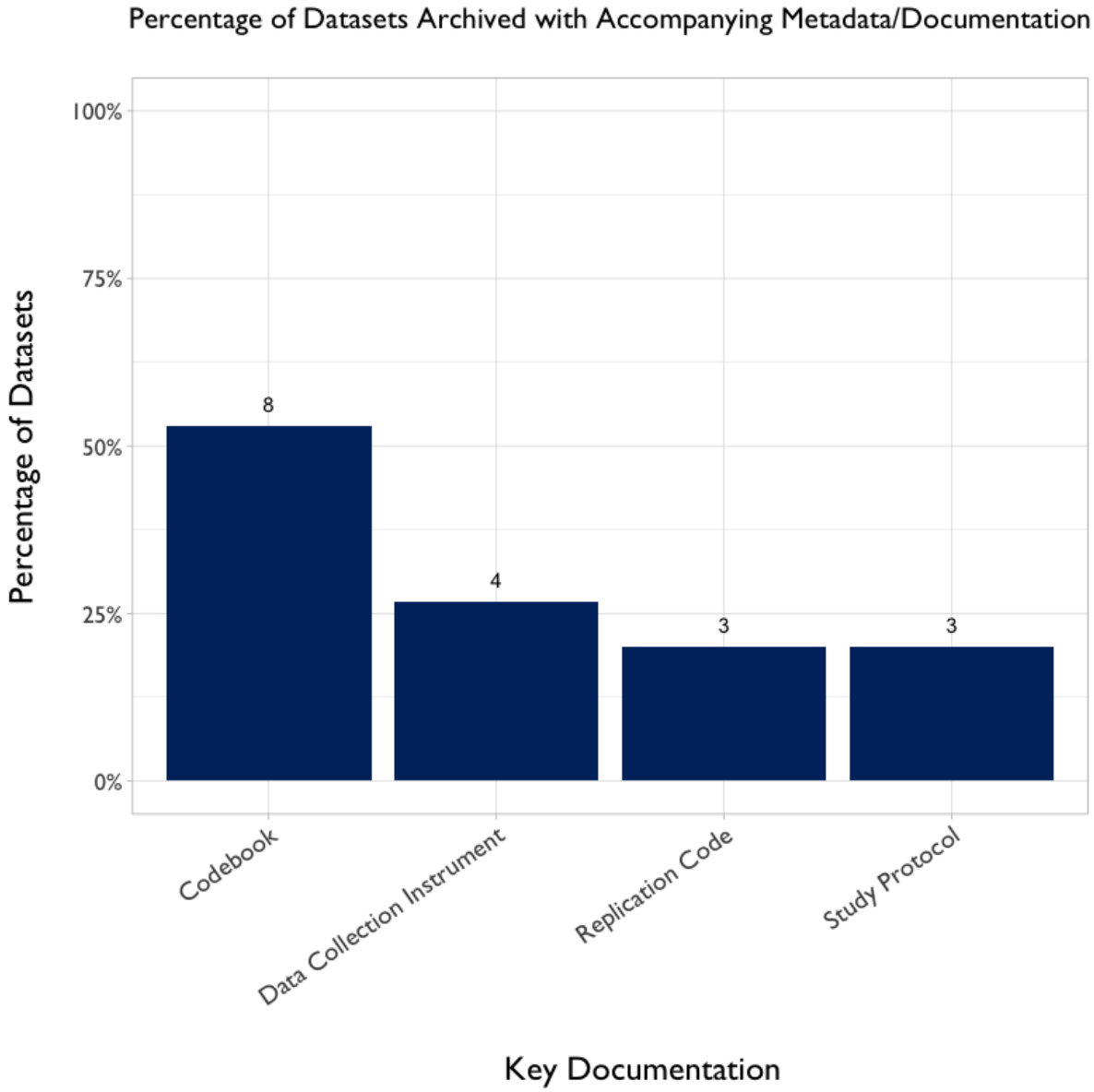
Most of the recovered datasets were deposited in a general or multidisciplinary open access repository that archived data from multiple scientific fields in various formats. **Figure 2** displays the distribution of recovered datasets in the open access repositories in which they were archived.

Figure 2: Location of Primary Datasets in Various Open Access Repositories.



As shown in **Figure 3**, only eight of the 15 datasets recovered had been archived with accompanying codebooks,[11,14-16,23-26] which are critical pieces of documentation that contain the structure and layout of a data file, including the variable names/descriptions, response options/values, and value labels. Four studies were archived with corresponding data collection instrument(s) used,[14,15,23,24] three with replication codes for verification of statistical analyses,[23-25] and three with study protocols,[13,23,24] which contain methodological information regarding the data collection process. **Annex Table 1** details the location of recovered datasets and accompanying documentation recovered with the datasets.

Figure 2: Availability of Accompanying Documentation (codebooks, data collection instrument, replication code and study protocol) for Recovered Datasets.



2.1.3. STUDY REVIEW AND COMPARISON

Before considering recovered datasets, all 15 corresponding publications were reviewed to compare key details of the studies for similarities that would elicit pertinent research questions to be addressed by meta-analyses. To be combined for pooled/meta-analyses, clinical trials/studies should address the same or similar research questions, in similar populations and study settings, employing comparable intervention components and address similar or the same outcome variables. [27]

Following REFINE inclusion criteria, all studies were interventional studies or publications from clinical trials seeking to assess the effectiveness or efficacy of a specialized, nutritionally-enhanced food assistance product on malnutrition outcomes. Eleven of the 15 studies (73%) employed a randomized-control trial design.[11,12,14-17,20,21,23-25] Of these, six employed a cluster-randomized

design[11,14,15,21,23,24] and five used a single-center or individual-level randomization design.[12,16,17,20,25] All studies employed several sampling designs for data collection. Most collected longitudinal data (n=12) with different sampling frequencies.[11-16,19-21,23-25] Others employed a cross-sectional survey design (n=4).[14,15,18,26]

Most studies targeted young children. Eleven studies (73%) specified an enrollment population that included children aged 0-59 months.[11,13-18,23-26] Three studies enrolled children aged three to 16 years,[16,19,21] and four studies enrolled pregnant women.[12,20,23,24] Among the studies that enrolled children aged 0-59 months, two studies specified enrollment criteria of a SAM diagnosis (MUAC <11.5 cm and/or WLZ <-3);[16,17] one specified enrollment of children with a MAM diagnosis (MUAC≥11.5 cm and <12.5 cm and/or WHZ ≥-3 and <-2);[13] and one specified inclusion of only children who were stunted (LAZ<-2.5).[25]

Nine different specialized, nutritionally-enhanced food assistance products were employed among all studies. Details regarding the specific specialized, nutritionally-enhanced food assistance product and their nutrient content are reported in **Table 3**. The most frequently investigated specialized, nutritionally-enhanced food assistance product was a lipid-based nutrient supplement (LNS). LNS was supplied in nine studies to either pregnant women (n=1)[12] or children six months and older (n=8).[11,13-15,18,21,23,24] Nine studies employed a number of complementary interventions. These include IYCF counseling (n=3),[23,24,26] BCC strategies for improved nutrition and sanitation behavior (n=5),[14,15,18,23,24] WASH-related interventions such as construction of latrines, handwashing stations, provision of water purification tablets (n=2),[23,24] and treatment for diarrhea, fever and/or malaria (n=2).[11,12] Additionally, two studies delivered the specialized, nutritionally-enhanced food assistance product through a school-feeding platform (n=2).[19,21]

This review of publications highlighted few interventions or other design elements that were analogous across studies that might allow for meta-analyses to derive pooled effect estimates. As a number of studies were comparable with respect to the use of an LNS (n=8),[11-15,18,23,24] and age population of 0-59 months (n=11),[11,13-18,23-26] the corresponding open access datasets were scanned to further assess possibilities for pooling data for analysis. This process is described in Section 2.2.

Table 3: Overview of Design Aspects of Studies with Open Access Datasets

(AUTHOR, YEAR) COUNTRY	TARGET POPULATION	STUDY DESIGN	INTERVENTION	PRIMARY OUTCOMES	DATA SET
Zhang et al., 2016[26] China	Children aged 6-23 months and their primary caregiver	Controlled intervention study	2 years of daily complementary food supplement (Ying Ying Bao)	Prevalence of anemia (hemoglobin <110 g/L) and prevalence of stunting	Variables reported include those in "YYB Surveys" (survey instrument not provided). Nearly all variables coded as categorical, save height and hemoglobin data.
Hess et al., 2015[11] Burkina Faso	Children aged 9 months	Partially masked, placebo-controlled, cluster-randomized intervention study	9 months of SQ-LNS + zinc at various mg amounts daily	Growth (length, weight, z scores, MUAC, head circumference), incidence of diarrhea, incidence of malaria, pZn	Minimum dataset reported for the analyses presented in this manuscript only; not representative of the whole of the iLiNS trial.
Fabiansen et al., 2017[13] Burkina Faso	Children aged 6-23 months, diagnosed with MAM	Randomized 2x3x3 factorial trial	3 months of one of 12 formulations of a CSB+ or LNS-based supplementary food for the treatment of MAM	Free fat mass index increment (FFMI, which equates FFM divided by length)	Data at enrollment and end-line only. Minimal dataset of variables reported for the analyses present in this manuscript only, not representative of the TreatFood trial.
Nkhoma et al., 2019[12] Malawi	Pregnant women	Randomized controlled intervention trial	SQ-LNS for the duration of the pregnancy	Gestational weight gain per week, pregnancy duration, birth weight, LAZ, WAZ, WLZ, HCZ, inadequate GWG, pre-term birth, low birth weight, stunting, wasting, underweight, small head circumference, and small for gestational age	Minimal dataset of variables reported in this analysis only. No specified timepoint of data collection reported. No data dictionary or codebook.
Becquey et al., 2019[14] Burkina Faso	Children 0-59 months of age and their primary caregiver	Two-arm, cluster-randomized, nonblinded, effectiveness trial (repeated cross-sectional and longitudinal evaluation designs)	Preventative AM package integrated into CMAM programming (BCC, SQ-LNS) for 2 years	Cross-sectional: 1. AM screening coverage, 2. AM treatment coverage, and 3. AM prevalence. Longitudinal: 1. AM screening coverage over 18-month follow-up, 2. AM treatment coverage over follow-up, and 3. incidence of first AM episode over follow-up	Comprehensive dataset for the full trial available. Includes information for both cross-sectional (baseline and end-line) and longitudinal studies (18 months of follow-up).

