Simulating the Joint Exposure Distributions of Child Growth Failure:

A Copula Approach

Anoushka Isabel Millear

A thesis

submitted in partial fulfillment of the

requirements for the degree of

Master of Public Health

University of Washington

2018

Committee:

Nicholas Kassebaum

Ashkan Afshin

Program Authorized to Offer Degree:

Public Health

University of Washington

Abstract

Simulating the Joint Exposure Distributions of Child Growth Failure:

A Copula Approach

Anoushka Isabel Millear

Chair of the Supervisory Committee:

Nicholas Kassebaum

Department of Global Health

As the global health community prioritizes reducing the global burden of child growth failure (stunting, underweight, and wasting [CGF]) in the next decade, comprehensive estimation of CGF should consider the joint distributions of these conditions, as by definition the burden of one is connected to the burden of the others. This study uses anthropometric data from 618 population surveys to assess the marginal distributions of the CGF indicators, the copulas that connect the joint distributions of those indicators, and the difference in CGF prevalence between empirical and simulated joint distributions. The results show that a single distribution and copula family cannot be used to describe the different age-sex patterns of the joint distributions of CGF. Future joint distribution CGF analyses should incorporate these results and iterate on the analyzed distributions to improve simulation of the joint distribution of CGF.

Introduction

Child growth failure (CGF; more commonly known as undernutrition) is a major contributor to childhood morbidity and mortality, accounting for 19.26% of global under-five all-cause disability adjusted life years (DALYs) in 2016 (second only to low birthweight and short gestation as the largest contributing risk factor) (Institute for Health Metrics and Evaluation, 2016). Stunting, wasting, and underweight, which make up CGF, are associated with an increased risk of diarrhea, measles, and lower respiratory infections (Olofin et al., 2013); therefore, any reduction in stunting, wasting, or underweight should reduce the burden of those diseases as well.

In 2012, the World Health Organization (WHO) Member States (via World Health Assembly resolution WHA 65/6) adopted the *Comprehensive implementation plan on maternal, infant, and young child nutrition*, which outlines six global nutrition targets to guide Member States and international partners in their efforts to reduce the burden of malnutrition (known as the Global Targets 2025) (World Health Organization, 2012; WHO, 2018). These targets include a 40% reduction in global under five stunting and reducing and maintaining global child wasting to less than 5% (WHO, 2014). Additionally, the global Sustainable Development Goals (SDGs) set in the *Agenda for Sustainable Development* by the United Nations in 2015 include SDG 2 (end hunger, achieve food security and improved nutrition and promote sustainable agriculture) and SDG 3 (ensure healthy lives and promote wellbeing for all at all ages), with a target date of 2030 (WHO, 2018; United Nations, n.d.).

Both the tremendous disease burden and political importance of addressing child growth failure highlight the necessity of comprehensive burden estimation. Measuring CGF entails using anthropometric data (height and weight) as well as age and sex, then comparing those values against one another to determine if a child is growing adequately. Stunting is a measure of height-for-age, underweight assesses weight-for-age, and wasting addresses weight-for-height (figure 1). Height, weight, and age are assessed against a standard reference population by calculating a "z-score", a number that describes how far that individual child's height and weight lies from the reference mean or median (WHO, n.d.).

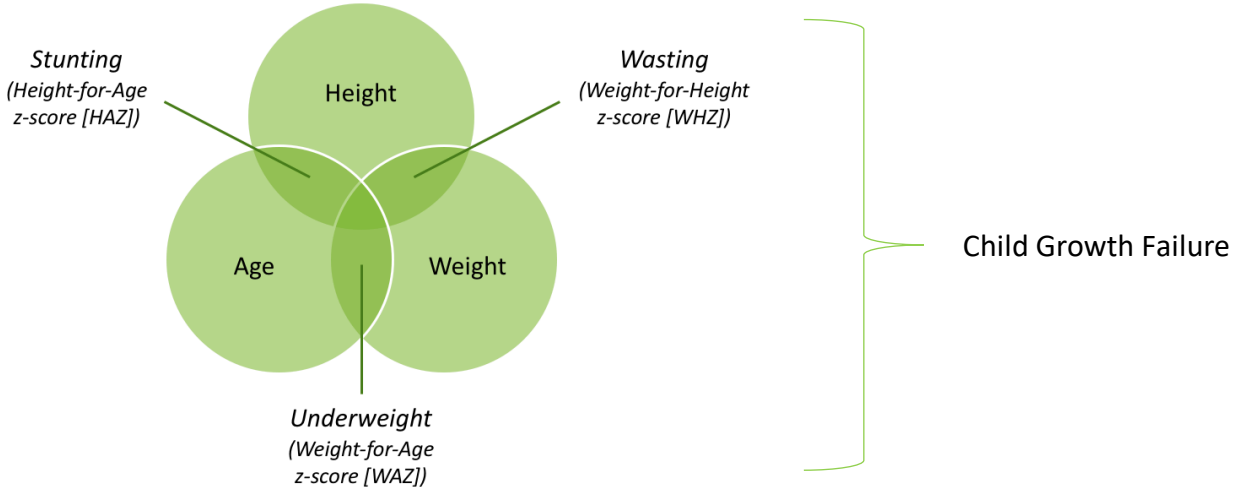


Figure 1: The intersections of height, age, and weight, and their corresponding z-scores

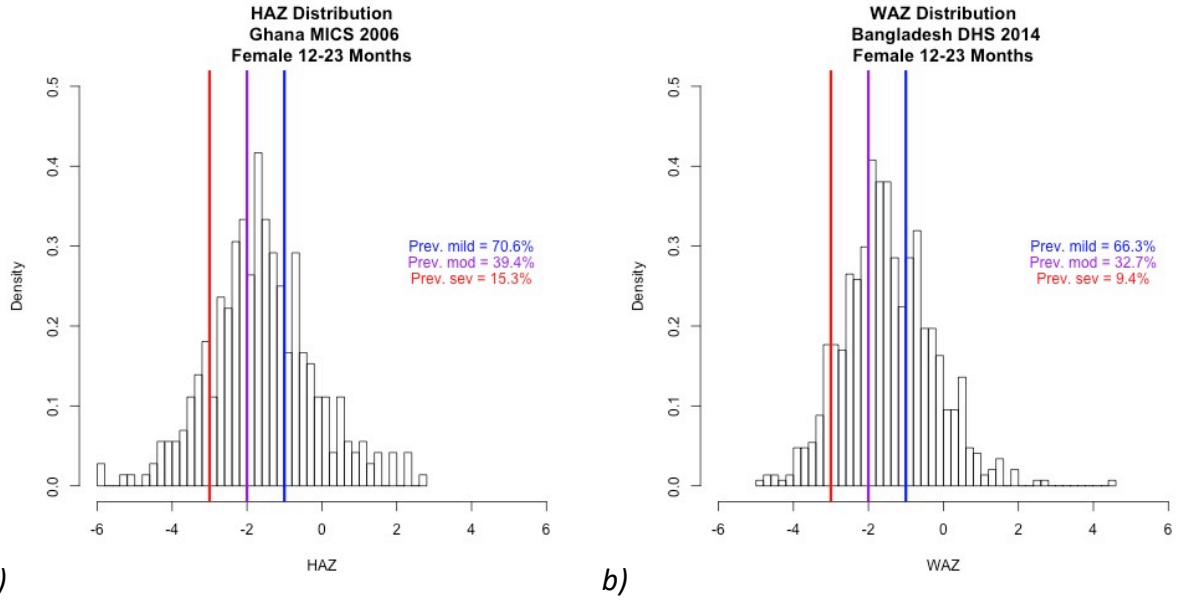
To determine how closely an individual child's growth matches the range of "normal" (typical) growth, the child's height and weight are compared to those of children of a similar age in an international reference population through calculation of a standard metric, the z-score, which allows for comparison across age, sex, geography, and time. Because the mean and standard deviation of the international reference population are used in z-score calculation, the subsequent z-scores necessarily reflect the distribution of that reference population (Wang and Chen, 2012).