(AUTHOR, YEAR) COUNTRY	TARGET POPULATION	STUDY DESIGN	INTERVENTION	PRIMARY OUTCOMES	DATA SET
Huybregts et al., 2019[15] Mali	Children aged 6-23 months and their primary caregiver	Two-arm, cluster-randomized, nonblinded, effectiveness trial (repeated cross-sectional and longitudinal evaluation designs)	Preventative AM package integrated into CMAM programming (BCC, SQ-LNS) for 2 years.	Cross-sectional: 1. AM screening coverage, 2. AM treatment coverage, and 3. AM prevalence. Longitudinal: 1. AM screening coverage over 18-month follow-up, 2. AM treatment coverage over follow-up, and 3. incidence of first AM episode over follow-up.	Comprehensive data set for the full trial available. Includes information for both cross-sectional (baseline and end-line) and longitudinal studies (18 months of follow-up)
Martinez et al., 2018[25] Guatemala	Stunted children aged 6-23 months and their caregivers	Single-center, individually randomized, parallel-group superiority trial	Individualized complementary feeding education for caregivers for 6 months.	Change in LAZ.	Variables reported for the analyses presented in this manuscript only. Baseline and end-line survey data.
Iannotti et al., 2016[21] Haiti	Children aged 3-16 years	A cluster-matched, randomized longitudinal study	Daily Vita Mamba supplement (fortified peanut butter paste) for 26 weeks (6.5 months).	Change in hemoglobin concentration, anemia status.	Variables reported for the analyses presented in this manuscript only. Baseline data, and end-line anthropometric and anemia data. No data dictionary.
Luby et al., 2018[24] Bangladesh	Pregnant women and the child in-utero at enrollment, and children <3 years residing in the compound observed for diarrhea outcomes	Cluster-randomized trial	Randomized to either a water, sanitation, handwashing, or nutrition intervention, or a combination of these interventions, for 2 years.	[23]Caregiver-reported diarrhea in the past 7 days among children (in utero or <3 years at enrollment) and LAZ at year 2 for index children.	Comprehensive data set for clinical trial available. Includes variables collected at enrollment, year 1 and year 2 surveys.
Null et al., 2018[23] Kenya	Pregnant women and the child in-utero at enrollment, and children <3 years residing in the compound observed for diarrhea outcomes	Cluster-randomized trial	Randomized to either a water, sanitation, handwashing, or nutrition intervention, or a combination of these interventions, for 2 years.	Caregiver-reported diarrhea in the past 7 days among children (in utero or <3 years at enrollment) and LAZ at year 2 in index children born to the enrolled pregnant women.	Comprehensive data set for clinical trial available. Includes variables collected at enrollment, year 1 and year 2 surveys.

(AUTHOR, YEAR) COUNTRY	TARGET POPULATION	STUDY DESIGN	INTERVENTION	PRIMARY OUTCOMES	DATA SET
Bandsma et al., 2019[16] Kenya & Malawi	Children 6 months - 13 years of age with complicated SAM	Randomized, double-blind controlled trial	A reduced carbohydrate, lactose-free F-75 formula for the treatment of complicated SAM until stabilization.	Number of days from hospital admission to first stabilization.	Data set includes all variables used in the analysis. Baseline, follow-up, and daily monitoring data included.
Kangas et al., 2019[17] Burkina Faso	Children aged 6 to 59 months	Randomized controlled clinical trial	Reduced RUTF dose for uncomplicated SAM upon admission to CMAM program until discharge.	Weight gain velocity (g/kg/d) from admission to discharge.	Minimal data set of anthropometric variables at baseline and end-line. No data dictionary/codebook.
Svefors et al., 2018[20] Bangladesh	Pregnant women	Randomized controlled trial with factorial design (6 groups)	Early or usual initiation timing of food and micronutrient supplementation for the duration of pregnancy.	DALYs averted (sub-analysis of clinical trial).	Minimal data set, longitudinal HAZ scores, and some anthropometry at 4.5- and 10-years post-intervention. No data dictionary/codebook.
Style et al., 2017[18] Kenya & Djibouti	Refugee children aged 6-23 months	Cross-sectional surveys (two-stage cluster sampling). Surveys covered children 6-23 months and 6-59 months of age	Nutributter LNS supplements to children. Duration ranged from 1 to 9 months depending on the refugee camp.	Hemoglobin concentration, prevalence of anemia, and stunting.	Minimal data set of variables used in the analysis. Data per survey year, per camp at baseline and end-line. No data dictionary/codebook.
Adams et al., 2017[19] Bangladesh	Children aged 6-11years	Mixed methods study (quantitative and qualitative cohort pre-post research survey design)	Daily fortified biscuit to school children for14 weeks.	Micronutrient status and qualitative acceptability.	Minimal data set of variables used in the analysis. Baseline and end-line survey data. No data dictionary/codebook.

2.2. SECTION II: DATASET ANALYSIS

2.2.1. DATASET COMPARISON

An overview of the variables reported in each open access dataset affiliated with each publication are presented in **Figure 4**. Nominal and binary variables are presented as they are reported in the dataset. The remaining variables are categorized for ease of visualization in the figures and do not necessarily reflect how the variable appears in the dataset. For example, "Dietary Data" may include information regarding frequency of individual food intake, frequency of food group intake, minimum acceptable diet, or other metrics. Information regarding how each of these variable categories was represented and coded in the dataset is summarized and included in Annex 6.1.2.

To establish the feasibility of pooling data for analysis, the datasets were compared by the study food used in the intervention, target population, and outcomes most frequently collected and reported, in addition to whether the outcome was primary, secondary, or tertiary in the study design. The precise outcomes from all studies are summarized in **Table 4**. The datasets, codebooks, and other available documentation were used to inform this analysis. For some datasets[13,17,21], which were published without a corresponding codebook but had a Stata file extension *'.dta'*, a codebook could be generated using Stata[28] to decipher the data collected for these studies. This aided in interpretation of some data coding schemes; however, instances of unlabeled variables and missing value labels persisted and created challenges for interpretation. This is expanded upon in Section 2.2.2.

Only four recovered datasets were comprehensive.[14,15,23,24] The rest were minimal in content and shared only material that was defined and required by the scientific journal in which study results were published (**Table 2**). These datasets only provided the variables and measurement timepoints (i.e. baseline and end-line) included in the statistical analyses detailed in the article and not the full data collected over the duration of the trial, as indicated by the trial design and protocol. Additionally, some articles reported secondary or sub analyses of primary data, providing access to even less complete data.[12,20] This provision of only a subset of the complete study or clinical trial data hindered the use of most recovered datasets. The information presented here is the most accurate interpretation of the datasets as they appeared in the open access format. A visual comparison and representation of the datasets is displayed in **Figures 4-9**.

As shown in **Figures 4-9**, few variables were reported in the open access datasets, and there was minimal overlap across studies. The most frequently reported variables were measures of anthropometry. This would be expected, as 73% of studies reported the effect of an intervention on an indicator of child growth as a primary or secondary outcome. Ninety three percent of studies (n=14) reported age[11-19,21,23-26] and sex;[11,13-21,23-26] 86% (n=13) reported height or length;[11,13-21,23,24,26,29] 73% (n=11) reported LAZ;[11,12,16-21,23,24,26] and 67% (n=10) reported weight,[11,13-20,24] WAZ,[11-18,23,24] and/or WHZ.[11-13,16-18,21,23-25] Forty percent (n=6) reported MUAC for the child;[11,14,15,17,18] 40% (n=6) reported hemoglobin concentration;[11,14-18] and 33% (n=5) reported presence of child anemia.[11,14,15,18,21] Fewer than 40% of studies reported maternal variables: maternal age was reported in 40% of studies (n=6),[12,14,23-26] 33% (n=5) reported maternal height and education,[12,15,21,23,24] and 13% (n=2) reported maternal weight[15,24] or maternal BMI.[12,24] Sixty percent (n=9) reported dietary data broadly (including data on breastfeeding),[11,13-16,19,21,25,26] 46% (n=7) reported asset ownership

information,[11,14,15,17,21,23,24] 33% (n=5) reported WASH data broadly,[14,15,21,24,26] and 33% (n=5) reported information regarding household construction.[14,15,21,23,24] These generalized variable categories (dietary data, asset data, WASH data and household construction) had various coding schemes, which are expanded upon in Section 2.2.2. Detailed information on coding schemes for these variable categories is included in Annex 6.1.2.

In addition to the heterogeneity in variables reported, data collection methods and survey designs (i.e. cross-sectional vs. longitudinal) differed among studies. As well, study duration varied, resulting in collection and reporting of child anthropometric data at differing timepoints across these datasets. Two studies reported all monthly child anthropometric measurements collected over an 18-month period.[14,15] One study reported only HAZ monthly for 12 months, later at 15, 18, 21, 24 and 54 months, and lastly 4.5 and 10 years of age.[20] Four studies employed a cross-sectional survey design and reported anthropometric measurements among different children at baseline and end-line only.[14,15,18,26] Five studies employed a longitudinal survey design where they followed the same children over time, but only baseline and end-line anthropometric measurements were reported.[11] Two studies also employed a longitudinal survey design, but the open access datasets only included measurements of child anthropometry at midline and end-line.[23,24] The remaining studies reported a variable number of anthropometric measurements per child: at admission, 2 weeks and exit during community-based treatment of uncomplicated SAM,[17] and at least 3 daily measurements during an inpatient SAM treatment program.[16] These differences created challenges for discovering ways to harmonize datasets with respect to child anthropometry.

Figure 3: Exposure Variables Reported in Recovered Datasets.

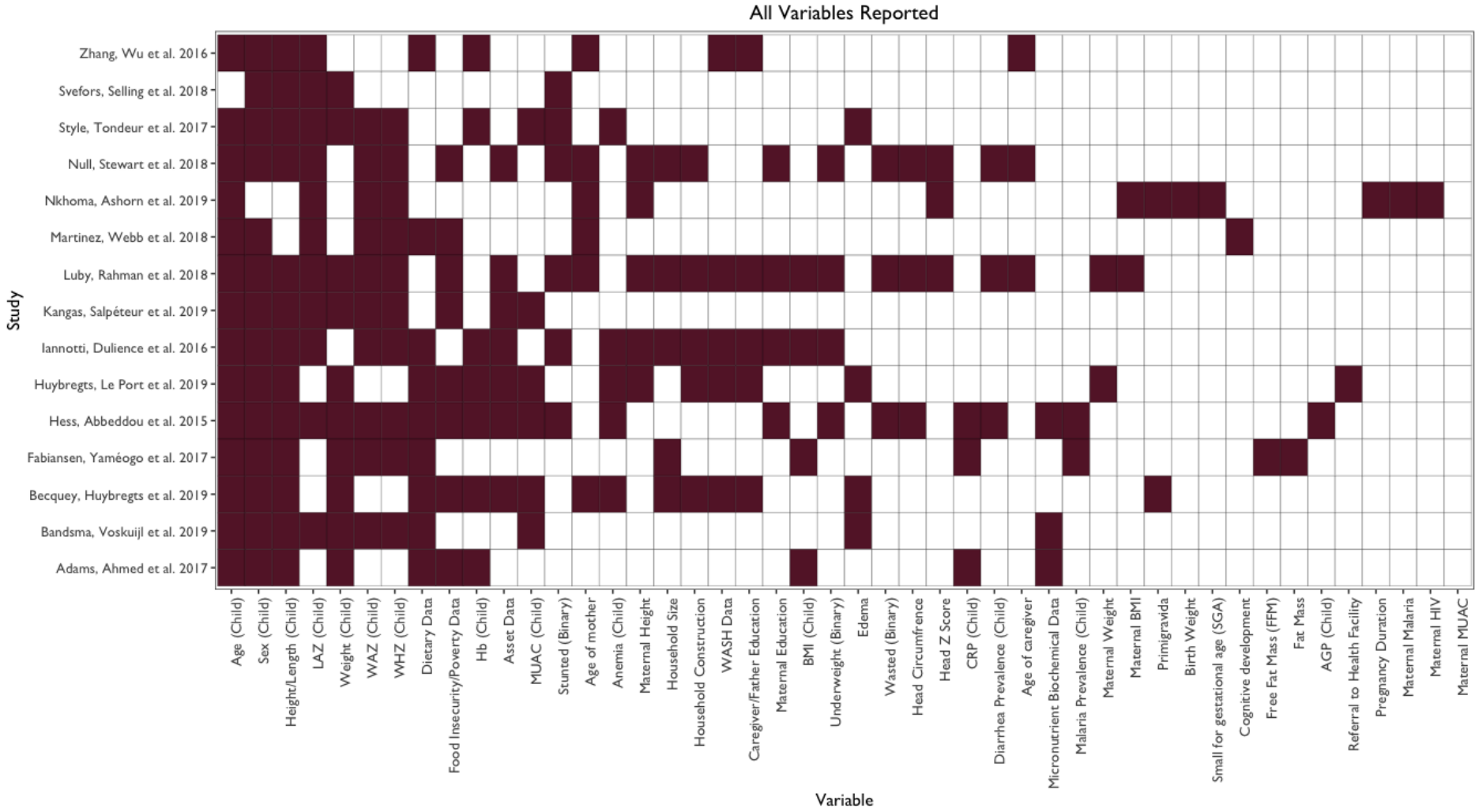


Figure 4: Exposure Variables Reported in Recovered Datasets, Grouped by Type of Specialized, Nutritionally-enhanced Food Assistance Product used in Corresponding Study.