In 1977, the United States National Center for Health Statistics (NCHS) published a set of growth charts for use in assessing child growth for 0-18 year old children. The CDC computed z-scores based on the 1977 NCHS percentiles in 1978, which were then recommended by the WHO as a global standard by which to measure child growth and malnutrition. However, concerns about the representativeness of the underlying population on which 1978 NCHS standards were based lead to further iteration on the growth standards. In 2000, the CDC released revised growth charts based on US data (from the National Health and Nutrition Examination Survey III), and in 2006, the WHO published new charts based on data from six urban sites in Brazil, Ghana, India, Norway, Oman, and California (from the 1997-2003 WHO Multicentre Growth Reference Study). The 2006 WHO Growth Standards "describe how healthy children should grow under optimal environmental and health conditions" (Grummer-Strawn, Reinold, Krebs, et al, 2010). The underlying data on which those standards are based constitute a more representative population than the 1978 NCHS standards, and thus, that 1997-2003 Multicentre Growth Reference Study population now serves as the international reference population to which children are compared in CGF estimation.

The distribution of many individual z-scores is used to assess the population level burden of CGF. The proportion of children that have a z-score below certain thresholds can be thought of as that population's "exposure" to CGF, where the -1, -2, and -3 z-score thresholds correspond to mild, moderate, and severe CGF, respectively (figure 2). The population level exposure to CGF informs our understanding of the risk of diarrhea, measles, and lower respiratory infection morbidity and mortality in that population.

Stunting, wasting, and underweight are inherently interconnected, as they share the same input components (height, weight, and age), and share common outcomes (i.e., the same metrics [prevalence of z-scores <-1, <-2, <-3] are used to measure stunting, wasting, and underweight). Thus, comparison of any two of the scores, e.g., stunting and underweight, should reflect that relationship. To do this, the stunting and underweight z-scores of an individual child are compared against one another; to compare the population level distributions, the stunting and underweight z-scores of an entire population are used. This joint distribution, or multivariate probability distribution, consists of three components: the two individual distributions (also known as univariate marginal distributions), and a copula, which describes the form of the connection between the two distributions (the dependence structure) (Schmidt, 2007). Just as different distributions (e.g., normal, Weibull, log-logistic) can be used to

best describe the shape of data, different copula families can be used describe the shape of that connection based on the input data and parameters.



a) *b)*

Figure 2: The distribution of z-scores for a) stunting [height-for-age z-score] in a 2006 Ghana Multiple Indicator Cluster Survey and b) underweight [weight-for-age z-score] in a 2014 Bangladesh Demographic and Health Survey. The blue, purple, and red lines show the cut-off points for mild, moderate, and severe CGF, respectively. The legend shows the proportion of the survey population with a z-score below that threshold, indicating the population level exposure to that particular form of CGF (e.g., in figure 2a, 39.4% of females aged 12-23 months old are moderately stunted).

For any combination of two of the CGF indicators, the combination of the distribution of the first indicator, the distribution of the second indicator, and information about their relationship (the copula) describes the shape of the joint distribution of those CGF indicators. Given the parameters of those two distributions and their corresponding copula, a sample input dataset can be simulated, and can be compared against the original data to see how closely the parameterized version matches the input data. Just as the input distributions can be integrated at the <-1, <-2, and <-3 z-score thresholds to ascertain mild, moderate, and severe prevalence, the same can happen with the joint distributions, to determine the joint prevalence of CGF (e.g., what proportion of the joint distribution is both moderately stunted and moderately wasted?). Importantly, estimated indicator distributions could be used for this simulation, and wouldn't necessarily require distributions from empirical microdata. Finally, because the two input distributions are inherently connected, a range of values for the third distribution can be ascertained, as only a certain range of height, weight, and age values can produce a particular combination of z-scores from each indicator.

In the existing literature, a few studies have employed copulas for undernutrition related research questions. Using 2004 Demographic and Health Survey from Bangladesh, Dancer,

Rammohan and Smith (2007) used the copula approach to establish "whether or not there exists a dependence structure between […] infant mortality and child nutrition" as part of a larger research question on gender differences in survival probabilities and any subsequent differences in child nutrition. Munyamahoro (2016) used Archimedean copulas to assess the relationship between under-five mortality rate and gross domestic product in Rwanda from 1981 to 2015. A study on nonparametric tests of independence tested copulas (among other methods) to assess the nonlinear dependence between childhood malnutrition indices and possible determinants in India (Herwartz and Maxand, 2018). Finally, Klein and Kneib (2015) further the methods used in this paper to "propose simultaneous Bayesian inference for both the marginal distributions and the copula using computationally efficient Markov chain Monte Carlo simulation techniques", using data on childhood undernutrition and macroecology as an illustration. While these studies use copula methods for research related to undernutrition, none assess the relationships between child growth failure indicators themselves.

The research questions for this study are as follows. For a given set of sample survey data, when simulating the joint distributions of child growth failure using copulas: 1) which distributions best describe and fit the univariate distribution of each indicator? 2) can the same copula family be used for each indicator combination, sex, and age, or are these subset data best fit by varying copula families? And 3) for each optimate combination of distribution and copula family, what is the difference between the joint prevalence as described in the microdata and the joint prevalence from the simulated copula data? These research questions lay the foundation for further CGF burden estimation work, as they help outline the assumptions around distribution and copula choice that can or cannot be made when simulating the joint burdens of child growth failure.

Methods

Data

This study uses microdata (individual level data) primarily from nationally-representative household studies such as the Demographic and Healthy Survey (DHS) series, the Multiple Indicator Cluster Survey (MICS) series, and the Living Standards Measurement Study (LSMS) series. These surveys typically employ two-stage stratified cluster sampling, in which aggregation of the households selected for surveying are representative of the country and first administrative unit (e.g., province, state). These surveys are relatively uniform in how data are collected, processed, and presented; standardized questionnaires and data collection procedures make the results more comparable and consistent. Smaller survey series, country specific surveys, and topic specific surveys are also included in this analysis. This analysis includes anthropometric measurements of 3,600,179 children under five years old (0-59 months) from 618 surveys, collected in 115 countries from 1986 to 2017.

Each survey dataset contains information about individual survey respondents, including their height and weight, as well as their birthdate and date of interview. Age is calculated from the difference between birthdate and date of interview; if only month and year are available, then age is months is used. Age is transformed into age in weeks for 13 weeks old or less, and into

months for >13 weeks to 5 years old, for use in z-score calculation. Height and weight are reported in centimeters and kilograms, respectively. Guidelines and most study protocols recommend that recumbent length measurements are taken for children under 2 years old and standing height measurements are taken for children 2-5 years old. Although anthropometric data collection is standardized across surveys, there may still be inconsistencies or bias present in the data due to poorly calibrated instruments or restlessness of study participants.

Data processing and cleaning removed any records for which height, weight, or age were missing, as well as any height or weight measurements that contained error codes (e.g., 99999 as an indicator in DHS surveys that the original entry was unreliable in some way). Z-scores are calculated using the LMS (lambda, mu, and sigma) method, in which age-, sex-, and indicator-specific lambda, mu, and sigma values (from the 2006 WHO Growth Charts), as well as the individual anthropometric measurements (*y*), are used in the following formula:

$$z = \frac{\left(y/M\right)^{L} - 1}{SL}$$

(Wang and Chen, 2012).

The formula produces a z-score for that specific indicator, e.g., using the LMS values from the height-for-age growth chart produces a height-for-age z-score. Finally, any indicator-specific extreme values (i.e., <-6 or >6 for HAZ, <-6 or >5 for WAZ, or <-5 or >5 for WHZ) were dropped, as they don't meet statistical plausibility criteria (Crowe, Seal, Grijalva-Eternod, et al., 2014; while there may be extreme cases outside of these bounds that are lost due to this exclusion, these cutoffs align with the cutoffs used in the CGF analysis from the Global Burden of Disease study (Gakidou, Afshin, Abajobir, et al., 2017). The final dataset includes age, sex, HAZ, WAZ, WHZ, location, year, and survey for each individual child.

Once z-score calculation is complete, the data are subset into age- and sex-specific groups. The age groups of analysis are 0-6 days, 7-27 days, 28-364 days, 12-23 months, and 2-4 years. The first three groups align with the age groups of the GBD 2016, while the latter two are split apart from GBD 2016's 1-4 years old age group, under a working assumption that CGF affects children 12-23 months and 2-4 years differently. Contextual, environmental, and biological conditions contribute to differential exposure to CGF in males and females; thus, this analysis is sex-specific. For computational efficiency, all age-sex subsets of data were reduced to a random sample of 100,000 individuals if the subset was greater than that.

Table 1: Number of surveys, total sample size, and median sample size per survey used in analysis, by super region and year

Super Region	1986	1987	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001
Central Europe, Eastern Europe, and Central Asia	0	0	0	0	0	0	0	0	1	1	1	1	1	1	6	1
	0	0	0	0	0	0	0	0	784	760	1105	1002	1279	586	13867	580
	0	0	0	0	0	0	0	0	784	760	1105	1002	1279	586	1670	580
High-income	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1309	0
	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1309	0
Latin America and Caribbean	4	2	0	2	2	3	2	3	3	4	5	3	9	7	10	5
	6498	3108	0	2746	4404	8311	4767	4817	5945	14595	26307	4485	31883	10597	36873	15057
	1659.5	1554	0	1373	2202	3336	2383.5	751	2573	2698	3880	1444	3618	606	2247	2696
North Africa and Middle East	0	1	2	1	1	1	3	2	0	1	1	1	2	0	4	0
	0	5494	4089	50	6925	4086	12411	3506	0	10560	336	5795	32157	0	51722	0
	0	5494	2044.5	50	6925	4086	4712	1753	0	10560	336	5795	16078.5	0	12575.5	0
South Asia	0	0	0	0	1	3	1	1	0	0	3	1	2	3	3	2
	0	0	0	0	989	11310	15703	22455	0	0	8313	2557	9851	34859	10022	9159
	0	0	0	0	989	3770	15703	22455	0	0	2704	2557	4925.5	10273	3362	4579.5
Southeast Asia, East Asia, and Oceania	0	2	0	0	0	1	1	3	1	0	1	2	1	1	7	1
	0	3891	0	0	0	1075	655	5162	286	0	988	3023	15321	784	29237	2145
	0	1945.5	0	0	0	1075	655	2139	286	0	988	1511.5	15321	784	3797	2145
Sub-Saharan Africa	4	5	7	4	1	2	9	6	5	5	8	6	10	10	24	9
	3235	4684	10359	3024	6183	6544	30520	19081	8223	13785	27764	25217	31357	34069	114207	39339
	868	982	1708	583	6183	3272	3433	3546.5	1532	1750	3045.5	4487.5	2556.5	2807.5	3499.5	4111

Table 1, cont.

Super Region	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017
Central Europe, Eastern Europe, and Central Asia	2	0	0	11	7	1	1	1	4	4	6	3	3	3	2	0
	1154	0	0	27268	17612	149	566	967	13608	4843	12752	7853	8316	9259	1434	0
	577	0	0	2024	2098	149	566	967	3860	1195	1332.5	1399	2431	3544	717	0
High-income	1	0	1	0	1	0	1	0	1	0	1	0	1	0	0	0
	1446	0	1476	0	1636	0	1462	0	1460	0	600	0	638	0	0	0
	1446	0	1476	0	1636	0	1462	0	1460	0	600	0	638	0	0	0
Latin America and Caribbean	4	4	5	6	11	5	5	5	6	7	11	6	7	3	3	0
	20104	31727	14908	19741	37807	19111	22855	22838	28964	22204	47738	15206	29569	39205	27150	0
	4295	5694	2919	2125.5	2455	3008	4651	1919	2030	2217	4058	903	3781	8057	4463	0
North Africa and Middle East	3	3	5	2	5	2	2	2	4	4	12	8	7	6	0	0
	13147	11660	74322	13524	50838	5465	13506	19772	24361	42199	57653	48274	51073	7401	0	0
	4754	4136	4685	6762	10834	2732.5	6753	9886	6033.5	2966.5	1780	3408.5	7256	813	0	0
South Asia	2	3	3	4	2	2	1	3	7	11	13	9	11	4	4	1
	105387	36332	109006	21929	46550	17500	20835	4921	17867	105365	31312	88735	170637	140670	118838	64
	52693.5	288	6212	4734	23275	8750	20835	2014	1122	2401	2134	841	4318	7722	1641	64
Southeast Asia, East Asia, and Oceania	1	3	2	2	5	4	4	6	4	3	3	1	2	2	2	0
	9080	11382	3021	5395	12006	6885	13050	22011	16803	8902	13831	3	6631	3885	10074	0
	9080	5376	1510.5	2697.5	1332	1384.5	2171.5	4121.5	3071	1221	3301	3	3315.5	1942.5	5037	0
Sub-Saharan Africa	4	9	13	16	25	12	11	17	31	23	21	21	26	10	12	4
	11664	67643	33992	55129	135183	57741	61245	31012	148053	92753	58030	80219	120011	54746	66892	7647
	2871	5379	2097	3115	4018	3996	2306	1324	2588	2089	2470	2509	2711	4339.5	3336.5	2315

Note: Total number of surveys, total sample size, and median number of children per survey by super region-year in each cell.

Analysis

The general analytical steps are split into three sections. The first portion consists of distribution fitting (for each individual indicator), while the second portion focuses on copula selection for the joint distributions of the original data. Finally, the parameters of the two input distributions and the selected copula family are used in data simulation for each joint distribution (combination of indicators). All analysis steps are conducted at the age-sex specific level, for a total of 10 different data subsets (2 sexes * 5 age groups).