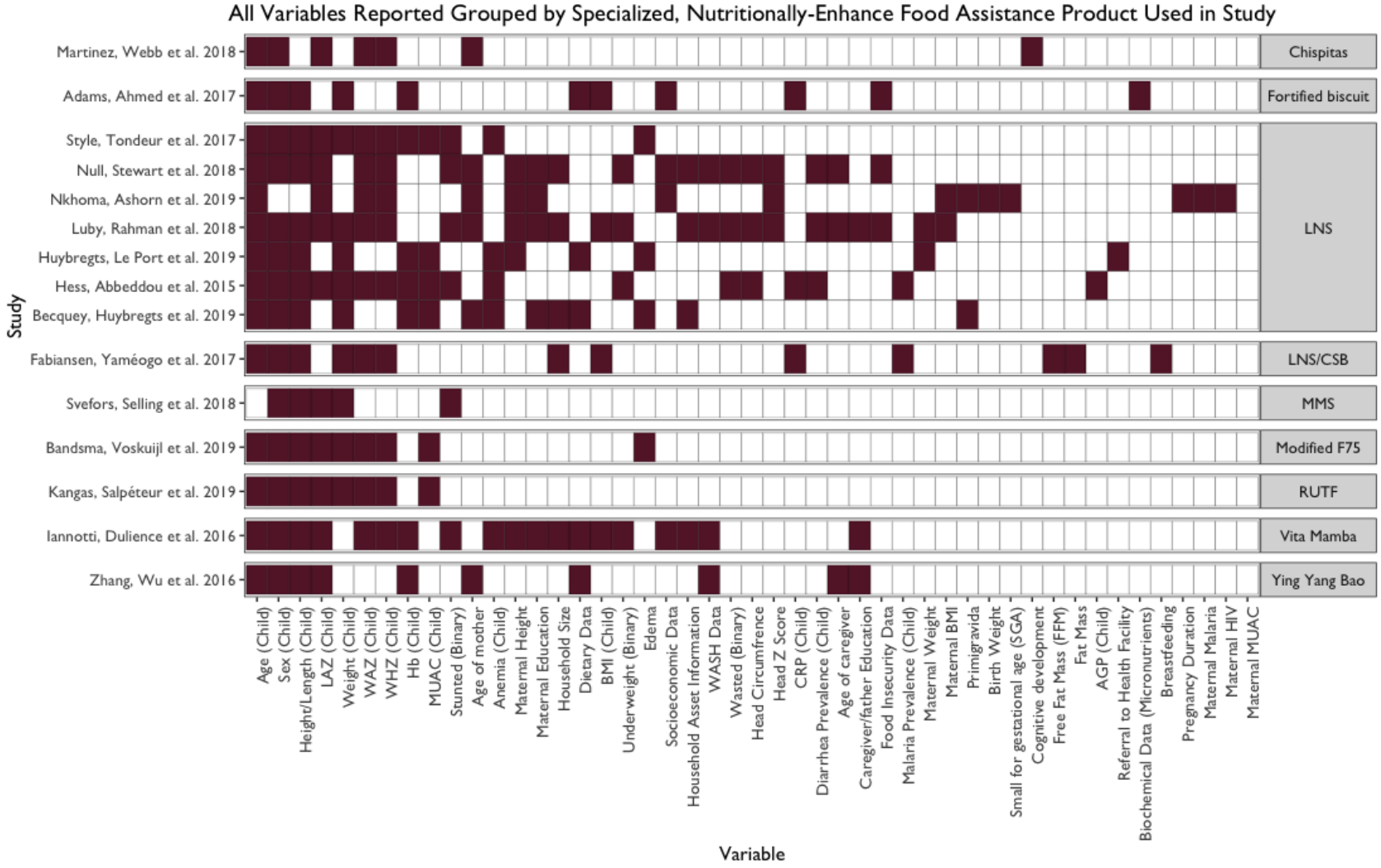


Figure 5: Exposure Variables Reported in Recovered Datasets, Grouped by Age Group and Population Recruited for Corresponding Study.

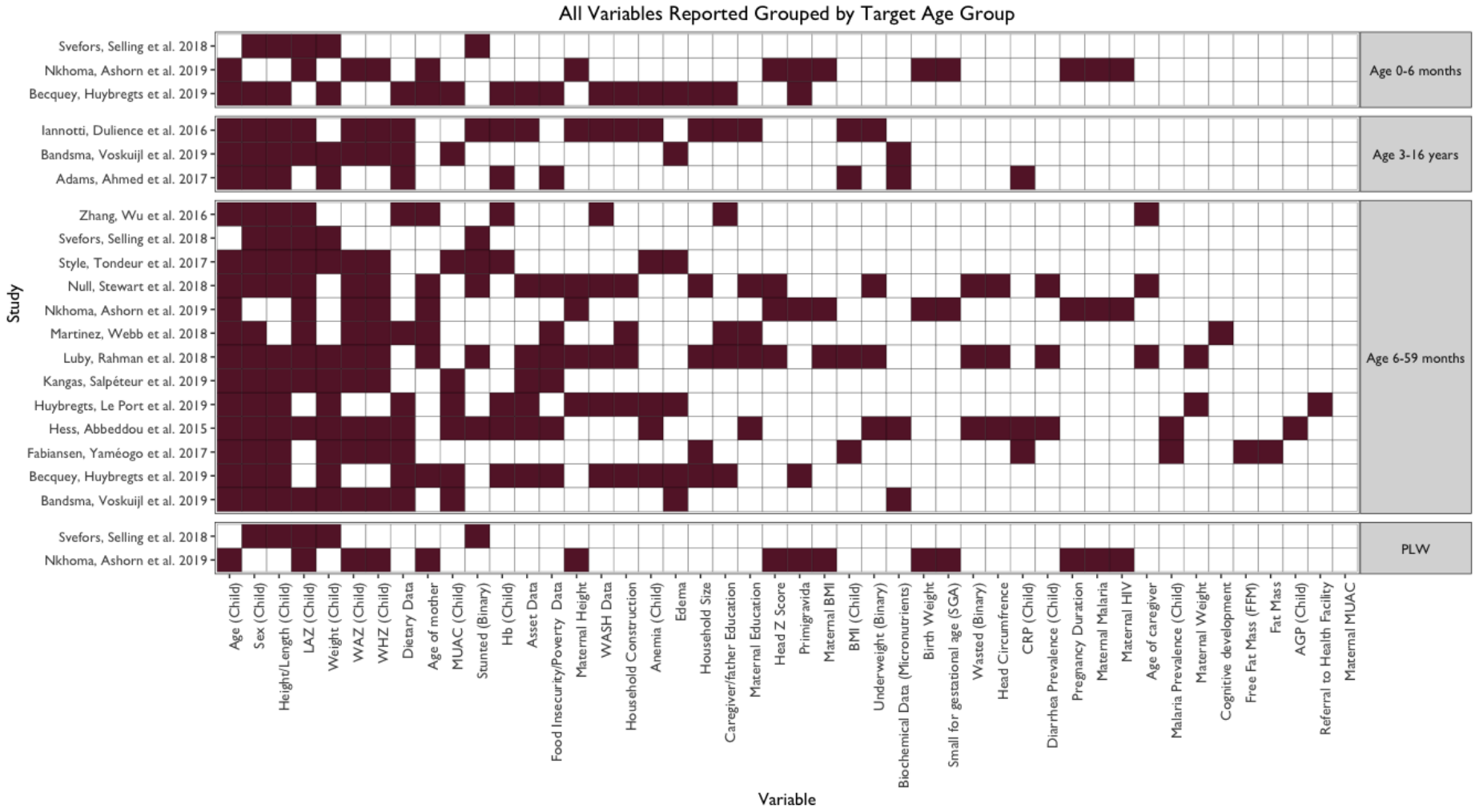


Figure 6: Exposure Variables Reported for Studies with HAZ as an Outcome.