Initially, a series of different distributions are fit to the data, identifying a set of parameters that describes that distribution fit. The distributions included in this analysis are the Weibull, log-logistic, and gamma distributions. These distributions were chosen for two reasons. Firstly, CGF estimation in the Global Burden of Disease study formally identified these (and similar) distributions as ones that might fit the distributions of child growth failure (Gakidou et al., 2017). Secondly, these distributions are defined within the R packages used for analysis, unlike other distributions that would require manual definition.

The joint distribution of each combination of CGF indicators (stunting vs underweight, stunting vs wasting, wasting vs underweight) is analyzed using the BiCopSelect() function from the VineCopula package. This function uses maximum likelihood estimation to fit every candidate copula family to the data, and the copula family which returns the lowest Akaike and Bayesian Information Criteria is returned as the selected copula family. That copula family is combined with the parameters of the distribution fit (as identified above) to generate the copula (or copula object) that connects the two indicators.

Finally, 100,000 random samples are taken from the copula object to simulate the joint distribution. That simulated data is compared to the empirical data by calculating the proportion of each dataset that experiences moderate and severe child growth failure, then subtracting to find the difference between the empirical and simulated data. Generally, it is assumed that the smallest difference between the two indicates the best performing combination of distribution and copula.

The analysis was conducted using R version 3.5.1 in RStudio 1.1.456, with the copula, VineCopula, and fitdistrplus packages.

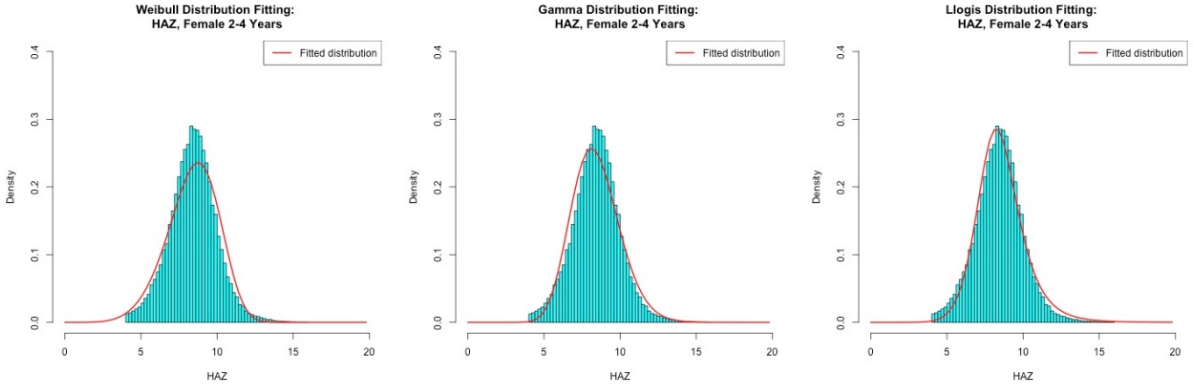
Results

Univariate Distribution Fitting

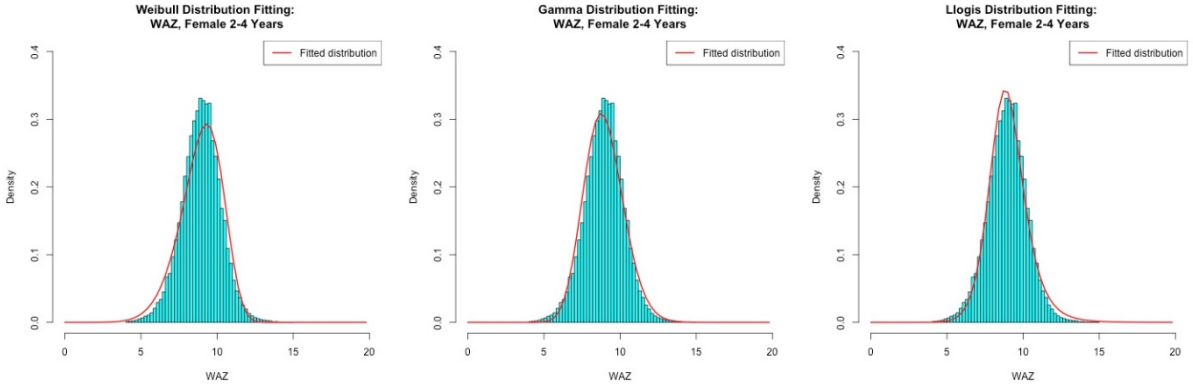
The results of univariate distribution fitting are shown graphically via probability density functions overlaid on microdata (figure 3) and numerically via Akaike Information Criterion (table 2) below. When examined graphically, the log-logistic distribution appears to fit the data the best, as seen for females aged 2-4 years old in figure 3; the log-logistic distribution captures both the left half of the distribution (from which mild, moderate, and severe CGF prevalence are integrated) and the peak of the distribution more faithfully than the Weilbull or the gamma. When compared visually, across all indicators, the distributions fit the 28-364 days, 12-23 months, and 2-4 years age groups better than the 0-6 days and 7-27 days, which may be an artifact of smoother underlying microdata explained by larger sample sizes for the older age groups. Across sexes, the fits for each distribution appear similar.

The Akaike Information Criterion, presented in table 2 below, show a more nuanced picture of which distributions best fit each indicator-sex-age group. For stunting, the log-logistic distribution produced the lowest AIC for all female age groups, and all but one male age group (the gamma distribution had the lowest AIC for males 2-4 years old). For underweight, the log-logistic distribution had the lowest AIC for six of the age groups, while the Weibull and gamma produced the lowest AIC twice each, with no clear age or sex pattern. For wasting, the gamma distribution produced the lowest AIC for the three youngest age groups (both females and males), while the log-logistic produced the lowest AIC for the two older age groups. Notably, although there are differences in the AIC between each distribution, they are relatively similar and don't clearly indicate which distribution is truly the best performer. The disconnect between the visual depiction of distribution fits, which show the log-logistic as the best fit, and the AIC, which lack the clarity of the visual depictions, is noted, and more investigation is needed to explain this discrepancy.

a) Stunting (HAZ) distribution fits, females, 2-4 years old, for Weibull, gamma, and log-logistic distributions



b) Underweight (WAZ) distribution fits, females, 2-4 years old, for Weibull, gamma, and log-logistic distributions



c) Wasting (WHZ) distribution fits, females, 2-4 years old, for Weibull, gamma, and log-logistic distributions

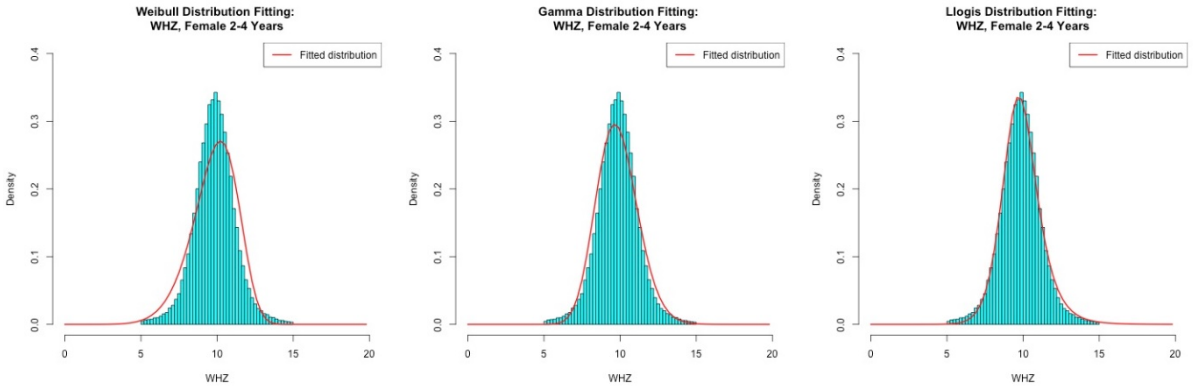


Figure 3: Stunting (HAZ), underweight (WAZ), and wasting (WHZ) distribution fits for females 2-4 years old, using the Weibull, gamma, and log-logistic distributions, with frequency on the y-axis and z-score on the x-axis (offset by 10 to avoid negative z-scores in analysis, e.g., z-score of -3 offset by 10 becomes 7). (See appendices for marginal distributions fits for each indicator, distribution, sex, and age group.)

a) Stunting (HAZ) Akaike Information Criterion, by Sex, Age, and Distribution

Sex	Age	N	Weibull	Gamma	Log-logistic
Female	0-6 Days	8141	32293	30492	*30471*
	7-27 Days	18911	75214	72145	*71991*
	28-364 Days	100000	401482	393543	*391254*
	12-23 Months	100000	398682	386515	*384773*
	2-4 Years	100000	377503	372500	*372387*
Male	0-6 Days	8276	32654	30972	*30944*
	7-27 Days	19468	78271	75878	*75862*
	28-364 Days	100000	409705	403954	*403073*
	12-23 Months	100000	404598	393117	*392729*
	2-4 Years	100000	380440	*375617*	375934

b) Underweight (WAZ) Akaike Information Criterion, by Sex, Age, and Distribution

Sex	Age	N	Weibull	Gamma	Log-logistic
Female	0-6 Days	8141	29835	28422	*28410*
	7-27 Days	18911	67087	65822	*65716*
	28-364 Days	100000	*353629*	354655	353641
	12-23 Months	100000	351273	*351190*	351265
	2-4 Years	100000	338418	335513	*335118*
Male	0-6 Days	8276	30035	*28780*	28794
	7-27 Days	19468	71271	70287	*70159*
	28-364 Days	100000	*364665*	364974	365056
	12-23 Months	100000	358636	*354481*	355556
	2-4 Years	100000	342442	*336898*	336991

c) Wasting (WHZ) Akaike Information Criterion, by Sex, Age, and Distribution

Sex	Age	N	Weibull	Gamma	Log-logistic
Female	0-6 Days	8141	32622	*32459*	32578
	7-27 Days	18911	76094	*75993*	76218
	28-364 Days	100000	382272	*377744*	377882
	12-23 Months	100000	361679	357498	*356388*
	2-4 Years	100000	351283	343917	*340689*
Male	0-6 Days	8276	33382	*33107*	33264
	7-27 Days	19468	80011	*79952*	80344
	28-364 Days	100000	391135	*387957*	388925
	12-23 Months	100000	369620	366771	*366723*
	2-4 Years	100000	359324	354189	*352143*

Table 2: Akaike Information Criterion for each distribution fit by indicator, sex, and age. Each AIC should be evaluated only against the AIC for the other distributions for each indicator, age, and age (e.g., the Weibull, gamma, and log-logistic AIC for stunting in females aged 2-4 years old). The distribution with a lower AIC (italicized) indicates a more suitable fit than the other distributions.

Copula Family Selection

Amongst the different indicator combinations, sex, and age groups, a total of six different copula families were chosen, as shown in table 3. The "t" copula family was most frequently chosen (21 times), followed by the BB1, Rotated BB1 90 Degrees, Rotated BB7 90 Degrees, and Rotated Tawn type 1 180 Degrees, all selected twice. Copula families were chosen consistently across age, with one exception (wasting versus underweight, age 7-27 days).

a) Stunting vs Underweight (HAZ vs WAZ)

Sex	Age	N	Correlation	Copula Family
Female	0-6 Days	8141	0.615	BB1
	7-27 Days	18911	0.604	t
	28-364 Days	100000	0.579	t
	12-23 Months	100000	0.633	t
	2-4 Years	100000	0.674	t
Male	0-6 Days	8276	0.634	BB1
	7-27 Days	19468	0.632	t
	28-364 Days	100000	0.593	t
	12-23 Months	100000	0.634	t
	2-4 Years	100000	0.68	t

b) Stunting vs Wasting (HAZ vs WHZ)

Sex	Age	N	Correlation	Copula Family
Female	0-6 Days	8141	-0.279	Rotated BB1 90 degrees
	7-27 Days	18911	-0.369	t
	28-364 Days	100000	-0.164	Rotated BB7 90 degrees
	12-23 Months	100000	0.064	t
	2-4 Years	100000	-0.03	t
Male	0-6 Days	8276	-0.288	Rotated BB1 90 degrees
	7-27 Days	19468	-0.37	t
	28-364 Days	100000	-0.157	Rotated BB7 90 degrees
	12-23 Months	100000	0.083	t
	2-4 Years	100000	0.023	t

c) Wasting vs Underweight (WHZ vs WAZ)

Sex	Age	N	Correlation	Copula Family
Female	0-6 Days	8141	0.514	Rotated Tawn type 1 180 degrees
	7-27 Days	18911	0.446	t
	28-364 Days	100000	0.642	t
	12-23 Months	100000	0.769	t
	2-4 Years	100000	0.663	t
Male	0-6 Days	8276	0.485	Rotated Tawn type 1 180 degrees
	7-27 Days	19468	0.415	Gaussian
	28-364 Days	100000	0.635	t
	12-23 Months	100000	0.785	t
	2-4 Years	100000	0.701	t

Table 3: The selected copula family for each indicator combination, sex, and age. Each candidate copula family is assessed against the input microdata, and the copula family that results in the lowest Akaike and Bayesian Information Criteria is returned as the selected copula.

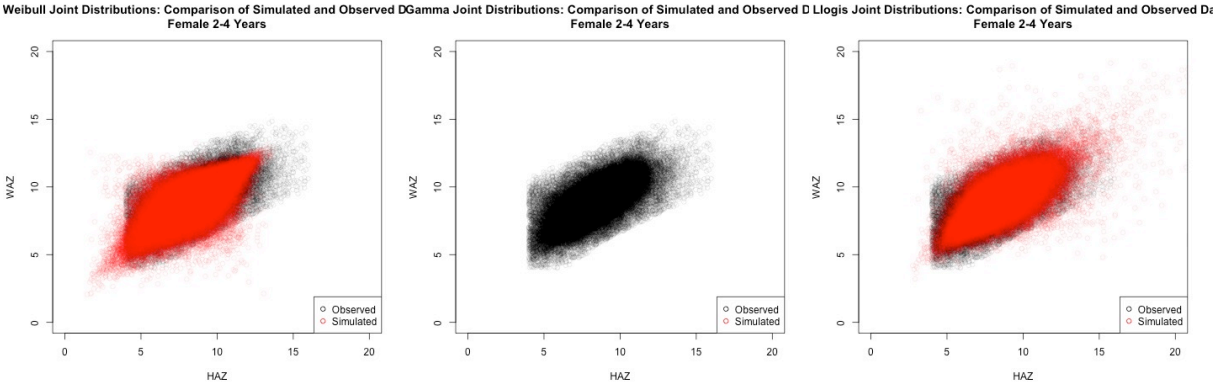
Empirical versus Simulated Joint Distributions

To evaluate the performance of the simulated joint distribution, it is compared to the empirical joint distribution for each combination of CGF indicators both visually and numerically. In figure 4, the microdata of the empirical distribition are shown in black, while the simulated joint distribution is shown in red. Immediately obvious in this comparison is the lack of simulated data for the gamma distributions; the simulated data are extraordinarily high (e.g., greater than 100), which indicates that either the gamma distribution is inappropriate for this analysis or that a specification is incorrect in the code. Regardless, more work is required to assess the suitability of the gamma distribution for this type of analysis.