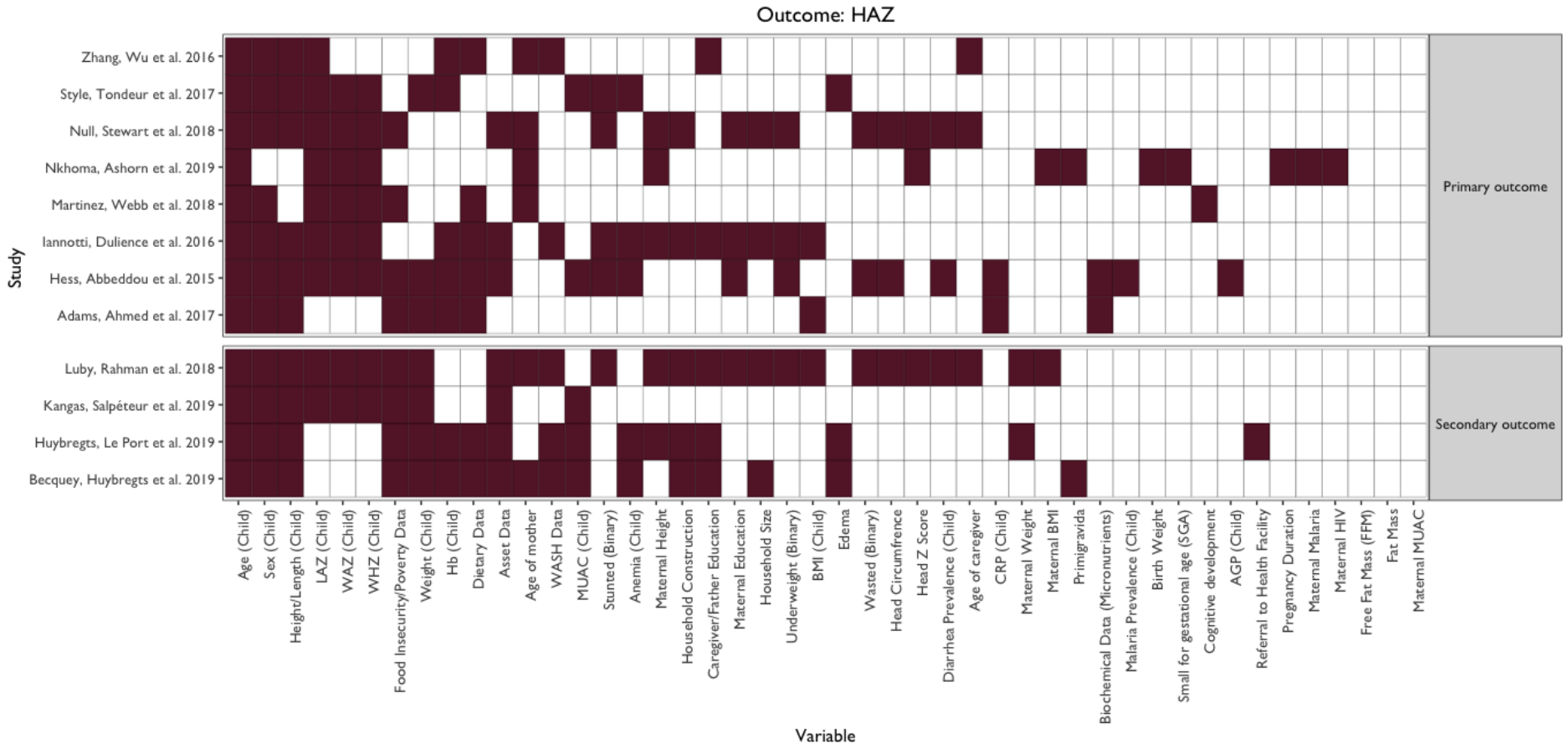


Figure 8: Exposure Variables Reported for Studies with WLZ as an Outcome.

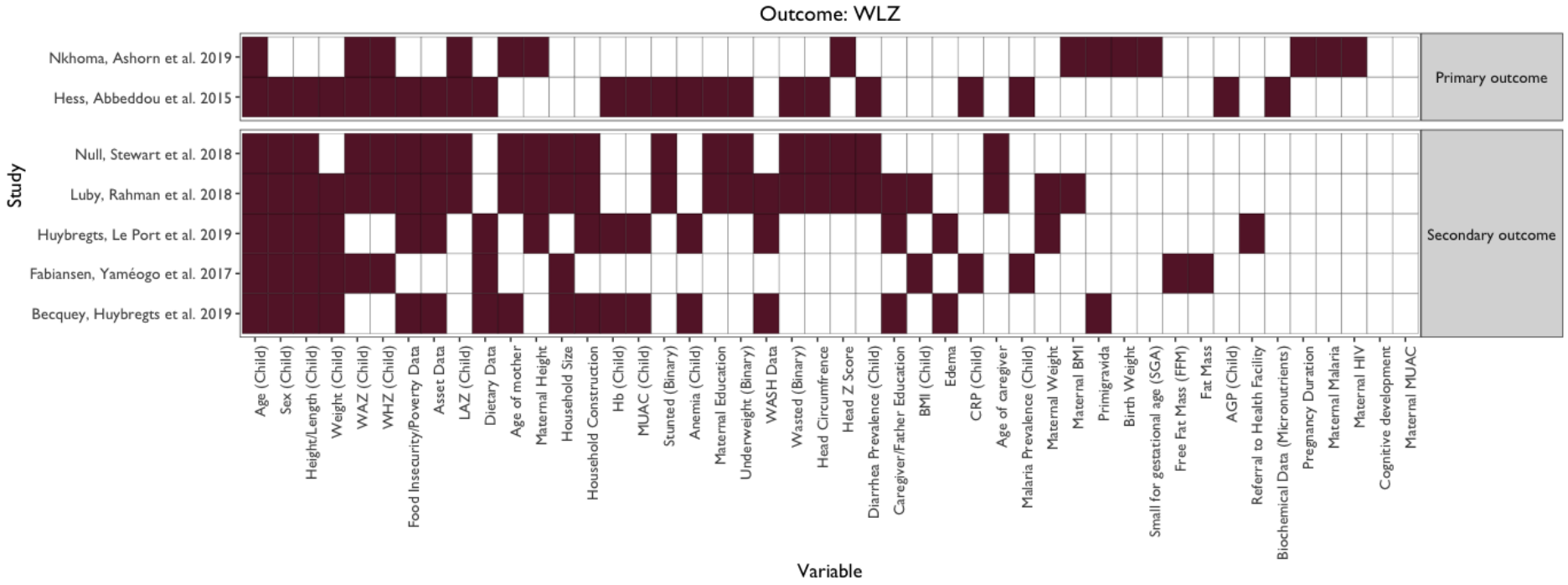


Figure 9: Exposure Variables Reported for Studies with WAZ as an Outcome.

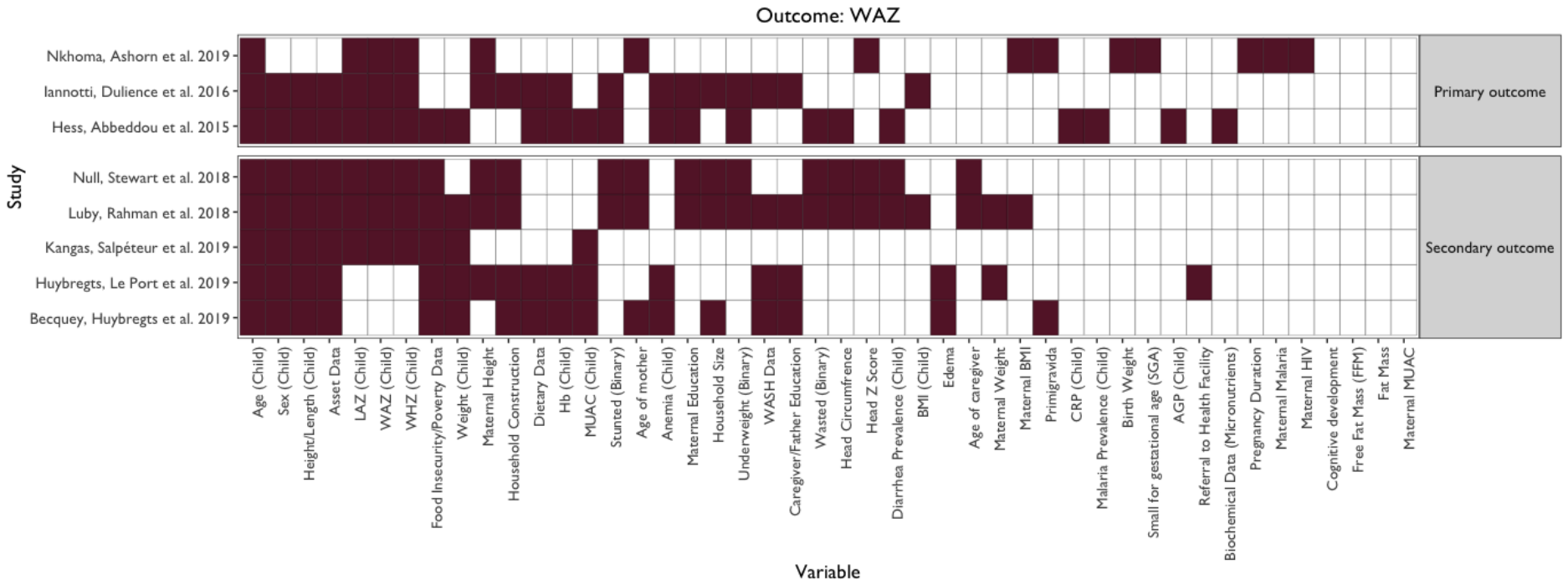


Table 4: All Outcomes Reported in Studies with Open Access Datasets.

(AUTHOR, YEAR) COUNTRY	ALL OUTCOMES (PRIMARY, SECONDARY AND TERTIARY)
Zhang et al., 2016[26] China	**Primary**: Anemia (determined by Hb), stunting **Secondary**: Indicators of infant feeding practices, adherence to complimentary food supplement
Hess et al., 2015[11] Burkina Faso	**Primary**: Growth, incidence of diarrhea, incidence of malaria and pZn Measured by: LAZ, WAZ, WLZ, stunting, underweight and wasting, MUAC, HC (growth), incidence of diarrhea (diarrhea was defined as the presence of ≥0.25 loose or liquid stools per 0.54 hr), malaria by positive RDT, pZn (plasma-measured zinc adjusted for inflammation (CRP and AGP) **Secondary**: Hb, ZPP
Fabiansen et al., 2017[13] Burkina Faso	**Primary**: FFMI increment measured using the deuterium dilution technique **Secondary**: Increments in FFM, FM, FMI, weight, length, knee-heel length, MUAC, triceps skinfold, and nutritional recovery
Nkhoma et al., 2019[12] Malawi	**Primary**: Rate of GWG (g/ week), pregnancy duration (gestation weeks), birth weight (grams), LAZ, WAZ, WLZ, HCZ, inadequate GWG (GWG below the lower limit of the recommended weight gain according to the Institute of Medicine's (IOM) recommendations, PTB (delivery occurring <0.257 gestation weeks), LBW (birth weight <0.5500g), stunting (LAZ <-0.5 at birth) underweight (WAZ<-0.5 at birth), wasting (WLZ<-0.5), small head circumference (HCZ <-0.5 at birth) and SGA (birth weight below the 10th centile of the Intergrowth Standards)
Becquey et al., 2019[14] Burkina Faso	**Primary**: Prevalence of AM (defined by a WLZ < -0.5 or a MUAC < 10.55 mm (only in children older than 6 months) or the presence of bilateral pitting edema, AM screening coverage (number of children screened in the month preceding the survey (as reported by the caregiver) over the total number of eligible study children; and AM treatment compliance defined as the number of AM children under appropriate treatment at the time of the survey over the total number of AM cases identified in the study sample **Secondary**: Anthropometric outcomes mean WLZ, MUAC and Length-for-Age Z-score(LAZ), the prevalence of child stunting (LAZ < -0.5 SD) and severe stunting (LAZ < -0.25 SD), the prevalence of SAM (defined by WLZ < -0.25 SD, a MUAC < 115 mm (children older than 6 months), or bilateral pitting edema)); mean hemoglobin (Hb) concentration, child anemia (Hb < 11 g/dL) and severe anemia (Hb <7 g/dL); caregivers' ENA and EHA and IYCF knowledge and practices
Huybregts et al., 2019[15] Mali	**Primary**: Incidence of AM (same definition of AM as above); monthly AM screening coverage (the number of children screened each month over the total number of eligible study children); AM treatment compliance (the number of AM children adhering to weekly or bi-weekly treatment until discharged over the total number of AM children that were scheduled for treatment) **Secondary**: Relapse rate after treatment of MAM and SAM (number of MAM or SAM cases detected after being successfully discharged from MAM or SAM treatment), ponderal and linear growth (monthly WLZ and LAZ increment, respectively), incidence of stunting (LAZ < -0.5 SD), MUAC gain (MUAC increment per month), longitudinal prevalence (ie. number of days of illness divided by the total number of days of observation for each child) of child morbidity (acute respiratory infections, fever, diarrhea, vomiting, and malaria), patterns of child morbidity prevalence over time, caregiver's knowledge and practices related to IYCF and ENA/EHA