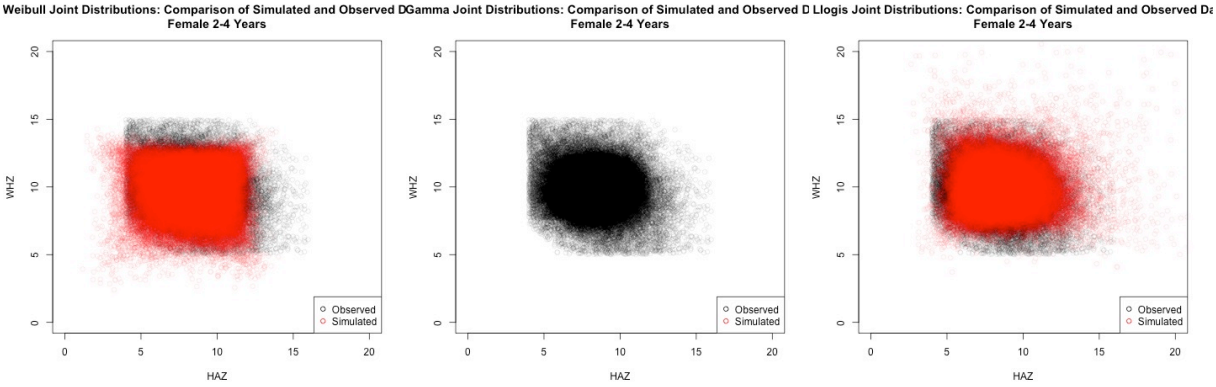
Both the Weibull and the log-logistic distributions overestimate the joint prevalence of the CGF indicators. The difference is generally larger in the older age groups, and is greater for moderate CGF than for severe CGF. This suggests that the simulated data may be capturing the tails (i.e., <-3 z-scores) better than the middle (i.e., <-2) of the joint distribution. As seen in the distribution fitting and copula selection, there doesn't appear to be any appreciable difference for males and females.

When comaring the joint distributions produced using the Weibull and log-logisitic marginal distributions, the log-logistic joint distributions appears to overlap the empirical data more completely than the Weibull joint distributions. This is reflected in table 4, which presents the difference in joint prevalence between the empirical and simulated data. Across the three indicator combinations, the log-logistic distribution minimizes this difference.

a) Stunting versus underweight (HAZ vs WAZ) joint distributions, empirical vs simulated data, females, 2-4 years old, for Weibull, gamma, and log-logistic distributions



b) Stunting versus wasting (HAZ vs WHZ) joint distributions, empirical vs simulated data, females, 2-4 years old, for Weibull, gamma, and log-logistic distributions



c) Wasting versus underweight (WHZ vs WAZ) joint distributions, empirical vs simulated data, females, 2-4 years old, for Weibull, gamma, and log-logistic distributions

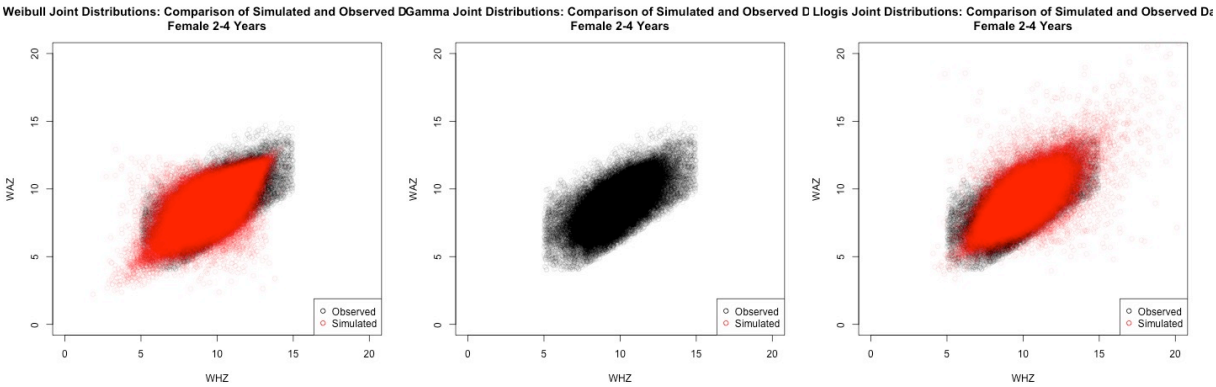


Figure 4: Empirical (black points) versus simulated (red points) joint distributions for each indicator combination and distribution, for females aged 2-4 years old. (All axes offset by 10 to avoid negative z-scores in analysis, e.g., z-score of -3 offset by 10 becomes 7). (See appendices for empirical versus simulated joint distributions for each indicator combination, distribution, sex, and age group.)

a) Stunting vs Underweight (HAZ vs WAZ)

Sex	Age	N	Weibull		Gamma		Log-logistic	
			Moderate	*Severe*	*Moderate*	*Severe*	*Moderate*	*Severe*
Female	0-6 Days	8141	-8.50%	-3.40%	1.80%	0.00%	-1.90%	-0.60%
	7-27 Days	18911	-10.90%	-5.00%	4.20%	0.90%	-4.20%	-0.50%
	28-364 Days	100000	-12.40%	-5.40%	8.20%	2.20%	-6.90%	-0.90%
	12-23 Months	100000	-17.20%	-8.10%	14.80%	4.20%	-13.20%	-2.40%
	2-4 Years	100000	-20.10%	-9.40%	17.60%	4.70%	-16.20%	-2.60%
Male	0-6 Days	8276	-9.10%	-4.00%	2.90%	0.20%	-2.50%	-0.50%
	7-27 Days	19468	-12.30%	-5.70%	6.70%	1.60%	-5.20%	-1.10%
	28-364 Days	100000	-15.10%	-7.10%	11.40%	3.50%	-10.40%	-1.80%
	12-23 Months	100000	-20.10%	-10.30%	19.50%	5.80%	-17.70%	-3.90%
	2-4 Years	100000	-20.10%	-9.60%	18.20%	5.00%	-17.00%	-3.00%

b) Stunting vs Wasting (HAZ vs WHZ)

Sex	Age	N	Weibull		Gamma		Log-logistic	
			Moderate	*Severe*	*Moderate*	*Severe*	*Moderate*	*Severe*
Female	0-6 Days	8141	-2.00%	-0.30%	0.70%	0.00%	0.00%	0.00%
	7-27 Days	18911	-2.10%	-0.70%	0.90%	0.10%	-0.70%	-0.10%
	28-364 Days	100000	-2.80%	-0.60%	1.80%	0.20%	-0.30%	0.10%
	12-23 Months	100000	-7.70%	-3.10%	4.00%	0.60%	-3.50%	-0.70%
	2-4 Years	100000	-7.30%	-2.40%	2.60%	0.30%	-2.70%	-0.40%
Male	0-6 Days	8276	-1.70%	-0.40%	0.80%	0.00%	-0.10%	-0.10%
	7-27 Days	19468	-2.40%	-0.60%	1.30%	0.20%	-1.20%	-0.20%
	28-364 Days	100000	-3.60%	-0.80%	2.90%	0.50%	-1.10%	0.10%
	12-23 Months	100000	-9.20%	-3.90%	6.00%	1.10%	-4.60%	-0.90%
	2-4 Years	100000	-8.10%	-2.90%	3.60%	0.50%	-3.20%	-0.50%

c) Wasting vs Underweight (WHZ vs WAZ)

Sex	Age	N	Weibull		Gamma		Log-logistic	
			Moderate	*Severe*	*Moderate*	*Severe*	*Moderate*	*Severe*
Female	0-6 Days	8141	-4.80%	-1.60%	3.40%	0.50%	-0.40%	0.10%
	7-27 Days	18911	-7.90%	-3.10%	3.60%	0.80%	-2.70%	-0.30%
	28-364 Days	100000	-10.80%	-4.80%	6.40%	1.50%	-5.10%	-1.00%
	12-23 Months	100000	-13.10%	-5.40%	7.00%	1.60%	-5.90%	-1.00%
	2-4 Years	100000	-12.80%	-4.90%	5.40%	1.10%	-4.50%	-0.50%
Male	0-6 Days	8276	-4.50%	-1.50%	3.90%	0.40%	-0.40%	0.00%
	7-27 Days	19468	-6.30%	-2.30%	4.50%	0.90%	-2.10%	-0.30%
	28-364 Days	100000	-11.80%	-5.50%	8.40%	2.20%	-6.40%	-1.30%
	12-23 Months	100000	-15.00%	-7.00%	9.60%	2.40%	-7.80%	-1.40%
	2-4 Years	100000	-13.00%	-5.40%	6.60%	1.40%	-5.40%	-0.80%

Table 4: The difference in moderate and severe prevalence between the empirical and simulated joint distributions, for each indicator combination, distribution, sex, and age. Note that the gamma results should be disregarded due to the extreme values produced by the simulation.

Discussion

These results highlight several important considerations for copula fitting and subsequent simulation of the joint distributions of child growth failure. First and foremost, this analysis shows that using a single copula family to describe the joint distribution for all indicator combinations and ages is inappropriate, as analysis of each indicator group at the age-sex level identifies different copula families for each group. However, the identified copulas are consistent across sexes (i.e., the same copula family is selected for both sexes at the same age), with one exception (WHZ vs WAZ, 7-27 days, for which the "t" copula family and "gaussian" copula family were selected for females and males, respectively). This indicates that age is a stronger determinant than sex when it comes to the joint distribution's copulas.

Additionally, the work emphasizes the importance of appropriate univariate distribution fitting, particularly when it comes to distribution selection. The construction of the simulated joint distribution is obviously dependent on which univariate distributions are used to describe the underlying microdata. For simplicity's sake, this analysis uses single distributions, but ensemble distributions (like the ones employed in the GBD CGF analysis) would fit each marginal better. The AIC shown in table x are quite high; generally, the Weibull distribution produces the highest AIC and the log-logistic the lowest, but across the distributions the AIC are relatively uniform, or rather, they're so high that any difference between them in relatively minimal. Additionally, this analysis lacks any metrics around predictive error; inclusion of these metrics would strengthen the analysis and enhance identification of which distribution is most appropriate. This indicates that further iteration on distribution selection is needed.

Assessment of the simulated joint distributions shows that the log-logistic distribution most consistently minimizes the difference in prevalence between the simulated and empirical data, as seen in table 5 below. Given that the gamma distribution produces such large simulated values, it can't reasonably be considered an option until further investigation into those values determines why they are so high. Disregarding the gamma distribution, the average difference in prevalence between the empirical and simulated joint distributions (as seen in table 5 below) is minimized in the stunting vs wasting (HAZ vs WHZ) joint distributions, with log-logistic marginal distributions as the input. This indicates that the stunting/wasting joint distribution may be the optimal indicator combination to use when assessing all three CGF indicators from the joint distribution of just two input indicators. Notably, the positive correlation between stunting and underweight, and wasting and underweight, can be clearly seen in these joint distribution scatters (see figure 4 for scatters and table 3 for correlation values), while the correlation between stunting and wasting is more multi-directional. This occurs because a sudden increase in height without a simultaneous increase in weight can cause an increase in stunting z-score but a decrease in wasting z-score. Any future work on the relationship between stunting and wasting should investigate this further.

Indicator Combination	Weibull		Log-logistic	
	Moderate	*Severe*	*Moderate*	*Severe*
Stunting vs Underweight (HAZ vs WAZ)	-14.58%	-6.80%	-9.52%	-1.73%
Stunting vs Wasting (HAZ vs WHZ)	-4.69%	-1.57%	-1.74%	-0.27%
Wasting vs Underweight (WHZ vs WAZ)	-10.00%	-4.15%	-4.07%	-0.65%

Table 5: Average percent difference (all ages, both sexes) of simulated vs empirical moderate and severe prevalence, from table 4.

This main strength of this study is found in the size of the dataset. This analysis uses individual level anthropometric data from 3.6 million children and over 600 surveys. Additionally, the data underwent several rounds of rigorous cleaning and testing, including removing implausible values, transforming from non-Gregorian calendars to the Gregorian calendar (for age calculation), and unit conversions to ensure the highest possible data quality. The smallest subset of data used in analysis (females, 0-6 days, wasting [WHZ]) is 8,219, while many subsets of data were restricted to 100,000 random records due to computational limitations. While these can't be claimed as globally representative from a statistical point of view, the size, geographic- and temporal-breadth of the data are as comprehensive has possible given the current landscape of published anthropometric data. Despite the current geographic limitations of this analysis, it could be generalized globally using the Global Burden of Disease CGF risk factor methodological techniques. In that estimation process, microdata and tabulated data are used to model the mean z-score and prevalence of <-2 and <-3 z-scores of each indicator for each age-sex-location-year population, which are then used as inputs into modeling an age-sex-location-year specific distribution for each indicator. If those distributions were input to this analysis for copula fitting and data simulation, then this work could be liberated from its current microdata limitation.

While the dataset is quite large, a limitation of the study is use of these data without further geographical disaggregation (e.g., running the analysis at the super-region level instead of a location aspecific analysis). However, given that the aim of this study was to test if there is a clear "best" combination of two CGF indicators from which the third indicator could be estimated, not to maximize the predictive validity of the simulated estimates, the geographic aggregation is appropriate in this situation. Additionally, the study doesn't consider survey weights included in many of the surveys; these are used to ensure that aggregated population-level metrics are representative at the country and first administrative unit level. While the effect of this is likely to be small, given the number of children included in analysis, the possibility of sampling bias should be considered. While any missing height, weight, or age data led to dropping that record for analysis, the survey data (as published) were not examined for any potential patterns of missingness, which may lead to bias. Although restricting the sample to 100,000 individuals was done randomly, the use of all data would be preferable. The data are simulated only once, whereas a more robust analysis would simulate multiple times to find uncertainty. Finally, this exploratory analysis does not employ any fit statistics for the data simulation; while this presents a larger statistical question outside the scope of this paper, it nonetheless remains a limitation.