(AUTHOR, YEAR) COUNTRY	ALL OUTCOMES (PRIMARY, SECONDARY AND TERTIARY)
Martinez et al., 2018[25] Guatemala	**Primary**: change in length/height-for-age Z score **Secondary**: minimum dietary diversity, minimum meal frequency and minimal acceptable diet (derived from WHO population-level feeding indicators)
Iannotti et al., 2016[21] Haiti	**Primary**: Hb concentration, anemia status (binary) **Secondary**: Changes in anthropometry and undernutrition status (HAZ, WAZ, BMIZ)
Arnold et al., 2013[30] Bangladesh & Kenya	**Primary**: Primary outcomes include length-for-age Z-scores (LAZ) measured 0.54 months after intervention initiation in target children and diarrhea prevalence in compound children <0.256 months old at enrolment **Secondary**: Two additional measures of linear growth (differences in LAZ and stunting prevalence), child development measures (Ages and Stages Questionnaire, communication, gross motor and personal/social domains) and measures of EE (urinary lactulose-to-mannitol ratio, faecal myeloperoxidase, fecal α-1-antitrypsin, fecal neopterin and plasma total IgG as biomarkers) **Tertiary**: Soil-transmitted helminths (Ascaris lumbricoides, Trichuris trichiura, hookworm) and protozoans (Giardia lamblia, Cryptosporidium parvum, Entamoeba histolytica), WAZ, WLZ, underweight, wasted, severely stunted, head circumference, verbal communicative development inventory, WHO motor milestones, acute upper respiratory illness, all-cause mortality
Bandsma et al., 2019[16] Kenya & Malawi	**Primary**: Clinical stabilization (absence of any WHO "danger" sign, loss of edema, tolerating full F75)
Kangas et al., 2019[17] Burkina Faso	**Primary**: Weight gain velocity (g/kg/day) from admission to discharge **Secondary**: weight gain velocity after two weeks, length of stay (LoS), discharge anthropometrics (weight, height, MUAC, WHZ, HAZ, WAZ), linear and MUAC growth, treatment outcome, morbidity (respiratory illness, malaria, diarrhea), and relapse
Svefors et al., 2018[20] Bangladesh	**Primary**: Cost (direct costs for the intervention included micronutrient and food supplements, staff, training and administration, capital, community volunteer time, and recurrent activities, indirect costs comprised the cost of the participants' time and estimated as cost of a laborer when labor cost was the lowest), and , DALYs, which were equal to YLL due to all-cause premature mortality plus YLD due to stunting
Style et al., 2017[18] Kenya & Djibouti	**Primary**: Hb concentration, anemia (binary), stunting severity, GAM (global acute malnutrition)
Adams et al., 2017[19] Bangladesh	**Primary**: Micronutrient status (Hb, plasma ferritin, folic acid, B10.5, plasma retinol, zinc, iodine, vitamin D), anemia, iodine deficiency, vitamin D deficiency, zinc deficiency (all binary)

2.2.2. VARIABLE CODING

Comparison of variables and coding schemes across recovered datasets revealed that studies collected and reported different variables and used different variable names and codes for the same exposure variable, resulting in inconsistencies in variable names, labels, code values, and value labels for the same variable across the datasets. Variables that differed across datasets included socioeconomic status (SES) indicators such as education, and other economic household status indicators like asset, animal and land ownership, as well as dietary and biochemical data. This created challenges for identifying methods for variable concatenation in a manner conducive to pooling.

2.2.2.1. EDUCATION

Higher parental/caregiver education has been shown to be inversely correlated to child malnutrition indicators like stunting and wasting, and is considered a key predictor of child undernutrition.[31] Despite this, coding the education variable continues to be a challenge, owing to variability in education systems across countries and regions.[32] Of the 15 recovered datasets, education was reported as a variable in eight, as shown in **Table 4** below.[11,14,21,26] Four studies reported maternal education only;[11,21,23,25] one reported paternal education only;[26] one reported both maternal and paternal education;[24] and two reported caregiver education.[14] Harmonizing these variables under a 'caregiver education' variable would not be possible, because only a few studies reported who the primary caregiver was.[14,15,26]

Even for studies that reported a similar variable, combining variables was not possible due to differences in response categories used for the same variable across the datasets. For the five studies that reported maternal education, one only reported the number of 'years of schooling,'[25] and the other four categorized this variable into different education levels, such as primary, secondary, and university.[11,21,23,24] Harmonizing education level would require knowledge of each school system in each study setting. Moreover, equivalency to a standard qualification (e.g., a high school diploma) is not defined for studies reporting years of schooling, nor is context for non-traditional students (e.g., part time students, students with learning disabilities, participants in adult education or literacy programs). The response categories between studies are also inconsistent, as one study permitted the following responses: 'no education', 'no formal education', and 'at least one year of formal education', which differ from more commonly used education categories like 'primary,' 'secondary,' and 'university.' The remaining three studies could be harmonized at the level of the maternal education variable;[21,23,24] however, heterogeneity persists across other variables.

2.2.2.2. DIET QUALITY

Harmonization of dietary data also presented challenges. Dietary variables were reported in eight of the 15 recovered datasets.[11,13-16,21,25,26] Six of the eight reported dietary data for the child only,[11,13,16,21,25,26] and two reported for the child and caregiver,[14,15] either as individual foods or food groups. Of the eight datasets that reported child dietary data, three reported on consumption of food groups for measuring dietary diversity at baseline only.[14,15,21] However, one dataset was missing value labels in the codebook for numerical response options.[21] One dataset reported food groups at baseline and end-line, and one reported food groups at baseline, midterm and end-line.[26] Only one reported the Minimum Dietary Diversity (MDD) score at baseline and exit but did not report the food groups used to calculate this.[25] The remaining three only reported on breastfeeding

status,[11,13,16] with two reporting on whether child was breastfeeding,[13,16] and one reporting on the frequency of breastfeeding.[11] Some reported consumption of foods in the previous 24 hours,[14,15,21,25,26] while others reported on beverage consumption following birth.[14,15]

2.2.2.3. WEALTH

Other variable differences were seen in asset ownership, where some studies collected detailed data on ownership of household items,[14,15,21,23,24] some reported on ownership of land,[21,24] others also included ownership of animals,[21,23] and one only reported a rural and urban wealth score but not its components.[17]

2.2.2.4. BIOCHEMICAL DATA

Six studies reported biochemical data, which differed significantly across datasets.[11,13,15,16,21,26]

Publishing of only minimal datasets, exclusion of codebooks/data dictionaries and data collection instruments that define variable construction or derivation, and missing definitions for numeric response values in datasets exacerbated difficulties in prospects for pooling data for further analyses. **Annex Table 2** details further examples of variability in coding across recovered datasets.

Table 5: Coding of Caregiver/Parental Education Variable Across Datasets

(AUTHOR, YEAR)	VARIABLE NAME	VARIABLE DESCRIPTION	RESPONSE CATEGORIES
Mother's Education			
Hess et al., 2015[11]	womaneducation	Mother's education at enrollment	0 – No education 1 – No formal education (Koranic school, alphabetization, or professional) or formal education for less than 1 year in a system without degree 2 – At least 1 year of formal education
Iannotti et al., 2016[21]	Nivo	Maternal education	1 – No education 2 – Primary 3 – Secondary 5 – University
Null et. al., 2018[23] - Kenya	mother_edu	-	0 – Incomplete primary 1 – Complete primary 2 – Any secondary 9 – Don't know
Luby et. al., 2018[24] - Bangladesh	momedu	Mother's education	0 – No education 1 – Primary (1-5y) 2 – Secondary (>5y)
	momeduy	Mother, years of education	-
Martinez et al., 2018[25]	anosescolmadre	Mother's number of years of schooling	-
Father's Education			
Luby et. al., 2018[24] - Bangladesh	dadeduy	Father, years of education	-
Zhang et al., 2016[26]	HHM16b	What highest academic level was father attending?	1 – Primary school 2 – Middle school 3 – High school 4 – College 5 – Master 88 – Don't know
Caregiver's Education			
Becquey et al., 2019[14] – Burkina Faso	mo_s2_q4	Caregiver has been to school (formal or informal)	0 – No 1 - Yes
Huybregts et al., 2019[15]- Mali	mo_s2_q4	Caregiver has been to school (formal or informal)	0 – No 1 – Yes

3. RESULTS

This study identified that most 'open access' data for studies curated by REFINE (relating to food aid and nutrition) are not openly accessible, nor are those available useful for pooled analyses because of the heterogeneity in definitions and variables used. This needs to be actively addressed by research funding agencies like USAID to make the effective mining of already-collected data both feasible and useful to help further inform funding decisions.

Following an in-depth review of 15 publications from our sample, which have archived underlying datasets in an openly accessible format, we found that the studies that employed an LNS as a key aspect of the intervention appeared to be most similar by design (**Figure 3**). However, among these studies, the targeted population, intervention duration, use of LNS (treatment or prevention of AM), complementary interventions, survey design (longitudinal or cross-sectional) and variables reported in the open access dataset differed substantially. This created difficulties for identifying potential research questions that could be addressed with meta-analyses of pooled data from studies that employed an LNS. Additionally, not all studies that met the REFINE search criteria and employed an LNS could be included in such an analysis, as not all were published with an open access dataset. As well, among studies that targeted a population of children aged 0-59 months, variation in survey design and length of study duration resulted in differences in anthropometric data collection and reporting across open access datasets. This resulted in difficulties identifying research questions that could address growth outcomes among this age population with pooled data derived from these datasets.

A review of the open access datasets archived from these studies and visualized in **Figures 7-9** demonstrated the heterogeneity in variables reported among the studies with pre-specified outcomes for LAZ, WLZ, and WAZ scores. Maternal anthropometry, maternal education, as well as household wealth have been observed to be among the strongest correlates for all three measures of child growth, substantiated by several recent meta-analytic studies of DHS and longitudinal growth data from research studies conducted in lower- middle-income countries.[33,34] However, few studies included these covariates in their analyses of these outcomes. Future studies should consider collecting these variables in order to improve effect estimates of interventions designed to promote child growth and prevent or treat undernutrition.

Additionally, the diversity of coding schemes across studies for similar variable categories, such as dietary data, asset data, household food insecurity and poverty data, household construction and WASH data is notable. This created difficulties for harmonization and called to attention the different metrics being used to quantify these predictors of undernutrition. Future work might consider guidelines for more uniform coding for key variables to facilitate meta-analyses of data from food assistance for nutrition research studies, and thereby improve effect estimates for such exposure variables on growth and nutrition outcomes.

4. CONCLUSIONS AND RECOMMENDATIONS

This desk study of open data from recent publications housed within the REFINE database illustrated the challenges that persist when attempting to locate, interpret, and reuse open access data within the global nutrition and food assistance research community. It exemplified how differing data sharing expectations and practices diminished the availability and usability of open data. Just 15 of the 179 publications analyzed for this study shared underlying datasets in an open access format. The absence of some accompanying documentation (codebooks, study protocols, data collection instruments, and replication codes), and the sharing of only minimal datasets resulted in an inability to interpret these data and identify opportunities for further analyses.