This analysis lays the groundwork for future work, with multiple areas for further testing and improvement. For example, the analysis could be expanded to include other distributions, including ensemble distributions (weighted combinations of different distributions), as seen in the Global Burden of Disease risk factors analysis (Gakidou et al., 2017). The work could also be furthered by selecting the distribution that best fits each indicator, then using that indicator-specific distribution when simulating the data, rather than the same distribution for both indicators (e.g., instead of Weibull vs Weibull, using a Weibull vs gamma distributions). Another avenue of analysis would test other growth and nutrition indicators, e.g., analyzing body mass index versus stunting, or height versus age directly.

Another extension of this would be to use the two univariate distributions and the copula to derive the plausible range of values of the third indicator. For example, when considering stunting and underweight, given two z-scores for a child at a specific age there is a finite range of height and weight values that could return both of the initial z-scores, i.e., only certain height and weight values for a child of particular age would produce both the stunting z-score and the underweight z-score in question. Given that range of height and weight values, a plausible range of z-scores could be produced for wasting. Thus, given the burden of two indicators, the burden of all three can be ascertained. A subsequent research question would be which combination of two indicators leads to the least overall measurement error when extrapolating to the third indicator, and how should measurement error be defined in this work. Possibilities include the predicted prevalence of the entire distribution or the predicted prevalence of the tails of the distribution (whether moderate, severe, or both). Another option would be to estimate population attributable fractions per the GBD methodological framework.

This analysis builds on several years of child growth failure modeling work done for the Global Burden of Disease study and highlights several areas for substantial growth in modeling the joint distribution, demonstrating the immense task associated with modeling the global burden of child growth failure. Although the volume of data included in this analysis is quite large, it should be noted that this dataset encapsulates much of the publicly available and accessible child anthropometric data from around the world, and yet many locations and years are quite underrepresented in the dataset. The global health work required to reach the WHO's World Targets 2025 and the UN's Sustainable Development Goals is immense, but without sustained attention to data collection (and publication) and comprehensive burden estimation, any appraisal of progress will be insufficient. The interconnected nature of child growth failures means that efforts to reduce the burden of any indicator necessarily impacts the burden of the other indicators; this analysis helps to lay the groundwork for measuring those changes for child growth failure's collective burden.

References

Crowe, S., Seal, A., Grijalva-Eternod, C., & Kerac, M. (2014). Effect of nutrition survey 'cleaning criteria'on estimates of malnutrition prevalence and disease burden: secondary data analysis. *PeerJ*, *2*, e380.

Dancer, D., Rammohan, A., & Smith, M. D. (2007, May). Infant mortality and Child Nutrition in Bangladesh: Modelling Sample Selection Using Copulas. In *Royal Statistical Society Conference, York, UK*.

Delignette-Muller, M. & Dutang, C. (2015). fitdistrplus: An R Package for Fitting Distributions. Journal of Statistical Software, 64(4), 1-34. URL http://www.jstatsoft.org/v64/i04/.

Gakidou, E., Afshin, A., Abajobir, A. A., Abate, K. H., Abbafati, C., Abbas, K. M., ... & Abu-Raddad, L. J. (2017). Global, regional, and national comparative risk assessment of 84 behavioural, environmental and occupational, and metabolic risks or clusters of risks, 1990–2016: a systematic analysis for the Global Burden of Disease Study 2016. *The Lancet*, *390*(10100), 1345-1422.

Grummer-Strawn, L. M., Reinold, C. M., Krebs, N. F., & Centers for Disease Control and Prevention (CDC). (2010). Use of World Health Organization and CDC growth charts for children aged 0-59 months in the United States.

Herwartz, H., & Maxand, S. (2018). Nonparametric tests for independence: a review and comparative simulation study with an application to malnutrition data in India. *Statistical Papers*, 1-27.

Hofert, M., Kojadinovic, I., Maechler, M., and Yan, J. (2017). copula: Multivariate Dependence with Copulas. R package version 0.999-18 URL https://CRAN.R-project.org/package=copula

Institute for Health Metrics and Evaluation (IHME). GBD Compare Data Visualization. Seattle, WA: IHME, University of Washington, 2016. Available from http:// vizhub.healthdata.org/gbd-compare. (Accessed August 18, 2018)

Klein, N., & Kneib, T. (2016). Simultaneous inference in structured additive conditional copula regression models: a unifying Bayesian approach. *Statistics and Computing*, *26*(4), 841-860.

Munyamahoro, F. (2016). Copula-Based Dependence Measures for Under-Five Mortality Rate in Rwanda. *ARCHIVOS DE MEDICINA*, *2*(4), 34.

Olofin, I., McDonald, C. M., Ezzati, M., Flaxman, S., Black, R. E., Fawzi, W. W., ... & Nutrition Impact Model Study (anthropometry cohort pooling. (2013). Associations of suboptimal growth with all-cause and cause-specific mortality in children under five years: a pooled analysis of ten prospective studies. PloS one, 8(5), e64636.

R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

RStudio Team (2016). RStudio: Integrated Development for R. RStudio, Inc., Boston, MA URL http://www.rstudio.com/.

Schepsmeier, U., Stoeber, J., Christian Brechmann, E., Graeler, B., Nagler, T., and Erhardt, T. (2018). VineCopula: Statistical Inference of Vine Copulas. R package version 2.1.6. https://CRAN.R-project.org/package=VineCopula

Schmidt, T. (2007). Coping with copulas. Copulas-From theory to application in finance, 3-34.

United Nations. (N.d.). *Sustainable Development Goals.* Retrieved from https://sustainabledevelopment.un.org/sdgs.

Wang, Y., & Chen, H. J. (2012). Use of percentiles and z-scores in anthropometry. In *Handbook of anthropometry* (pp. 29-48). Springer, New York, NY.

World Health Organization. (2006). *The WHO Child Growth Standards*. Retrieved from http://www.who.int/childgrowth/standards/en/.

World Health Organization. Comprehensive implementation plan on maternal, infant and young child nutrition. 65th World Health Assembly, World Health Organization, Geneva (2012) http://www.who.int/nutrition/topics/WHA65.6_annex2_en.pdf

World Health Organization. (2014). *Comprehensive Implementation Plan on Maternal, Infant and Young Child Nutrition*. Retrieved from http://apps.who.int/iris/bitstream/handle/10665/113048/WHO_NMH_NHD_14.1_eng.pdf?ua=1.

World Health Organization. (N.d.). *Global Database on Child Growth and Malnutrition: The Z-score or standard deviation classification system*. Retrieved from http://www.who.int/nutgrowthdb/about/introduction/en/index4.html.

World Health Organization. (February 16, 2018). *Malnutrition*. Retrieved from http://www.who.int/news-room/fact-sheets/detail/malnutrition.