This effort calls attention to the work that remains among researchers and funders to collaborate on efforts to enhance the utility of shared research data. Archiving datasets in an open repository is not enough to ensure the usability of those data.[35] Recalling the FAIR principles, open access data used in

this study were findable and accessible but were frequently neither interoperable nor reusable. This indicates that robust guidance and infrastructure for data sharing is needed in order to make open access to data truly meaningful.

Of note, the WASH Benefits trials reviewed in this study shared much of the necessary information and documentation needed for interpretation of underlying datasets and replication of statistical analyses. All datasets, codebooks, and replication code for all publications stemming from data collected as a part of the WASH Benefits trial, are clearly archived on an open access data repository (available at https://osf.io/tprw2/). The WASH Benefits trial may serve as an exemplar for good data sharing practices from a clinical study addressing a food assistance for nutrition intervention.

USAID has encouraged its partners to make research data open access with their Development Data Policy[36] and also maintains a platform for data-sharing known as the Data Development Library (DDL) (https://data.usaid.gov). USAID's Open Data Policy indicates that its staff, contractors and award recipients "must submit any dataset created or collected with USAID funding to the DDL in accordance with the terms and conditions of their awards." However, the policy specifies that research datasets need not be housed in the DDL if they are archived in another "publicly accessible research database…commonly accessed among scientific and academic communities," although "USAID staff or implementing partners" must submit a notice to the DDL regarding "where and how that data can be accessed."[36]

Of the 35 articles archived in REFINE within the last five years that were funded by USAID, only one indicated that the datasets were available on the DDL.[37] Even in that case, the datasets were not found in a search of the DDL. Two USAID-funded studies had open access datasets archived in a repository separate from the DDL, but these studies did not indicate that the data were also available on the DDL, nor were they referenced or found on the DDL. Information regarding any of these 35 articles and their underlying datasets was not available on the DDL itself. In addition, data from FAQR field studies that were provided in the past to USAID for deposition in the DDL are not found in the DDL at this time.

USAID should consider clearer and more concise guidance for partners to make data more accessible from the research it funds. USAID should create a culture of good data stewardship among its partners, while also facilitating transparency, innovation, and discovery among the global research community. As a public funder of research, USAID has a responsibility to ensure data are openly available, intelligible, and usable in order to honor the populations who participated in USAID-funded research studies by maximizing data use and reuse.

This study identified key recommendations for funding organizations, including USAID, which focus on steps to guide partners in fulfilling open data requirements. Implementation of the following provisions will improve the transparency, accessibility and maximization of USAID's research investments, and the valuable data generated by research activities.

Key Recommendations:

1. **Provide guidance on prospective planning for data sharing**
   - Require a detailed data management plan (DMP) specifying data sharing procedures.
   - Include in DMPs: data collection methods, defined roles for research staff assigned to data archival, and plans for long-term accessibility of data.
   - Provide guidance and example DMPs, a mechanism for evaluation of DMPs, and training for evaluators.
   - Include in participant consent forms a statement of intent to openly share de-identified data.

2. **Provide guidance on curation of data for dissemination, repository deposition, and inclusion of accompanying documentation**
   - Specify timeframe to archive underlying data following article publication.
   - Ensure datasets are comprehensive and de-identified.
   - Archive documentation needed for interpretation of dataset contents:
     a. Complete codebooks/data dictionaries with variable names, descriptions, response options/values and value labels, and criteria for indicator calculation and data exclusion.
     b. Data collection instruments with items corresponding to dataset variables and response options.
     c. Syntax used to run analyses to allow for reproduction of results (replication code).
     d. Study protocols detailing when and how datasets were collected (including time points).
   - Clarify preferred location for archiving data:
     a. Require partners to archive complete datasets and accompanying documentation in a publicly accessible database.
     b. Provide a list of preferred open access repositories for archiving of data.
   - Monitor data sharing compliance:
     a. Confirm datasets have been appropriately archived by the deadline and follow up with non-compliant partners.

5. REFERENCES

1. Lavrakas PJ. Encyclopedia of Survey Research Methods. 2008.
2. Pinfield S. A mandate to self archive? The role of open access institutional repositories. *Serials.* 2005;18(1).
3. Wilkinson MD, Dumontier M, Aalbersberg IJ, et al. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data.* 2016;3(1):1-9.
4. van Panhuis WG, Paul P, Emerson C, et al. A systematic review of barriers to data sharing in public health. *BMC Public Health.* 2014;14(1):1144.
5. Chan A-W, Song F, Vickers A, et al. Increasing value and reducing waste: addressing inaccessible research. *The Lancet.* 2014;383(9913):257-266.
6. Bull S, Cheah PY, Denny S, et al. Best Practices for Ethical Sharing of Individual-Level Health Research Data From Low- and Middle-Income Settings. *Journal of Empirical Research on Human Research Ethics.* 2015;10(3):302-313.
7. Walport M, Brest P. Sharing research data to improve public health. *Lancet.* 2011;377(9765):537-539.
8. Wilkinson MD, Dumontier M, Aalbersberg IJ, et al. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data.* 2016;3(1):160018.
9. Carr D, Littler K. Sharing Research Data to Improve Public Health: A Funder Perspective. *J Empir Res Hum Res Ethics.* 2015;10(3):314-316.
10. Lang T. Advancing Global Health Research Through Digital Technology and Sharing Data. *Science.* 2011;331(6018):714.
11. Hess SY, Abbeddou S, Jimenez EY, et al. Small-quantity lipid-based nutrient supplements, regardless of their zinc content, increase growth and reduce the prevalence of stunting and wasting in young burkinabe children: a cluster-randomized trial. *PLoS One.* 2015;10(3):e0122242.
12. Nkhoma M, Ashorn P, Ashorn U, Dewey KG, Gondwe A, Maleta K. Lipid based nutrient supplements during pregnancy may improve foetal growth in HIV infected women - A cohort study. *PLoS One.* 2019;14(5):e0215760.
13. Fabiansen C, Yaméogo CW, Iuel-Brockdorf AS, et al. Effectiveness of food supplements in increasing fat-free tissue accretion in children with moderate acute malnutrition: A randomised 2 × 2 × 3 factorial trial in Burkina Faso. *PLoS Med.* 2017;14(9):e1002387.
14. Becquey E, Huybregts L, Zongrone A, et al. Impact on child acute malnutrition of integrating a preventive nutrition package into facility-based screening for acute malnutrition during well-baby consultation: A cluster-randomized controlled trial in Burkina Faso. *PLoS Med.* 2019;16(8):e1002877.
15. Huybregts L, Le Port A, Becquey E, et al. Impact on child acute malnutrition of integrating small-quantity lipid-based nutrient supplements into community-level screening for acute malnutrition: A cluster-randomized controlled trial in Mali. *PLoS Med.* 2019;16(8):e1002892.
16. Bandsma RHJ, Voskuijl W, Chimwezi E, et al. A reduced-carbohydrate and lactose-free formulation for stabilization among hospitalized children with severe acute malnutrition: A double-blind, randomized controlled trial. *PLoS Med.* 2019;16(2):e1002747.
17. Kangas ST, Salpéteur C, Nikièma V, et al. Impact of reduced dose of ready-to-use therapeutic foods in children with uncomplicated severe acute malnutrition: A randomised non-inferiority trial in Burkina Faso. *PLoS Med.* 2019;16(8):e1002887.

18. Style S, Tondeur M, Grijalva-Eternod C, et al. Assessment of the effectiveness of a small quantity lipid-based nutrient supplement on reducing anaemia and stunting in refugee populations in the Horn of Africa: Secondary data analysis. *PLoS One.* 2017;12(6):e0177556.
19. Adams AM, Ahmed R, Latif AH, et al. Impact of fortified biscuits on micronutrient deficiencies among primary school children in Bangladesh. *PLoS One.* 2017;12(4):e0174673.
20. Svefors P, Selling KE, Shaheen R, Khan AI, Persson L, Lindholm L. Cost-effectiveness of prenatal food and micronutrient interventions on under-five mortality and stunting: Analysis of data from the MINIMat randomized trial, Bangladesh. *PLoS One.* 2018;13(2):e0191260.
21. Iannotti L, Dulience SJ, Joseph S, et al. Fortified Snack Reduced Anemia in Rural School-Aged Children of Haiti: A Cluster-Randomized, Controlled Trial. *PLoS One.* 2016;11(12):e0168121.
22. Data Availability. PLoS Journals. https://journals.plos.org/plosone/s/data-availability. Accessed2020.
23. Null C, Stewart CP, Pickering AJ, et al. Effects of water quality, sanitation, handwashing, and nutritional interventions on diarrhoea and child growth in rural Kenya: a cluster-randomised controlled trial. *The Lancet Global Health.* 2018;6(3):e316-e329.
24. Luby SP, Rahman M, Arnold BF, et al. Effects of water quality, sanitation, handwashing, and nutritional interventions on diarrhoea and child growth in rural Bangladesh: a cluster randomised controlled trial. *The Lancet Global Health.* 2018;6(3):e302-e315.
25. Martinez B, Webb MF, Gonzalez A, Douglas K, Grazioso MDP, Rohloff P. Complementary feeding intervention on stunted Guatemalan children: a randomised controlled trial. *BMJ Paediatr Open.* 2018;2(1):e000213.
26. Zhang Y, Wu Q, Wang W, et al. Effectiveness of complementary food supplements and dietary counselling on anaemia and stunting in children aged 6–23 months in poor areas of Qinghai Province, China: a controlled interventional study. *BMJ Open.* 2016;6(10):e011234.
27. Bangdiwala SI, Bhargava A, O'Connor DP, et al. Statistical methodologies to pool across multiple intervention studies. *Transl Behav Med.* 2016;6(2):228-235.
28. *Stata Statistical Software* [computer program]. College Station, TX: StataCorp, LLC; 2019.
29. Svefors P, Sysoev O, Ekstrom EC, et al. Relative importance of prenatal and postnatal determinants of stunting: data mining approaches to the MINIMat cohort, Bangladesh. *BMJ Open.* 2019;9(8):e025154.
30. Arnold BF, Null C, Luby SP, et al. Cluster-randomised controlled trials of individual and combined water, sanitation, hygiene and nutritional interventions in rural Bangladesh and Kenya: the WASH Benefits study design and rationale. *BMJ Open.* 2013;3(8):e003476.
31. Vollmer S, Bommer C, Krishna A, Harttgen K, Subramanian S. The association of parental education with childhood undernutrition in low- and middle-income countries: comparing the role of paternal and maternal education. *International Journal of Epidemiology.* 2016;46(1):312-323.
32. Ortmanns V, Schneider SL. Harmonization Still Failing? Inconsistency of Education Variables in Cross-National Public Opinion Surveys. *International Journal of Public Opinion Research.* 2015;28(4):562-582.
33. Li Z, Kim R, Vollmer S, Subramanian SV. Factors Associated With Child Stunting, Wasting, and Underweight in 35 Low- and Middle-Income Countries. *JAMA Network Open.* 2020;3(4):e203386-e203386.
34. Mertens A, Benjamin-Chung J, Colford JM, et al. Causes and consequences of child growth failure in low- and middle-income countries. *medRxiv.* 2020:2020.2006.2009.20127100.

35. Pisani E, Aaby P, Breugelmans JG, et al. Beyond open data: realising the health benefits of sharing data. *BMJ*. 2016;355:i5295.
36. USAID. ADS Chapter 579, USAID Development Data. USAID. https://www.usaid.gov/ads/policy/500/579. Published 2020. Accessed.
37. Wilner L, Suri DJ, Langlois BK, Walton SM, Rogers BL. Effective delivery of social and behavior change communication through a Care Group model in a supplementary feeding program. *J Health Popul Nutr*. 2017;36(1):34-34.

6. ANNEXES

6.1.1. LOCATION OF RECOVERED OPEN ACCESS DATASETS AND ACCOMPANYING DOCUMENTATION

Annex Table 1. Location of Open Access Datasets and Availability of Separate Codebook, Data Collection Instrument, Study Protocol and Replication Code

AUTHOR, YEAR (COUNTRY)	DATASET LOCATION	SEPARATE CODEBOOK	DATA COLLECTION INSTRUMENTS	STUDY PROTOCOL	REPLICATION CODE
Zhang et al., 2016[26]	Dryad Digital Repository	☼	✈	✈	✈
Hess et al., 2015[11]	Figshare	☼	✈	✈	✈
Fabiansen et al., 2017[13]	Figshare, Within paper and supporting information files	✈	✈	☼	✈
Nkhoma et al, 2019[12]	OSF	✈	✈	✈	✈
Becquey et al., 2019[14]	IFPRI Dataverse (Harvard Dataverse)	☼	☼	✈	✈
Huybregts et al., 2019[15]	IFPRI Dataverse (Harvard Dataverse)	☼	☼	✈	✈
Martinez et al., 2018[25]	Harvard Dataverse	☼	✈	✈	☼
Iannotti et al., 2016[21]	Figshare public repository	✈	✈	✈	✈
Luby et al., 2018[24]	OSF	☼	☼	☼	☼
Null et al., 2018[23]	OSF	☼	☼	☼	☼
Bandsma et al., 2019[16]	Harvard Dataverse	☼	✈	✈	✈
Kangas et al., 2019[17]	Zenodo data repository	✈	✈	✈	✈
Svefors et al., 2018[20]	Figshare, Within paper and supporting documents	✈	✈	✈	✈
Style et al., 2017[18]	UCL Discovery	✈	✈	✈	✈
Adams et al., 2017[19]	Figshare, Within paper and supplemental files	✈	✈	✈	✈

6.1.2. VARIABLE CODING COMPARISON

Annex Table 2. Heterogeneity in Variable Coding

AUTHOR, YEAR	DIETARY DATA	ASSET DATA (HOUSEHOLD ASSETS, LAND AND ANIMAL OWNERSHIP)	FOOD INSECURITY AND POVERTY DATA	HOUSEHOLD CONSTRUCTION	ALL BIOCHEMICAL DATA (MICRONUTRIENTS, CRP, AGP AND HB)	WASH
Zhang et al., 2016[26]	Dietary data collected for: child only Data collection tool: WHO Maternal, Newborn and Child Health (MNCH)Household Survey and the Demographic Health Survey (DHS) Types of questions: Questions on whether or not child was breastfed or consumed specific beverages and food groups in the previous 24 hours (baseline, midterm and end-line) Response options: 1 - Yes 2 - No 8 - Don't know	-	-	-	Hemoglobin level of child in g/L	Questions on main source of drinking water, type of toilet facility used by household members, presence of water and soap, detergent or local cleaning agent at handwashing station. **Variable examples:** Variable name: HH18 Variable description: What is the main source of drinking water for members of your household? Response options: 1 - Piped water 2 - Public tap/standpipe 3 - Hand press well 4 - Electric pump well 5 - Protected well 6 - Unprotected well 7 - Rainwater 8 - Cart with small tank 9 - Protected spring/tanker truck 10 - Unprotected spring 11 - Surface water 12 - Bottled water 88 - Others Variable name: HH21 Variable description: What kind of toilet facility do members of your household usually use? Response options: 1 - Flush or pour flush toilet 2 - Urine directing toilet 3 - Biogas toilet 4 - Ventilation improved 5 - Attic type composting latrine

AUTHOR, YEAR	DIETARY DATA	ASSET DATA (HOUSEHOLD ASSETS, LAND AND ANIMAL OWNERSHIP)	FOOD INSECURITY AND POVERTY DATA	HOUSEHOLD CONSTRUCTION	ALL BIOCHEMICAL DATA (MICRONUTRIENTS, CRP, AGP AND HB)	WASH
						6 - A deep pit antifreeze toilet 7 - Anti-seepage toilet 8 - Pit latrine with slab 9 - Pit latrine without slab 10 - Bucket toilet 11 - No facility/bush/field 12 - Double weng funnel toilet 13 - Three-squared septic tank 14 - Alternating double pit toilet 88 - Others Variable name: HH30a Variable description: Is there water around washing hand place? Response options: 1 - Yes 2 - No Variable name: HH30b Variable description: Is there any soap or detergent or local ussed cleaning agent around washing hand place? Response options: 1 - Soap/detergent 2 - Mud/sand 3 - Other
Hess et al., 2015[11]	Dietary data collected for: child only Data collection tool: - Types of questions: Continued breastfeeding at enrollment Response options: 0 - Child not breastfed or breastfed less than 6 times per day 1 - Child breastfed more frequently (>=6 times per day)	Variable name: asset_index Variable description: Household asset index at enrollment	Variable name: hfia_09adj Variable description: Household Food Insecurity Access Scale score adjusted for seasons	-	Reported at both 9 and 18 months: Hemoglobin concentration in g/dL Zinc Protoporphyrin adjusted for malaria in umol/mol heme Logarithmic zinc concentration C-reactive protein in mg/L α-1-acid glycoprotein in g/L	-
Fabiansen et al., 2017[13a]	Dietary data collected for: child only. Data collection tool: - Types of questions:	-	-	-	Reported at baseline: Serum CRP level in mg/L	-

AUTHOR, YEAR	DIETARY DATA	ASSET DATA (HOUSEHOLD ASSETS, LAND AND ANIMAL OWNERSHIP)	FOOD INSECURITY AND POVERTY DATA	HOUSEHOLD CONSTRUCTION	ALL BIOCHEMICAL DATA (MICRONUTRIENTS, CRP, AGP AND HB)	WASH
	Still breastfed? Response options: 0 - no value label 1 - no value label					
Nkhoma et al., 2019[12b]						
Becquey et al., 2019[14] PROMIS Burkina Faso (Longitudinal)	Dietary data collected for: caregiver Data collection tool: Types of questions All food groups for dietary diversity measurement (at enrollment) Response options: 0 - No 1 - Yes Dietary data collected for: child 1. After birth, amount of time until child was breastfed for first time (hours)	Questions about household assets owned only **Variable example:** 1. Asset == Axe Response options: 0 - No 1 - Yes	Questions about whether or not hh had food insecurity in the past 4 weeks and frequency **Variable examples:** 1) Past 4 wks, concerned hh did not have enough food? Response options: 0 - No 1 - Yes 2) HH did not have enough food, frequency Response options: 1 - Rarely (once or twice during the four weeks) 2 - Sometimes (three to ten times during the four weeks) 3 - Often (more than ten times during the four weeks)	Variable name: s4_q1 Variable description: Construction material of floor in primary building Response options: 11 - Earth 21 - Banco 31 - Baked bricks 32 - Linoleum (Plastic) 33 - Ceramic tiles, parquet 34 - Cement 35 - Carpet -96 - Other (specify)	-	**Variable examples:** Type of toilet facility used by hh members Response options: 22 - Pit latrine with slab 23 - Pit latrine without slab/open pit 41 - Bucket 51 - No facility/bush/field 52 - Pit manure heap -96 - Other Primary source of drinking water 11 - Piped water: Tap in the house 12 - Piped water: Tap in the concession 13 - Piped water: Public tap (ONEA) 14 - Piped water: Neighbor's tap 21 - Open well: Unprotected well in the house 22 - Open well: Unprotected well in concession 23 - Open well: Unprotected public well 24 - Open well: Neighbor's 31 - Protected well: Covered well in house 32 - Protected well: Covered well in concession 33 - Protected well: Covered public well 34 - Protected well: Neighbor's covered well 41 - Surface water: Spring 42 - Surface water: River/stream 43 - Surface water: Pond/Lake 44 - Surface water: Canal/Dam 51 - Rainwater

AUTHOR, YEAR	DIETARY DATA	ASSET DATA (HOUSEHOLD ASSETS, LAND AND ANIMAL OWNERSHIP)	FOOD INSECURITY AND POVERTY DATA	HOUSEHOLD CONSTRUCTION	ALL BIOCHEMICAL DATA (MICRONUTRIENTS, CRP, AGP AND HB)	WASH
						61 - Tanker truck 81 - Bottled or bagged water
Becquey et al., 2019[14] PROMIS Burkina Faso (Cross-sectional)	Dietary data collected for: caregiver and child Data collection tool: - Types of questions 1. All food groups for dietary diversity measurement (at enrollment) Response options: 0 - No 1 - Yes Dietary data collected for: child only Types of questions 2. Child was breastfed? 3. Child was given any food/liquid before being breastfed 4. Child was given any food/liquid other than breastmilk during first day 5. Child consumed during first day: honey, oil, water, sugar water, brew, animal milk, blessed water, herb infusion, light porridge 6. Child was given colostrum 7. Child is still breastfed Response options for all: 0 - No 1 - Yes 8. Since yesterday, how many times did child eat milk or dairy? 9. Since yesterday, how many times did (name) eat solid, semi-solid or soft food? 10. After birth, amount of time until child was	Questions about household assets owned only **Variable example:** 1. asset = Axe Response options: 0 – No 1 – Yes	Questions about whether or not households had food insecurity in the past 4 weeks and frequency **Variable examples:** 1) Past 4 wks, concerned hh did not have enough food? Response options: 0 - No 1 – Yes 2) HH did not have enough food, frequency Response options: 1 - Rarely (once or twice during the four weeks) 2 - Sometimes (three to ten times during the four weeks) 3 - Often (more than ten times during the four weeks)	Variable name: s4_q1 Variable description: Primary building material: Floor Response options: 11 - Earth 21 - Banco 31 - Baked bricks 32 - Linoleum (Plastic) 33 - Ceramic tiles 34 - Cement 35 - Carpet -96 - Other (specify)	-	**Variable examples:** 1) Type of toilet facility used by hh members Response options: 22 - Pit latrine with slab 23 - Pit latrine without slab/open pit 41 - Bucket 51 - No facility/bush/field 52 - Pit manure heap -96 - Other 2) Primary source of drinking water for HH 11 - Piped water: Tap in the house 12 - Piped water: Tap in the concession 13 - Piped water: Public tap (ONEA) 14 - Piped water: Neighbor's tap 21 - Open well: Unprotected well in the house 22 - Open well: Unprotected well in concession 23 - Open well: Unprotected public well 24 - Open well: Neighbor's 31 - Protected well: Covered well in house 32 - Protected well: Covered well in concession 33 - Protected well: Covered public well 34 - Protected well: Neighbor's covered well 41 - Surface water: Spring 42 - Surface water: River/stream 43 - Surface water: Pond/Lake 44 - Surface water: Canal/Dam 51 - Rainwater 61 - Tanker truck 81 - Bottled or bagged water

AUTHOR, YEAR	DIETARY DATA	ASSET DATA (HOUSEHOLD ASSETS, LAND AND ANIMAL OWNERSHIP)	FOOD INSECURITY AND POVERTY DATA	HOUSEHOLD CONSTRUCTION	ALL BIOCHEMICAL DATA (MICRONUTRIENTS, CRP, AGP AND HB)	WASH
	breastfed for first time (hours)					
Huybregts et al., 2019[15] PROMIS Mali Longitudinal	Dietary data collected for: caregiver Data collection tool: Types of questions: All food groups for dietary diversity measurement (at enrollment) Response options: 0 - No 1 - Yes	Questions about household assets owned only **Variable example:** 1) asset = Bed Response options: 0 - No 1 – Yes	Questions about whether or not hh had food insecurity in the past 4 weeks and frequency **Variable examples:** 1) Past 4 wks, concerned hh did not have enough food? Response options: 0 - No 1 - Yes 2) HH did not have enough food, frequency Response options: 1 - Rarely (once or twice during the four weeks) 2 - Sometimes (three to ten times during the four weeks) 3 - Often (more than ten times during the four weeks)	Variable name: s4_q1 Variable description: Construction material of floor in primary building Response options: 11 - Earth 21 - Banco 31 - Baked bricks 32 - Linoleum (Plastic) 33 - Ceramic tiles, parquet 34 - Cement 35 - Carpet -96 - Other (specify)	-	**Variable examples:** 1) Type of toilet facility used by hh members Response options: 22 - Pit latrine with slab 23 - Pit latrine without slab/open pit 41 - Bucket 51 - No facility/bush/field 52 - Pit manure heap -96 - Other 2) Primary source of drinking water Response options: 11 - Piped water: Tap in the house 12 - Piped water: Tap in the concession 13 - Piped water: Public tap (ONEA) 14 - Piped water: Neighbor's tap 21 - Open well: Unprotected well in the house 22 - Open well: Unprotected well in concession 23 - Open well: Unprotected public well 24 - Open well: Neighbor's 31 - Protected well: Covered well in house 32 - Protected well: Covered well in concession 33 - Protected well: Covered public well 34 - Protected well: Neighbor's covered well 41 - Surface water: Spring 42 - Surface water: River/stream 43 - Surface water: Pond/Lake 44 - Surface water: Canal/Dam 51 - Rainwater 61 - Tanker truck 81 - Bottled or bagged water

AUTHOR, YEAR	DIETARY DATA	ASSET DATA (HOUSEHOLD ASSETS, LAND AND ANIMAL OWNERSHIP)	FOOD INSECURITY AND POVERTY DATA	HOUSEHOLD CONSTRUCTION	ALL BIOCHEMICAL DATA (MICRONUTRIENTS, CRP, AGP AND HB)	WASH
Huybregts et al., 2019[15] PROMIS Mali Cross-sectional	Dietary data collected for: caregiver and child Data collection tool: Types of questions 1. All food groups for dietary diversity measurement (at baseline) Response options: 0 - No 1 - Yes Dietary data collected for: child only 1. Child was breastfed 2. Child was given any food/liquid before being breastfed 3. Child was given any food/liquid other than breast milk during first day 4. Child consumed during first day: honey, oil, water, sugar water, brew, animal milk, blessed water, herb infusion, light porridge 5. Gave liquid to child yesterday: breast milk, water, baby formula, animal milk (powder or fresh), yogurt, fruit juice, tea without milk, tea with milk, other Response options for all: 0 - No 1 - Yes 6. After birth, amount of time until child was breastfed for first time (hours)	Questions about household assets owned only **Variable example:** 1. Asset == Radio Response options: 0 - No 1 – Yes	Questions about whether or not households had food insecurity in the past 4 weeks and frequency **Variable examples:** 1) Past 4 wks, concerned hh did not have enough food? Response options: 0 - No 1 - Yes 2) HH did not have enough food, frequency Response options: 1 - Rarely (once or twice during the four weeks) 2 - Sometimes (three to ten times during the four weeks) 3 - Often (more than ten times during the four weeks)	Variable name: s4_q1 Variable description: Construction material of floor in primary building Response options: 11 - Earth 21 - Banco 31 - Baked bricks 32 - Linoleum (Plastic) 33 - Ceramic tiles, parquet 34 - Cement 35 - Carpet 36 - Grass/Thatch/Mud -96 - Other (specify)	Reported at baseline: Hemoglobin concentration of the child in g/dL	**Variable examples:** 1) Type of toilet facility used by hh members Response options: 22 - Pit latrine with slab 23 - Pit latrine without slab/open pit 41 - Bucket 51 - No facility/bush/field 52 - Pit manure heap 99 - Used neighbor's facilities 2) Primary source of drinking water Response options 11 - Piped water: Tap in the house 12 - Piped water: Tap in the concession 13 - Piped water: Public tap (ONEA) 14 - Piped water: Neighbor's tap 21 - Open well: Unprotected well in the house 22 - Open well: Unprotected well in concession 23 - Open well: Unprotected public well 24 - Open well: Neighbor's 31 - Protected well: Covered well in house 32 - Protected well: Covered well in concession 33 - Protected well: Covered public well 34 - Protected well: Neighbor's covered well 41 - Surface water: Spring 42 - Surface water: River/stream 43 - Surface water: Pond/Lake 44 - Surface water: Canal/Dam 51 - Rainwater 61 - Tanker truck 81 - Bottled or bagged water

AUTHOR, YEAR	DIETARY DATA	ASSET DATA (HOUSEHOLD ASSETS, LAND AND ANIMAL OWNERSHIP)	FOOD INSECURITY AND POVERTY DATA	HOUSEHOLD CONSTRUCTION	ALL BIOCHEMICAL DATA (MICRONUTRIENTS, CRP, AGP AND HB)	WASH
Martinez et al., 2018[25]	Dietary data collected for: child only Data collection tool: - Types of questions: 1) Was the subject breastfeeding during previous day? (at baseline, exit) 2) Is the subject eating solid, semisolid or soft foods? (at baseline only) 3) Did subject eat during previous day? (baseline and exit) 4) Minimum meal frequency (at baseline and exit) 5) Minimum Dietary Diversity (at baseline and exit) 6) Minimum Acceptable Diet (at baseline and exit) Response options for all: 0 - No 1 - Yes 7) Number of meals the subject had during previous day (baseline and exit)	-	Variable name: QPStotal Variable description: Household Quick Poverty Score	-	-	-
Iannotti et al., 2016[21a]	Dietary data collected for: Child Data collection tool: 24 h food frequency Consumption duration: last 24 hrs Types of questions: 1) Consumption of different individual foods and food groups (eg bread, eggs, beef, vegetables, fruits) 2) Child eats breakfast 3) Number of food groups (diversity) Response options for all: -	Questions on ownership of certain household assets, home, land, and farm animals Response options: no value labels	-	Variable name: ate Variable description: Floor material in home Response options: no value labels	Hemoglobin (data collection time points and units not outlined in minimal codebook)	Questions about drinking and washing water source, time taken to get to water source, mode of transportation used, number of times water is fetched daily, quantities used for drinking, cleaning, whether or not water is treated/sanitized before drinking, mode of water sanitization used, type of toilet used and number of people who use the toilet **Variable examples:** 1) Drinking water source Response options: 1 - tab/facet

AUTHOR, YEAR	DIETARY DATA	ASSET DATA (HOUSEHOLD ASSETS, LAND AND ANIMAL OWNERSHIP)	FOOD INSECURITY AND POVERTY DATA	HOUSEHOLD CONSTRUCTION	ALL BIOCHEMICAL DATA (MICRONUTRIENTS, CRP, AGP AND HB)	WASH
	no value labels in codebook					2 - public pump 7 - spring 8 - river/lake 2) HH toilet Response options: no value labels
Null et al., 2018[23] WASH Kenya	-	Questions on ownership of individual household items and number of farm animals owned **Variable examples:** 1) Radio Response options: 0 - No radio 1 - Has radio 9 - Missing/DK 2) How many goats belong to your household?	Variable name: HHS Variable description: Hunger score Response options: 1 - Little to none 2 - Moderate 3 - Severe 9 – Missing	Variable name: floor Variable description: unlabeled Response options: 0 - Earth/dung 1 - Cement/concrete 9 - Missing/DK	-	-
Luby et al., 2018[24] WASH Bangladesh	-	Questions on size of land, and individual asset ownership and quantities **Variable examples:** 1) Acres of land owned 2) Households have radio (no value labels for response options) 3) Number of assets (e.g. clocks, chairs)	Variable name: hfiacat Variable description: HFIA's food insecurity category Response options: 1 - Food secure 2 - Mildly Food Insecure 3 - Moderately Food Insecure 4 - Severely Food Insecure Variable name: hfias Variable description: HFIA's food insecurity score (0-27)	Variable name: floor Variable description: Improved floor (wood, concrete) Response options: 0 or 1 (no value labels)	-	Questions on primary water source (only one option), time to get to water source, water storage and treatment, availability of water in latrine and availability of soap and water at handwashing station **Variable examples:** 1) Primary water source: tubewell: 0 or 1 (no value labels) 2) Store drinking water: 0 or 1 (no value labels) 3) Reported treating water: 0 or 1 (no value labels) 4) Own their latrine (not shared): 0 or 1 (no value labels) 5) Has handwashing station w/in 6 steps of latrine/kitchen, w/ water, soap: 0 or 1 (no value labels)

AUTHOR, YEAR	DIETARY DATA	ASSET DATA (HOUSEHOLD ASSETS, LAND AND ANIMAL OWNERSHIP)	FOOD INSECURITY AND POVERTY DATA	HOUSEHOLD CONSTRUCTION	ALL BIOCHEMICAL DATA (MICRONUTRIENTS, CRP, AGP AND HB)	WASH
						6) Dist (mins) to primary water source
Bandsma et al., 2019[16]	Dietary data reported for: child Data collection tool: - **Variable examples:** 1. Is the child still breastfeeding at all? (Baseline) 2. Non-standard formula milk e.g. Soya (Daily during hospitalization) 3. Any EBM or breastfeeding in 24hrs (Daily during hospitalization) Response options for all: 0 - No 1 – Yes	-	-	-	Reported daily during hospitalization: (no units given in codebook) Hemoglobin Sodium Potassium Creatinine	-
Kangas et al., 2019[17a]	-	Variable name: pc1 Variable description: Rural wealth score Variable name: pc2 Variable description: Urban wealth score	Variable name: hfias_cat Variable description: HFIAs category of the household Response options: 1 - Food secure 2 - Mildly food insecure 3 - Moderately food insecure 4 - Severely food insecure	-	-	-
Svefors et al., 2018[20b]						
Style et al., 2017[18b]						
Adams et al., 2017[19b]						

[a]Codebook generated with Stata.

[b]No codebook archived with dataset.